Bioassessment in complex environments: designing an index for consistent meaning in different settings

Raphael D. Mazor[1,2,5], Andrew C. Rehn[2,6], Peter R. Ode[2,7], Mark Engeln[1,8], Kenneth C. Schiff[1,9], Eric D. Stein[1,10], David J. Gillett[1,11], David B. Herbst[3,12], and Charles P. Hawkins[4,13]

[1]Southern California Coastal Water Research Project, 3535 Harbor Boulevard, Suite 110, Costa Mesa, California 92626 USA
[2]Aquatic Bioassessment Laboratory, California Department of Fish and Wildlife, 2005 Nimbus Road, Rancho Cordova, California 95670 USA
[3]Sierra Nevada Aquatic Research Laboratory, University of California, 1016 Mt. Morrison Road, Mammoth Lakes, California 93546 USA
[4]Department of Watershed Sciences, Western Center for Monitoring and Assessment of Freshwater Ecosystems, and the Ecology Center, Utah State University, Logan, Utah 84322-5210 USA

Abstract: Regions with great natural environmental complexity present a challenge for attaining 2 key properties of an ideal bioassessment index: 1) index scores anchored to a benchmark of biological expectation that is appropriate for the range of natural environmental conditions at each assessment site, and 2) deviation from the reference benchmark measured equivalently in all settings so that a given index score has the same ecological meaning across the entire region of interest. These properties are particularly important for regulatory applications like biological criteria where errors or inconsistency in estimating site-specific reference condition or deviation from it can lead to management actions with significant financial and resource-protection consequences. We developed an index based on benthic macroinvertebrates for California, USA, a region with great environmental heterogeneity. We evaluated index performance (accuracy, precision, responsiveness, and sensitivity) throughout the region to determine if scores provide equivalent ecological meaning in different settings. Consistent performance across environmental settings was improved by 3 key elements of our approach: 1) use of a large reference data set that represents virtually all of the range of natural gradients in the region, 2) development of predictive models that account for the effects of natural gradients on biological assemblages, and 3) combination of 2 indices of biological condition (a ratio of observed-to-expected taxa [O/E] and a predictive multimetric index [pMMI]) into a single index (the California Stream Condition Index [CSCI]). Evaluation of index performance across broad environmental gradients provides essential information when assessing the suitability of the index for regulatory applications in diverse regions.

Key words: bioassessment, predictive modelling, predictive multimetric index, reference condition

A major challenge for conducting bioassessment in environmentally diverse regions is ensuring that an index provides consistent meaning in different environmental settings. A given score from a robust index should indicate the same biological condition, regardless of location or stream type. However, the performance (e.g., accuracy, precision, responsiveness, and sensitivity) of an index may vary in different settings, complicating its interpretation (Hughes et al. 1986, Yuan et al. 2008, Pont et al. 2009). Effective bioassessment indices should account for naturally occurring variation in aquatic assemblages so that deviations from reference conditions resulting from anthropogenic disturbance are minimally confounded by natural variability (Hughes et al. 1986, Reynoldson et al. 1997). When bioassessment indices are used in regulatory applications, such as measuring compliance with biocriteria (Davis and Simon 1995, Council of European Communities 2000, USEPA 2002, Yoder and Barbour 2009), variable meaning of an index score may lead to poor stream management, particularly if the environmental factors affecting index performance are unrecognized. Those who develop bioassessment indices or the policies that rely on them should evaluate index performance carefully across the different environmental gradients where an index will be applied.

A reference data set that represents the full range of environmental gradients where an index will be used is key for index development in environmentally diverse regions. In addition, reference criteria should be consistently defined so that benchmarks of biological condition are equivalent across environmental settings. Indices based on

E-mail addresses: [5]raphaelm@sccwrp.org; [6]andy.rehn@wildlife.ca.gov; [7]peter.ode@wildlife.ca.gov; [8]mengeln@gmail.com; [9]kens@sccwrp.org; [10]erics@sccwrp.org; [11]davidg@sccwrp.org; [12]david.herbst@lifesci.ucsb.edu; [13]chuck.hawkins@usu.edu

DOI: 10.1086/684130. Received 21 September 2014; Accepted 28 April 2015; Published online 22 October 2015.
Freshwater Science. 2016. 35(1):249–271. © 2016 by The Society for Freshwater Science.

benthic macroinvertebrates (BMI) for use in California were developed with reference data sets that used different criteria in different regions (e.g., Hawkins et al. 2000, Herbst and Silldorff 2009, Rehn 2009). For example, several reference sites used to calibrate an index for the highly urbanized South Coast region had more nonnatural land use than any reference site used to develop an index for the rural North Coast region (Ode et al. 2005, Rehn et al. 2005). Furthermore, lower-elevation settings were poorly represented in these reference data sets. In preparation for establishing statewide biocriteria, regulatory agencies and regulated parties desired a new index based on a larger, more consistently defined reference data set that better represented all environmental settings. Considerable effort was invested to expand the statewide pool of reference sites to support development of a new index (Ode et al. 2016). The diversity of stream environments represented in the reference pool necessitated scoring tools that could handle high levels of complexity.

Predictive modeling of the reference condition is an increasingly common way to obtain site-specific expectations for diverse environmental settings (Hawkins et al. 2010b). Predictive models can be used to set biological expectations at test sites based on the relationship between biological assemblages and environmental factors at reference sites. Thus far, predictive modeling has been applied almost exclusively to multivariate indices focused on taxonomic completeness of a sample, such as measured by the ratio of observed-to-expected taxa (O/E) (Moss et al. 1987, Hawkins et al. 2000, Wright et al. 2000), or location of sites in ordination space (e.g., **BE**nthic **A**ssessment of **S**edimen**T** [BEAST]; Reynoldson et al. 1995). Applications of predictive models to multimetric indices (i.e., predictive multimetric indices [pMMIs]) are relatively new (e.g., Cao et al. 2007, Pont et al. 2009, Vander Laan and Hawkins 2014). MMIs include information on the life-history traits observed within an assemblage (e.g., trophic groups, habitat preferences, pollution tolerances), so they may provide useful information about biological condition that is not incorporated in an index based only on loss of taxa (Gerritsen 1995). Predictive models that set site-specific expectations for biological metric values may improve the accuracy, precision, and sensitivity of MMIs when applied across diverse environmental settings (e.g., Hawkins et al. 2010a).

A combination of multiple indices (specifically, a pMMI and an O/E index) into a single index might provide more consistent measures of biological condition than just one index by itself. Variation in performance of an index would be damped by averaging it with a 2nd index, and poor performance in particular settings might be improved. For example, an O/E index may be particularly sensitive in mountain streams that are expected to be taxonomically rich, whereas a pMMI might be more sensitive in lowland areas, where stressed sites may be well represented in calibration data. Moreover, pMMIs and O/E indices characterize assemblage data in fundamentally different ways. Thus, they provide complementary measures of stream ecological condition and may contribute different types of diagnostic information. Taxonomic completeness, as measured by an O/E index, and ecological structure, as measured by a pMMI, are both important aspects of stream communities, and certain stressors may affect these aspects differently. For example, replacement of native taxa with invasive species may reduce taxonomic completeness, even if the invaders have ecological attributes similar to those of the taxa they displaced (Collier 2009). Therefore, measuring both taxonomic completeness and ecological structure may provide a more complete picture of stream health.

Our goal was to construct a scoring tool for perennial wadeable streams that provides consistent interpretations of biological condition across environmental settings in California, USA. Our approach was to design the tool to maximize the consistency of performance across settings, as indicated by evaluations of accuracy, precision, responsiveness, and sensitivity. We first constructed predictive models for both a taxon loss index (O/E) and a pMMI. Second, we compared the accuracy, precision, responsiveness, and sensitivity of the O/E, pMMI, and combined O/E + pMMI index across a variety of environmental settings. Our primary motivation was to develop biological indices to support regulatory applications in the State of California. However, our broader goal was to produce a robust assessment tool that would support a wide variety of bioassessment applications, such as prioritization of restoration projects or identification of areas with high conservation value.

METHODS

Study region

California contains continental-scale environmental diversity within 424,000 km^2 that encompass some of the most extreme gradients in elevation and climate found in the USA. It has temperate rainforests in the North Coast, deserts in the east, and chaparral, oak woodlands, and grasslands with a Mediterranean climate in coastal regions (Omernik 1987). Large areas of the state are publicly owned, but vast regions have been converted to agricultural (e.g., the Central Valley) or urban (e.g., the South Coast and the San Francisco Bay Area) land uses (Sleeter et al. 2011). Forestry, grazing, mining, other resource extraction activities, and intensive recreation occur throughout rural regions of the state, and the fringes of urban areas are undergoing increasing development. For convenience, we divided the state into 6 regions and 10 subregions based on ecoregional (Omernik 1987) and hydrologic boundaries (California State Water Resources Control Board 2013) (Fig. 1).

Compilation of data

We compiled data from >20 federal, state, and regional monitoring programs. Altogether, we aggregated data from

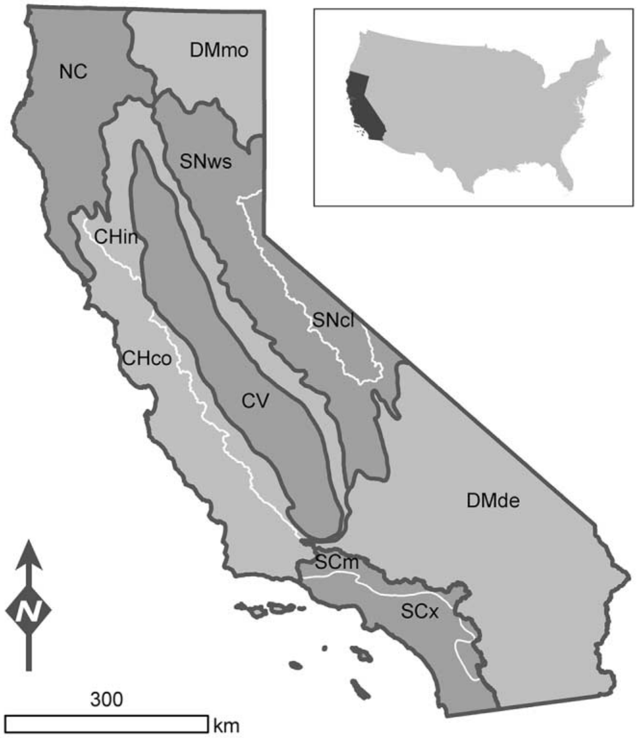

Figure 1. Regions and subregions of California. Thick gray lines indicate regional boundaries, and thin white lines indicate subregional boundaries. NC = North Coast, CHco = Coastal Chaparral, Chin = Interior Chaparral, SCm = South Coast mountains, SCx = South Coast xeric, CV = Central Valley, SNws = Sierra Nevada-western slope, SNcl = Sierra Nevada-central Lahontan, DMmo: Desert/Modoc-Modoc plateau, DMde =Desert/Modoc-deserts.

4457 samples collected from 2352 unique sites between 1999 and 2010 into a single database. We excluded BMI samples with insufficient numbers of organisms or taxonomic resolution (described below) from analyses. We treated observations at sites in close proximity to each other (within 300 m) as repeat samples from a single site. For sites with multiple samples meeting minimum requirements, we randomly selected a single sample for use in all analyses described below, and we withheld repeat samples from all analyses, except where indicated below. We used 1318 sites sampled during probabilistic surveys (e.g., Peck et al. 2006) to estimate the ambient condition of streams (described below).

Biological data

Fifty-five percent of the BMI samples were collected following a reach-wide protocol (Peck et al. 2006), and the other samples were collected with targeted riffle protocols, which produce comparable data (Gerth and Herlihy 2006, Herbst and Silldorff 2006, Rehn et al. 2007). For most samples, taxa were identified to genus, but this level of effort and the total number of organisms/sample varied among samples, necessitating standardization of BMI data. We used different data standardization approaches for the pMMI and the O/E. For the pMMI, we aggregated identifications to 'Level 1' standard taxonomic effort (most insect taxa identified to genus, Chironomidae identified to family) as defined by the Southwest Association of Freshwater Invertebrate Taxonomists (SAFIT; Richards and Rogers 2011) and used computer subsampling to generate 500-count subsamples. We excluded samples with <450 individuals (i.e., not within 10% of target). For the O/E index, we used operational taxonomic units (OTUs) similar to SAFIT Level 1 except that we aggregated Chironomidae to subfamily. We excluded ambiguous taxa (i.e., those identified to a higher level than specified by the OTU). We also excluded samples with >50% ambiguous individuals from O/E development, no matter how many unambiguous individuals remained. We used computer subsampling to generate 400-count subsamples, and we excluded samples with <360 individuals. A smaller subsample size was used for the O/E index than for the pMMI because exclusion of ambiguous taxa often reduced sample size to <500 individuals. A final data set of 3518 samples from 1985 sites met all requirements and was used for development and evaluation of both the O/E and pMMI indices.

Environmental data

We collected environmental data from multiple sources to characterize natural and anthropogenic factors known to affect benthic communities, such as climate, elevation, geology, land cover, road density, hydrologic alteration, and mining (Tables 1, 2). We used geographic information system (GIS) variables that characterized natural, unalterable environmental factors (e.g., topography, geology, climate) as predictors for O/E and pMMI models and variables related to human activity (e.g., land use) to classify sites as reference and to evaluate responsiveness of O/E and pMMI indices to human activity gradients. We calculated most variables related to human activity at 3 spatial scales (within the entire upstream drainage area [watershed], within the contributing area 5 km upstream of a site [5 km], and within the contributing area 1 km upstream of a site [1 km]) so that we could screen sites for local and catchment-scale impacts. We created polygons defining these spatial analysis units using ArcGIS tools (version 9.0; Environmental Systems Research Institute, Redlands, California).

Classification of sites along a human activity gradient

We were unable to measure stress directly with this data set, so instead, we used a human activity gradient under the assumption that it was correlated with stress (Yates and Bailey 2010). We divided sites into 3 sets for development and evaluation of indices: reference (i.e., low activity), moderate-, and high-activity sites. We defined reference

Table 1. Natural gradients and their importance (Gini = mean decrease in Gini index), MSE = % increase in mean squared error) for random-forest models for the observed (O)/expected (E) taxa index and each metric used in the predictive multimetric index (pMMI). Predictors that were evaluated but not selected for any model include % sedimentary geology, nitrogenous geology, soil hydraulic conductivity, soil permeability, S-bearing geology, calcite-bearing geology, and magnesium oxide-bearing geology. Sources: A = National Elevation Dataset (http://ned.usgs.gov/), B = PRISM climate mapping system (http://www.prism.oregonstate.edu), C = generalized geology, mineralogy, and climate data derived for a conductivity prediction model (Olson and Hawkins 2012). Dashes indicate that the predictors were not used to model the metric.

		O/E								
Variable	Description	Gini	MSE	Taxonomic richness MSE	% intolerant MSE	# Shredder taxa MSE	Clinger % taxa MSE	Coleoptera % taxa MSE	EPT % taxa MSE	Data source
Location										
New lat	Latitude	90.5	0.09	18.8	0.0063	1.26	0.0054	0.00079	0.0027	
New long	Longitude	–	–	25.3	0.0058	0.99	0.0030	–	0.0024	
SITE_ELEV	Elevation	89.5	0.11	11.8	–	–	–	0.00231	–	A
Catchment morphology										
LogWSA	Log watershed area	86.6	0.06	–	0.0020	1.23	–	–	–	A
ELEV_RANGE	Elevation range	–	–	2.4	–	–	0.0026	–	–	A
Climate										
PPT	10-y (2000–2009) average precipitation at the sampling point	74.8	0.07	8.4	0.0063	0.92	–	–	0.0016	B
TEMP	10-y (2000–2009) average air temperature at the sampling point	81.9	0.09	9.3	0.0052	–	0.0023	–	0.0019	B
SumAve_P	Mean June to September 1971–2000 monthly precipitation, averaged across the catchment	–	–	5.5	–	–	–	–	0.0033	B
Geology										
BDH_AVE	Average bulk soil density	–	–	5.7	–	–	0.0021	–	–	C
KFCT_AVE	Average soil erodibility factor (k)	–	–	6.2	–	–	0.0027	–	0.0025	C
Log_P_MEAN	Log % P geology	–	–	3.7	–	–	–	–	–	C

Table 2. Stressor and human-activity gradients used to identify reference sites and evaluate index performance. Sites that did not exceed the listed thresholds were used as reference sites. Sources A = National Landcover Data Set (http://www.epa.gov/mrlc/nlcd-2006.html), B = custom roads layer, C = National Hydrography Dataset Plus (http://www.horizon-systems.com/nhdplus), D = National Inventory of Dams (http://geo.usace.army.mil), E = Mineral Resource Data System (http://tin.er.usgs.gov/mrds), F = predicted specific conductance (Olson and Hawkins 2012), G = field-measured variables. WS = watershed, 5 km = watershed clipped to a 5-km buffer of the sampling point, 1 km = watershed clipped to a 1-km buffer of the sampling point, W1_HALL = proximity-weighted human activity index (Kaufmann et al. 1999), Code 21 = landuse category that corresponds to managed vegetation, such as roadsides, lawns, cemeteries, and golf courses. * indicates variable used in the random-forest evaluation of index responsiveness.

	Variable	Scale	Threshold	Unit	Data source
*	% agricultural	1 km, 5 km, WS	<3	%	A
*	% urban	1 km, 5 km, WS	<3	%	A
*	% agricultural + % urban	1 km, 5 km, WS	<5	%	A
*	% Code 21	1 km and 5 km	<7	%	A
*		WS	<10	%	A
*	Road density	1 km, 5 km, WS	<2	km/km^2	B
*	Road crossings	1 km	<5	crossings	B, C
*		5 km	<10	crossings	B, C
*		WS	<50	crossings	B, C
*	Dam distance	WS	<10	km	D
*	% canals and pipelines	WS	<10	%	C
*	Instream gravel mines	5 km	<0.1	mines/km	C, E
*	Producer mines	5 km	0	mines	E
	Specific conductance	Site	99/1[a]	prediction interval	F
	W1_HALL	Reach	<1.5	NA	G
	% sands and fines	Reach		%	G
	Slope	Reach		%	G

[a] The 99th and 1st percentiles of predictions were used to generate site-specific thresholds for specific conductance. The model underpredicted at higher levels of specific conductance (data not shown), so a threshold of 2000 μS/cm was used as an upper bound if the prediction interval included 1000 μS/cm.

sites as 'minimally disturbed' sensu Stoddard et al. (2006) and selected them by applying screening criteria based primarily on landuse variables calculated at multiple spatial scales (i.e., 1 km, 5 km, watershed; Table 2). We calculated some screening criteria at only 1 spatial scale (e.g., in-stream gravel mine density at the 5-km scale and W1_HALL, a proximity-weighted index of human activity based on field observations made within 50 m of a sampling reach; Kaufmann et al. 1999). We excluded sites thought to be affected by grazing or recreation from the reference data set, even if they passed all reference criteria. Identification of high-activity sites was necessary for pMMI calibration (described below) and for performance evaluation of both pMMI and O/E. We defined high-activity sites as meeting any of the following criteria: ≥50% developed land (i.e., % agricultural + % urban) at all spatial scales, ≥5 km/km^2 road density, or W1_HALL ≥ 5. We defined sites not identified as either reference or high-activity as moderate-activity sites. We further divided sites in each set into calibration (80%) and validation (20%) subsets and stratified assignment to calibration and validation sets by subregion to ensure representation of all environmental settings in both sets (Fig. 1). Only 1 reference site was found in the Central Valley, so that region was combined with the Interior Chaparral (whose boundary was within 500 m of the site) for stratification purposes.

Development of the O/E index

Development of an O/E index or pMMI follows the same basic steps: biological characterization, modeling of reference expectations from environmental factors, selection of metrics or taxa, and combining of metrics or taxa into an index. pMMI development has an additional intermediate step to set biological expectations for sites with high levels of activity (Fig. 2). Taxonomic completeness, as measured by O/E, quantifies degraded biological condition as loss of expected native taxa (Hawkins 2006). E represents the number of taxa expected in a specific sample, based on its environmental setting, and O represents the number of those expected taxa that were actually observed. We developed models to calculate the O/E index following the general approach of Moss et al. (1987). First, we defined groups of reference calibration sites based on their

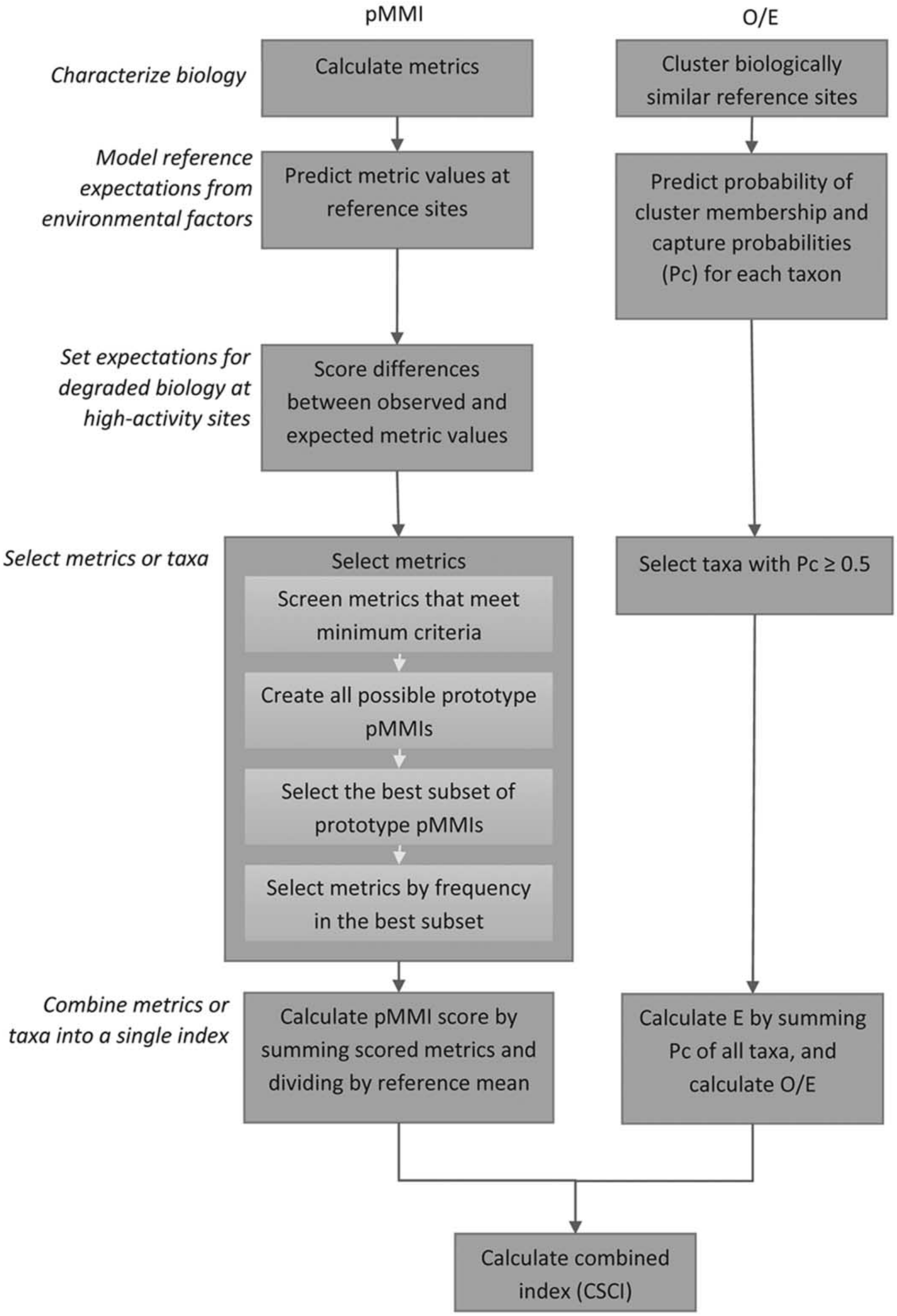
pMMI
O/E
Characterize biology
Calculate metrics
Cluster biologically similar reference sites
Model reference expectations from environmental factors
Predict metric values at reference sites
Predict probability of cluster membership and capture probabilities (Pc) for each taxon
Set expectations for degraded biology at high-activity sites
Score differences between observed and expected metric values
Select metrics or taxa
Select metrics
Screen metrics that meet minimum criteria
Create all possible prototype pMMIs
Select the best subset of prototype pMMIs
Select metrics by frequency in the best subset
Select taxa with Pc ≥ 0.5
Combine metrics or taxa into a single index
Calculate pMMI score by summing scored metrics and dividing by reference mean
Calculate E by summing Pc of all taxa, and calculate O/E
Calculate combined index (CSCI)

Figure 2. Summary of steps in developing the predictive multimetric index (pMMI) and observed (O)/expected (E) taxa index. Pc = probability of observing a taxon at a site, CSCI = California State Condition Index.

taxonomic similarity. Second, we developed a random-forest model (Cutler et al. 2007) to predict group membership based on naturally occurring environmental factors minimally affected by human activities. We used this model to predict cluster membership for test sites based on their natural environmental setting. The probability of observing a taxon at a test site (i.e., the capture probability) was calculated as the cluster-membership-probability-weighted frequencies of occurrence summed across clusters:

$$Pc_j = \sum_{i=1}^{k}(G_i F_i), \qquad \text{(Eq. 1)}$$

where Pc_j is the probability of observing taxon j at a site, G_i is the probability that a site is a member of group i, F_i is the relative frequency of the taxon in group i, and k is the number of groups used in modeling. The sum of the capture probabilities is the expected number of taxa (E) in a sample from a site:

$$E = \sum_{j=1}^{m} Pc_j, \qquad \text{(Eq. 2)}$$

where m is the number of taxa observed across all reference sites. We used Pc values ≥ 0.5 when calculating O/E because excluding locally rare taxa generally improves precision of O/E indices (Hawkins et al. 2000, Van Sickle et al. 2007). This model was used to predict E at reference and nonreference sites based on their natural environmental setting.

We used presence/absence-transformed BMI data from reference calibration sites to identify biologically similar groups of sites. We excluded taxa occurring in <5% of reference calibration samples from the cluster analysis because inclusion of regionally rare taxa can obscure patterns associated with more common taxa (e.g., Gauch 1982, Clarke and Green 1988, Ostermiller and Hawkins 2004). We created a dendrogram with Sørensen's distance measure and flexible β (β = −0.25) unweighted pair group method with arithmetic mean (UPGMA) as the linkage algorithm in R (version 2.15.2; R Project for Statistical Computing, Vienna, Austria) with the *cluster* package (Maechler et al. 2012) and scripts written by J. Van Sickle (US Environmental Protection Agency, personal communication). We identified groups containing ≥10 sites and subtended by relatively long branches (to maximize differences in taxonomic composition among clusters) by visual inspection of the dendrogram. We retained rare taxa that were excluded from the cluster analysis for other steps in index development.

We constructed a 10,000-tree random-forest model with the *randomForest* package in R (Liaw and Wiener 2002) to predict cluster membership for new test sites. We excluded predictors that were moderately to strongly correlated with one another (|Pearson's r| ≥ 0.7). When we observed correlation among predictors, we selected the predictor that was simplest to calculate (e.g., calculated from point data rather than delineated catchments) as a candidate predictor. We used an initial random-forest model based on all possible candidate predictors to identify those predictors that were most important for predicting new test sites into biological groups as measured by the Gini index (Liaw and Wiener 2002). We evaluated different combinations of the most important variables to identify a final, parsimonious model that minimized the standard deviation (SD) of reference site O/E scores at calibration reference sites with the fewest predictors.

We evaluated O/E index performance in 2 ways. First, we compared index precision with the lowest and highest precision possible given the sampling and sample-processing methods used (Van Sickle et al. 2005). SD of O/E index scores produced by a null model (i.e., all sites are in a single group, and capture probabilities for each taxon are the same for all sites) estimates the lowest precision possible for an O/E index. SD of O/E values based on estimates of variability among replicate samples (SDRS) estimates the highest attainable precision possible for the index. Second, we evaluated the index for consistency by regressing O against E for reference sites. Slopes close to 1 and intercepts close to 0 indicate better performance.

Development of the pMMI

We followed the approach of Vander Laan and Hawkins (2014) to develop a pMMI. In contrast to traditional MMIs, which typically attempt to control for the effects of natural factors on biological metrics via landscape classifications or stream typologies, a pMMI accounts for these effects by predicting the expected (i.e., naturally occurring) metric values at reference sites given their specific environmental setting. A pMMI uses the difference between the observed and predicted metric values when scoring biological condition, whereas a traditional MMI uses the raw metric for scoring. Traditional approaches to MMI development may reduce the effects of natural gradients on metric values through classification (e.g., regionalization or typological approaches; see Ode et al. 2005 for a California example), but they seldom produce site-specific expectations for different environmental settings (Hawkins et al. 2010b).

We developed the pMMI in 5 steps (Fig. 2): 1) metric calculation, 2) prediction of metric values at reference sites, 3) metric scoring, 4) metric selection, and 5) assembly of the pMMI. Apart from step 2, the process for developing a pMMI is comparable to that used for a traditional MMI (e.g., Stoddard et al. 2008). We developed a null MMI based on raw values of the selected metrics to allow us to estimate how much predictive modeling improved pMMI performance. The process was intended to produce a pMMI that was unbiased, precise, responsive, and able to characterize a large breadth of ecological attributes of the BMI assemblage.

Metric calculation We calculated biological metrics that characterized the ecological structure of BMI assemblages for each sample in the data set. We used custom scripts in R and the *vegan* package (Oksanen et al. 2013) to calculate a suite of 48 widely used bioassessment metrics, chosen because they quantify important ecological attributes, such as taxonomic richness or trophic diversity (a subset of which is presented in Table 3). Many of these metrics are widely used in other bioassessment indices (e.g., Royer et al. 2001, Stribling et al. 2008). Different formulations of metrics based on taxonomic composition (e.g., Diptera metrics) or traits (e.g., predator metrics) were assigned to thematic metric groups representing different ecological attributes (Table 3). These thematic groups were used to help ensure that the metrics included in the pMMI were ecologically diverse.

Prediction of metric values at reference sites We used random-forest models to predict values for all 48 metrics at reference calibration sites based on the same GIS-derived candidate variables that were used for O/E devel-

Table 3. Metrics evaluated for inclusion in the predictive multimetric index (pMMI). Only metrics that met all evaluation criteria are shown. EPT = Ephemeroptera, Plecoptera, and Trichoptera; Resp = direction of response; I = metric increases with human-activity gradients; D = metric decreases with human-activity gradients; Var Exp = % variance explained by the random-forest model; r^2 (cal) = squared Pearson correlation coefficient between predicted and observed values at reference calibration sites; r^2 (val) = squared Pearson correlation coefficient between predicted and observed values at reference validation sites; t (null) = t-statistic for the comparison of the raw metric between the reference and high-activity samples within the calibration data set; t (mod) = t-statistic for the comparison of the residual metric between the reference and high-activity samples within the calibration data set; F = F-statistic for an analysis of variance of metric residual values from reference calibration sites among regions shown in Fig. 1; S : N = signal-to-noise ratio; Freq = frequency of the metric among the best-performing combinations of metrics. Tolerance, functional feeding group, and habit data were from CAMLnet (2003). * indicates metric selected for inclusion in the pMMI.

Metric	Resp	Var Exp	r^2 (cal)	r^2 (val)	t (null)	t (mod)	F	S : N	Freq
Taxonomic diversity									
*Taxonomic richness	D	0.27	0.27	0.15	21.6	23.7	1.0	6.7	0.83
Functional feeding group									
Scrapers									
No. Scraper taxa	D	0.40	0.40	0.29	15.3	19.1	1.2	7.6	0.17
Shredders									
% Shredder taxa	D	0.27	0.27	0.46	17.6	10.6	1.0	4.1	0.33
* No. Shredder taxa	D	0.39	0.39	0.35	19.2	15.2	1.9	5.4	0.50
Habit									
Clingers									
* % Clinger taxa	D	0.34	0.34	0.42	21.7	14.6	0.2	4.8	1.00
No. Clinger taxa	D	0.39	0.40	0.32	26.0	25.3	0.5	11.1	0
Taxonomy									
Coleoptera									
* % Coleoptera taxa	D	0.30	0.31	0.22	10.3	15.8	1.0	5.0	0.83
No. Coleoptera taxa	D	0.34	0.34	0.29	13.6	20.9	0.6	6.2	0.17
EPT									
* % EPT taxa	D	0.31	0.32	0.46	30.0	23.1	0.4	6.0	0.67
No. EPT taxa	D	0.40	0.40	0.31	27.8	25.3	1.4	10.0	0.17
Tolerance									
* % Intolerant taxa	D	0.23	0.23	0.15	21.7	15.6	0.5	5.1	0.67
% Intolerant taxa	D	0.51	0.51	0.58	32.7	25.3	1.5	6.9	0.17
No. Intolerant taxa	D	0.52	0.52	0.53	28.4	21.8	1.5	9.6	0
Tolerance value	I	0.22	0.25	0.20	−21.5	−17.0	0.4	5.0	0
% Tolerant taxa	I	0.22	0.24	0.38	−26.1	−22.3	1.4	4.9	0.17

opment (Table 1). Manual refinement was impractical because of the large number of models that were developed, so we used an automated approach (recursive feature elimination [RFE]) to select the simplest model (the model with the fewest predictors) whose root mean square error (RMSE) was ≤2% greater than the RMSE of the optimal model (the model with the lowest RMSE). We considered only models with ≤10 predictors. Limiting the complexity of the model typically reduces overfitting and improves model validation (Strobl et al. 2007). We implemented RFE with the *caret* package in R using the default settings for random-forest models (Kuhn et al. 2012). We used the *randomForest* package (Liaw and Wiener 2002) to create a final 500-tree model for each metric based on the predictors used in the model selected by RFE. We then used these models to predict metric values for all sites. We used out-of-bag predictions for the reference calibration set (an out-of-bag prediction is based only on the subset of trees in which a calibration site was excluded during model training). To evaluate how well each model predicted metric values, we regressed raw observed values against predicted values for reference sites. Slopes close to 1 and intercepts close to 0 indicate better model performance. If the pseudo-R^2 of the model (calculated as 1 – mean squared error [MSE]/variance) was >0.2, we used the model to adjust metric values (i.e., observed – predicted), otherwise we used the observed metric values. Hereafter, 'metric' is used to refer to both raw and adjusted metric values.

Metric scoring Scoring is required for MMIs because metrics have different scales and different responses to stress (Blocksom 2003). Scoring transforms metrics to a standard scale ranging from 0 (i.e., most stressed) to 1 (i.e., identical to reference sites). We scored metrics following Cao et al. (2007). We scored metrics that decrease with human activity as

$$(\text{Observed}-\text{Min})/(\text{Max}-\text{Min}), \qquad \text{(Eq. 3)}$$

where Min is the 5th percentile of high-activity calibration sites and Max is the 95th percentile of reference calibration sites. We scored metrics that increase with human activity as

$$(\text{Observed}-\text{Max})/(\text{Min}-\text{Max}), \qquad \text{(Eq. 4)}$$

where Min is the 5th percentile of reference calibration sites, and Max is the 95th percentile of high-activity sites. We trimmed scores outside the range of 0 to 1 to 0 or 1. We used 5th and 95th percentiles instead of minimum or maximum values because they are more robust estimates of metric range than minima and maxima (Blocksom 2003, Stoddard et al. 2008).

Metric selection We selected metrics in a 2-phase process: 1) based on their individual performance, and 2) based on their frequency in high-performing prototype pMMIs. Evaluating the performance of many prototype pMMIs avoids selection of metrics with spuriously good performance and is preferable to selecting metrics or pMMIs based on performance evaluations conducted 1 metric at a time (Hughes et al. 1998, Roth et al. 1998, Angradi et al. 2009, Van Sickle 2010). Initial elimination of metrics based on their individual performance alleviates the computational challenge of evaluating large numbers of prototype pMMIs.

We used several performance criteria to eliminate metrics from further analysis. We assessed responsiveness to human activity by computing *t*-statistics based on comparisons of mean metric values at reference sites and sites with high levels of activity and eliminated metrics with a *t*-statistic < 10. We assessed bias by determining whether metric values varied among predefined geographic regions (Fig. 1). We considered metrics with an *F*-statistic > 2 derived from analysis of variance (ANOVA) by geographic region to have high regional bias and eliminated them. Other screening criteria were modified from Stoddard et al. (2008). We excluded metrics with >⅔ zero values across samples and richness metrics with range < 5. We also eliminated metrics with a signal-to-noise ratio (ratio of between-site to within-site variance estimated from data collected at sites with multiple samples) < 3.

We further screened metrics by evaluating the performance of all possible combinations as prototype pMMIs and selecting metrics that were frequent among prototypes with the best performance. First, we assembled all nonredundant combinations of metrics that met minimum performance criteria into prototype pMMIs. Limiting the redundancy of metrics increases the number of thematic groups included in prototypes, thereby improving the ecological breadth of the pMMI. Redundant combinations of metrics included those with multiple metrics from a single metric group (e.g., tolerance metrics; Table 3) or correlated metrics (|Pearson's $r \geq |0.7|$). Prototype pMMIs ranged in size from a minimum of 5 to a maximum of 10 metrics, a range that is typical of MMIs used for stream bioassessment (e.g., Royer et al. 2001, Fore and Grafe 2002, Ode et al. 2005, Stoddard et al. 2008, Van Sickle 2010). We calculated scores for these prototype pMMIs by averaging metric scores and rescaling by the mean of reference calibration sites, which allows comparisons among prototype pMMIs.

Subsequently, we ranked prototype pMMIs to identify those with the best responsiveness and precision. Biased metrics already had been eliminated from consideration, and none of the prototypes exhibited geographic bias (results not shown), so we did not use accuracy to rank prototype pMMIs. We estimated responsiveness as the *t*-statistic based on mean scores at reference and high-activity calibration sites and precision as the SD of scores from reference calibration sites. We identified the best subset of prototype pMMIs as those appearing in the top quartile for both criteria. Therefore, prototype pMMIs in the best subset possessed several desirable characteristics: ecological breadth, high responsiveness, and high precision.

We assembled the final pMMI by selecting metrics in order of their frequency in the best subset of prototype pMMIs. We added metrics in order of decreasing frequency and avoided adding metrics from the same thematic group or correlated (Pearson's $r \geq 0.7$) metrics. We excluded metrics that appeared in <⅓ of the best prototype pMMIs from the final pMMI.

Aggregation of the pMMI We calculated scores for the final pMMI by averaging metric scores and rescaling by the mean of reference calibration sites (as for prototype pMMIs). Rescaling of pMMI scores ensures that pMMI and O/E are expressed in similar scales (i.e., as a ratio of observed to reference expectations) and improves comparability of the 2 indices.

We calculated scores for a combined index (the California Stream Condition Index [CSCI]) by averaging pMMI and O/E scores. We calculated a null combined index by averaging null MMI and null O/E scores.

Performance evaluation Evaluation of index performance focused on accuracy, precision, responsiveness, and sensitivity (Table 4). We compared the performance of each index to that of its null counterpart. Many of our approaches to measuring performance also have been used widely in index development (e.g., Hawkins et al. 2000, 2010a, Clarke et al. 2003, Ode et al. 2008, Cao and Hawkins 2011). We scored all indices on similar scales (i.e., a minimum of 0, with a reference expectation of 1), so no adjustments were required to make comparisons (Herbst and Silldorff 2006, Cao and Hawkins 2011). We conducted all performance evaluations separately on calibration and validation data sets.

We regarded indices as accurate if scores at reference sites were not influenced by environmental setting or time of sampling. Precise indices were those with low variability among reference sites and among samples from repeated visits within sites. Responsive indices were those that showed large decreases in response to human activity. Sensitive indices were those that frequently found nonreference sites to be below an impairment threshold (e.g., 10th percentile of scores at reference sites).

Performance of the indices along a gradient of expected numbers of common taxa (E) The performance of an ideal index should not vary with E. For example, index accuracy should not be influenced by the expected richness of a site. We evaluated the accuracy, precision, and sensitivity of the indices against E by grouping sites into bins that ranged in the number of expected taxa (bin size = 4 taxa). We chose this bin size because it was the smallest number that allowed analysis of a wide range of values of E with large numbers of sites in each bin (i.e., ≥37 sites for accuracy and precision estimates and 15 sites for sensitivity estimates). We measured accuracy as the proportion of reference sites in each bin with scores ≥10th percentile of reference calibration sites. We measured precision as the SD of reference sites in each bin and sensitivity as the

Table 4. Summary of performance evaluations. SD = standard deviation.

Aspect	Description	Indication of good performance
Accuracy and bias	Scores are minimally influenced by natural gradients	• Approximately 90% of validation reference sites have scores >10th percentile of calibration reference sites • Landscape-scale natural gradients explain little variability in scores at reference sites, as indicated by a low pseudo-R^2 for a 500-tree random-forest model • No visual relationship evident in plots of scores at reference sites against field measurements of natural gradients
Precision	Scores are similar when measured under similar settings	• Low SD of scores among reference sites (1 sample/site) • Low pooled SD of scores among samples at reference sites with multiple sampling events
Responsiveness	Scores change in response to human activity gradients	• Large t-statistic in comparison of mean scores at reference and high-activity sites • Landscape-scale human activity gradients explain variability in scores, as indicated by a high pseudo-R^2 for a 500-tree random-forest model
Sensitivity	Scores indicate poor condition at high-activity sites	• High percentage of high-activity sites have scores <10th percentile of calibration reference sites

proportion of high-activity sites within each bin with scores <10th percentile of reference calibration sites. We repeated all analyses with scores from indices based on null models.

Unlike accuracy and precision, the sensitivity of an ideal index (if measured as described above) may vary with E, but only to the extent that stress levels vary with E. However, how stress levels truly varied with E is unknown because human activity gradients were used to approximate stressor gradients, and direct, quantitative measures of stress levels are not possible. Even direct measures of water chemistry or habitat-related variables are at best incomplete estimates of the stress experienced by stream communities, and these data were not available for many sites in our data set. Therefore, we supplemented analyses of sensitivity against E by evaluating the difference in sensitivity between the pMMI and O/E against E. We calculated the difference as the adjusted Wald interval for a difference in proportions with matched pairs (Agresti and Min 2005) with the *PropCIs* package in R (Scherer 2013). The difference between the indices should be constant if E has no influence on sensitivity, or if E affects both indices in the same way. In the absence of direct measures of stress levels, these analyses provide a good measure of the influence of E on index sensitivity.

Establishment of biological condition classes, and application to a statewide assessment

We created 4 condition classes based on the distribution of scores at reference calibration sites, with a recommended interpretation for each condition class: likely to be intact (>30th percentile of reference calibration site CSCI scores), possibly altered (10th–30th percentiles), likely to be altered (1st–10th percentile), and very likely to be altered (<1st percentile). We used the *qnorm*() function in R to estimate thresholds from the observed mean and SD of reference calibration site CSCI scores. We explored other approaches to setting thresholds, such as varying thresholds by ecoregion or setting thresholds from environmentally similar reference sites, but rejected these approaches because of their added complexity and minimal benefits (Appendix S1).

We applied thresholds to a subset of sites from probabilistic surveys (n = 1318 sites) to provide weighted estimates of stream condition in California and for each major region. We also used the thresholds to make unweighted estimates of reference, moderate-activity, and high-activity sites for each region of the state. We used unweighted estimates because few reference probabilistic samples were available in certain regions. For weighted estimates, we calculated site weights by dividing total stream length in each stratum by the number of sampled sites in that stratum (these strata were defined as the intersections of strata from each contributing survey). All weight calculations were conducted using the *spsurvey* package (Kincaid and Olsen 2013) in R (version 2.15.2). We used site weights to estimate regional distributions for environmental variables using the Horvitz–Thompson estimator (Horvitz and Thomson 1952). Confidence intervals for estimates of the proportion of California's stream length meeting reference criteria were based on local neighborhood variance estimators (Stevens and Olsen 2004).

RESULTS

Biological and environmental diversity of California

Biological assemblages varied markedly across natural gradients in California, as indicated by cluster analysis. We identified 11 groups that contained 13 to 61 sites (Fig. 3). A few of these groups were geographically restricted, but most were distributed across many regions of the state. For example, sites in group 10 were concentrated in the Transverse Ranges of southern California, and sites in group 7 were entirely within the Sierra Nevada. In contrast, sites in groups 1 and 4 were broadly distributed across the northern ⅔ of California.

Environmental factors differed among several groups. Groups 8 through 11, all in the southern portions of the state, were generally drier and hotter than other groups, whereas groups 1 through 5, predominantly in mountainous and northern regions, were relatively wet and cold. Expected number of taxa also varied across groups. For example, the highest median E (i.e., sum of capture probabilities > 0.5) (17.2) was observed in group 3, whereas the lowest (7) was observed in group 8. The median E was <10 for 3 of the 11 groups (groups 8, 10, and 11). Sites in low-E groups were preponderantly (but not exclusively) in the southern portions of the state.

Development of predictive models

Predicting the number of locally common taxa for the O/E index The random-forest model selected to predict assemblage composition used 5 predictors: latitude, elevation, watershed area, mean annual precipitation, and mean annual air temperature (Table 1). The model explained 74 and 64% of the variation in O at calibration and validation sites, respectively. Regression slopes (1.05 and 0.99 at calibration and validation sites, respectively) and intercepts (–0.36 and 0.52) were similar to those expected from unbiased predictions (i.e., slope = 1 and intercept = 0, p > 0.05). The random-forest model was modestly more precise (SD = 0.19) than the null model (SD = 0.21) but substantially less precise than the best model possible (SD = 0.13).

Predicting metric values and developing the pMMI Predictive models explained >20% of variance in 17 of the 48 metrics evaluated for inclusion in the pMMI (a subset of which are shown in Table 3). For 10 metrics, ≥30% of the variance was explained, and for 2 metrics (no. intolerant taxa and % intolerant taxa), >50% of the variance was explained. Squared correlation coefficients (r^2) between predicted and observed metric values ranged from near 0 (e.g., Simpson diversity) to >0.5 (no. and % intolerant taxa metrics). Results for validation reference sites were consistent with results for calibration sites, but r^2 values differed markedly between calibration and validation data sets for some metrics (Table 3). In general, models explained the most variance for %-taxa metrics, and the least for %-abundance metrics, but this pattern was not consistent for all groups of metrics.

Metrics selected for the pMMI Of the 48 metrics evaluated, 15 met all acceptability criteria (Table 3). The bias criterion was the most restrictive and eliminated 21 metrics, including all raw metrics and 2 modeled metrics (% climber taxa and % predators). The discrimination criterion eliminated 15 metrics, most of which were already eliminated by the bias criterion. Other criteria eliminated few metrics, all of which were already rejected by other criteria. The 15 acceptable metrics yielded 28,886 possible prototype pMMIs ranging in size from 5 to 10 metrics, but only 234 prototype pMMIs contained uncorrelated metrics or metrics belonging to unique metric groups (data not shown). All of these prototype pMMIs contained ≤7 metrics. Of these 234 prototypes, only 6 were in the top quartile for both discrimination between reference and high-activity calibration samples and for lowest SDs among reference calibration samples.

The final pMMI included 1 metric from each of 6 metric groups (Table 3). Some of the selected metrics (e.g., Coleoptera % taxa) were similar to those used in regional indices previously developed in California (e.g., Ode et al. 2005). However, other widely used metrics (e.g., noninsect metrics) were not selected because they were highly correlated with other metrics that had better performance (pairwise correlations not shown).

The random-forest models varied in how much of the variation in the 6 individual metrics they explained (Pseudo-R^2 range: 0.23–0.39). Regressions of observed on predicted values for reference calibration data showed that several intercepts were significantly different from 0 and slopes were significantly different from 1 (i.e., $p < 0.05$), but these differences were small. The number of predictors used in each of the 6 models ranged from 2 (for no. Coleoptera

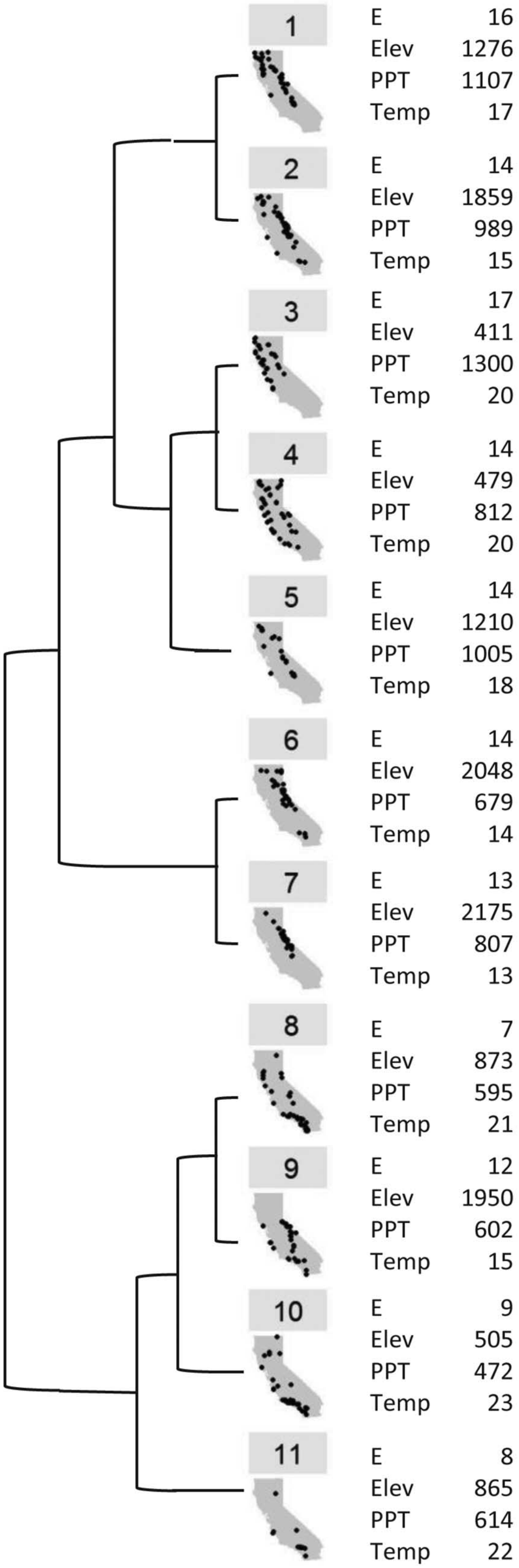

Figure 3. Dendrogram and geographic distribution of each group identified during cluster analysis. Numbers next to leaves are median values for expected number of taxa (E), elevation (Elev, m), precipitation (PPT, mm), and air temperature (Temp, °C).

taxa) to 10 (for taxonomic richness) (Table 1). Predictors related to location (e.g., latitude, elevation) were widely used, with latitude appearing in every model. In contrast, predictors related to geology (e.g., soil erodibility) or catchment morphology (e.g., watershed area) were used less often. In general, the most frequently used predictors also had the highest importance in the predictive models, as measured by % increase in mean square error. The least frequently used predictor (i.e., % P geology) was used in 1 model (taxonomic richness).

Performance of predictive models

Effects of predictive modeling on metrics For most metrics, reducing the influence of natural gradients through predictive modeling reduced the calculated difference between high-activity and reference sites, a result suggesting that stressor and natural gradients can have similar and confounded effects on many metric values (Table 3). For example, for 27 of the 48 metrics evaluated, the absolute *t*-statistic was much higher (difference in $|t| > 1$) for the raw metric than for the residuals. In contrast, the absolute *t*-statistic for residuals was higher for only 12 metrics.

Performance evaluation of the O/E, pMMI, and combined indices By all measures, predictive indices (whether used alone or combined) performed better than their null counterparts, particularly with respect to accuracy/bias (Table 5). For example, mean regional differences in null index scores at reference sites were large and significant (Fig. 4A, C, E), and responses to natural gradients were strong (Fig. 5A–O). In contrast, all measures of biases were greatly reduced for predictive indices (Fig. 4B, D, F).

Predictive modeling improved several aspects of precision. Variability of scores among reference sites was lower for all predictive indices than for their null counterparts, particularly for the pMMI (Table 5). Regional differences in precision were larger for the pMMI than O/E (both predictive and null models), and combining these 2 indices into the CSCI improved regional consistency in precision (Fig. 4B, D, F). Predictive modeling had a negligible effect on within-site variability (Table 5).

In contrast to precision and accuracy, responsiveness was more affected by index type than whether predictive or null models were used. Both predictive and null MMIs appeared to be slightly more responsive than the combined indices, which in turn were more responsive than O/E indices. This pattern was evident in all measures of responsiveness, such as magnitude of *t*-statistics, variance explained by multiple human-activity gradients in a random-forest model, and steepness of slopes against individual gradients (Table 5, Fig. 6A–I).

Analysis of sensitivity indicated stronger sensitivity of the pMMI than the O/E, and the combined index had intermediate sensitivity. Overall, 47% of nonreference sites had scores <10th percentile of reference calibration sites for the CSCI, in contrast with 52% of the pMMI and 35% of the O/E. Despite the overall difference between the pMMI and the O/E, agreement was relatively high (76%) when the 10th percentile was used as an impairment threshold (i.e., O/E ≥ 0.76 and pMMI ≥ 0.77). When the 1st percentile was used to set thresholds (i.e., O/E ≥ 0.56 and pMMI ≥ 0.58), the agreement rate was 90%.

Table 5. Performance measures to evaluate California State Condition Index (CSCI), MMI = multimetric index, and observed (O)/expected (E) taxa index at calibration (Cal) and validation (Val) sites. For accuracy tests, only reference sites were used. Ref mean = mean score of reference sites (* indicates value is mathematically fixed at 1), *F* = *F*-statistic for differences in scores at calibration sites among 5 regions (shown in Fig. 1, Central Valley excluded; residual df = 467), Var = variance in index scores explained by natural gradients at reference sites, among sites = standard deviation of scores at reference sites, within sites = standard deviation of within-site residuals for reference Cal (*n* = 220 sites) and Val (*n* = 60) sites with multiple samples, *t* = *t*-statistic for difference between mean scores at reference and high-activity sites, var = variance in index scores explained by human-activity gradients at all sites.

		Accuracy						Precision				Responsiveness			
		Ref mean		*F*		Var		Among sites		Within sites		*t*		Var	
Index	Type	Cal	Val	Cal	Val	Cal	Val	Cal	Val	Cal	Val	Cal	Val	Cal	Val
CSCI	Predictive	1.01	1.01	1.3	1.4	−0.08	−0.13	0.16	0.17	0.11	0.1	28.5	13	0.49	0.42
	Null	1*	1	52.9	4.7	0.41	0.12	0.21	0.2	0.11	0.11	28.6	14.8	0.64	0.58
MMI	Predictive	1*	0.98	0.8	1.3	−0.15	−0.09	0.18	0.19	0.12	0.12	30.9	14.4	0.54	0.48
	Null	1*	1	62.2	8.7	0.46	0.2	0.24	0.24	0.12	0.12	29.2	15.3	0.67	0.61
O/E	Predictive	1.02	1.03	1.2	1	0.01	−0.12	0.19	0.2	0.16	0.13	21.0	9.3	0.31	0.25
	Null	1*	1	23.5	0.9	0.23	−0.03	0.21	0.22	0.15	0.13	24.1	11.8	0.48	0.41

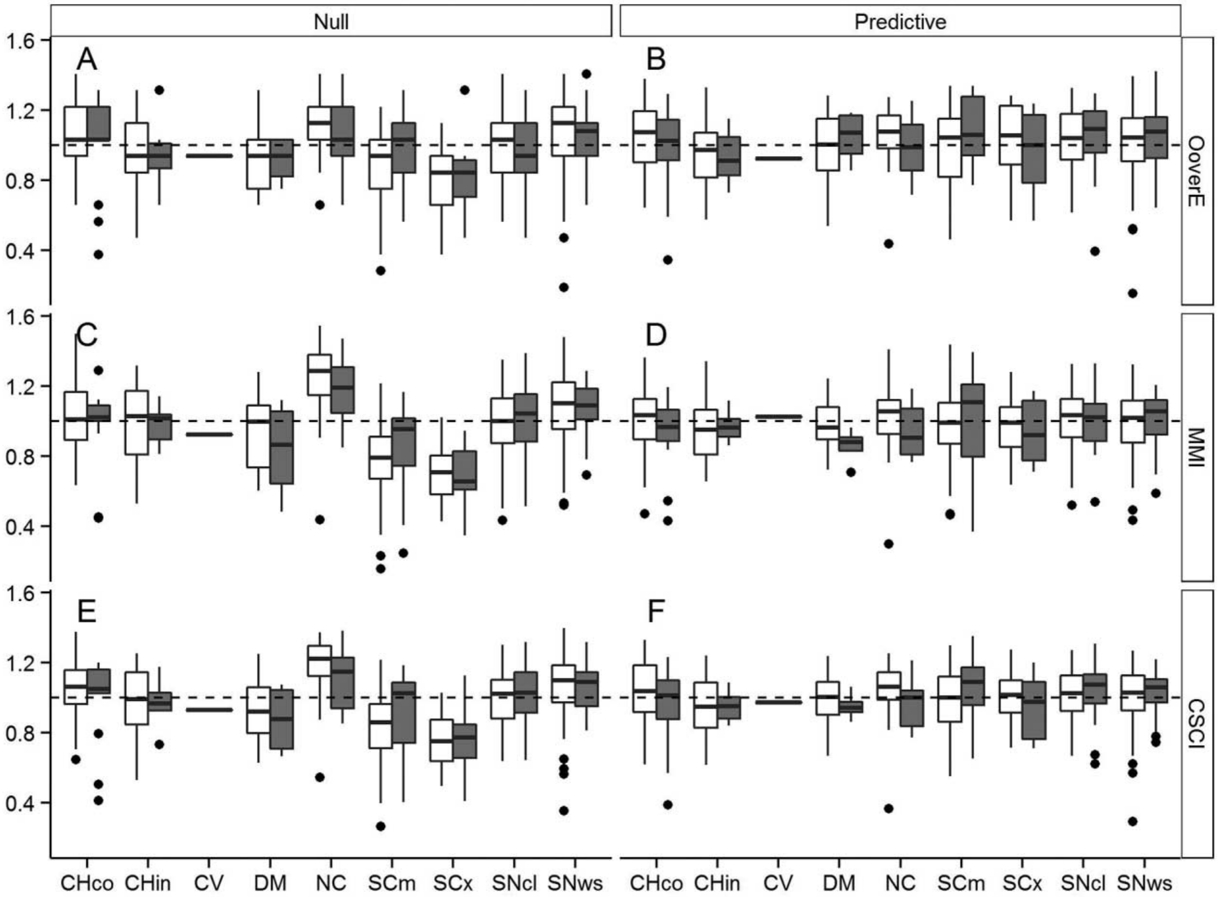

Figure 4. Box-and-whisker plots for distribution of scores for null (A, C, E) and predictive (B, D, F) models for the observed (O)/expected (E) taxon index (A, B), multimetric index (MMI) (C, D), and the combined index (CSCI) (E, F) scores by geographic region (see Fig. 1 for codes). White boxes indicate scores at calibration sites, and gray boxes indicate scores at validation sites. The horizontal dashed lines indicate the expected value at reference sites (= 1). Lines in boxes are medians, box ends are quartiles, whiskers are 1.5× the interquartile range, and dots are outliers (i.e., values >1.5× the interquartile range).

Effect of E on performance By most measures, performance was better at high-E than at low-E sites, but predictive indices were much more consistent than their null equivalents. For example, the accuracy of null indices was very poor at low-E sites (0.46–0.54 at E = 5; Fig. 7A), whereas predictive indices were much more accurate (0.73–0.86 at E = 5; Fig. 7E. At high-E sites, accuracy was >0.90 for both predictive and null indices. Precision was better at high-E sites for the pMMI and O/E index, but the CSCI had better and more consistent precision than the other indices at all values of E (Fig. 7B, F). For example, precision ranged from 0.22 to 0.15 (range = 0.07) for both the pMMI and the O/E, whereas it ranged from 0.18 to 0.14 (range = 0.04) for the CSCI.

In contrast to the weak associations between E and accuracy and precision, E was very strongly associated with sensitivity, as measured by the percentage of high-activity sites with scores <10th percentile threshold (Fig. 7C, G). The pMMI classified a larger proportion of sites as in nonreference condition across nearly all values of E than the O/E index did, but the difference was largest at low-E sites (Fig. 7D, H). For example, at the lowest values of E analyzed (5), the pMMI identified 87% of high-activity sites as biologically different from reference, whereas O/E identified only 47% of sites as in nonreference condition. As E increased, the difference between the 2 indices in proportion of sites classified as nonreference decreased. Wald's interval test indicated significant differences between the indices for values of E up to 13. At low-E sites, the sensitivity of the CSCI was between the 2 indices, but at high-E sites, CSCI was more similar to pMMI. All 3 indices showed that low-E sites were more pervasively in nonreference condition than high-E sites, and the proportion of sites with scores <10th percentile of reference calibration sites decreased as E increased. In contrast to precision and accuracy, sensitivity was more consistent across settings for

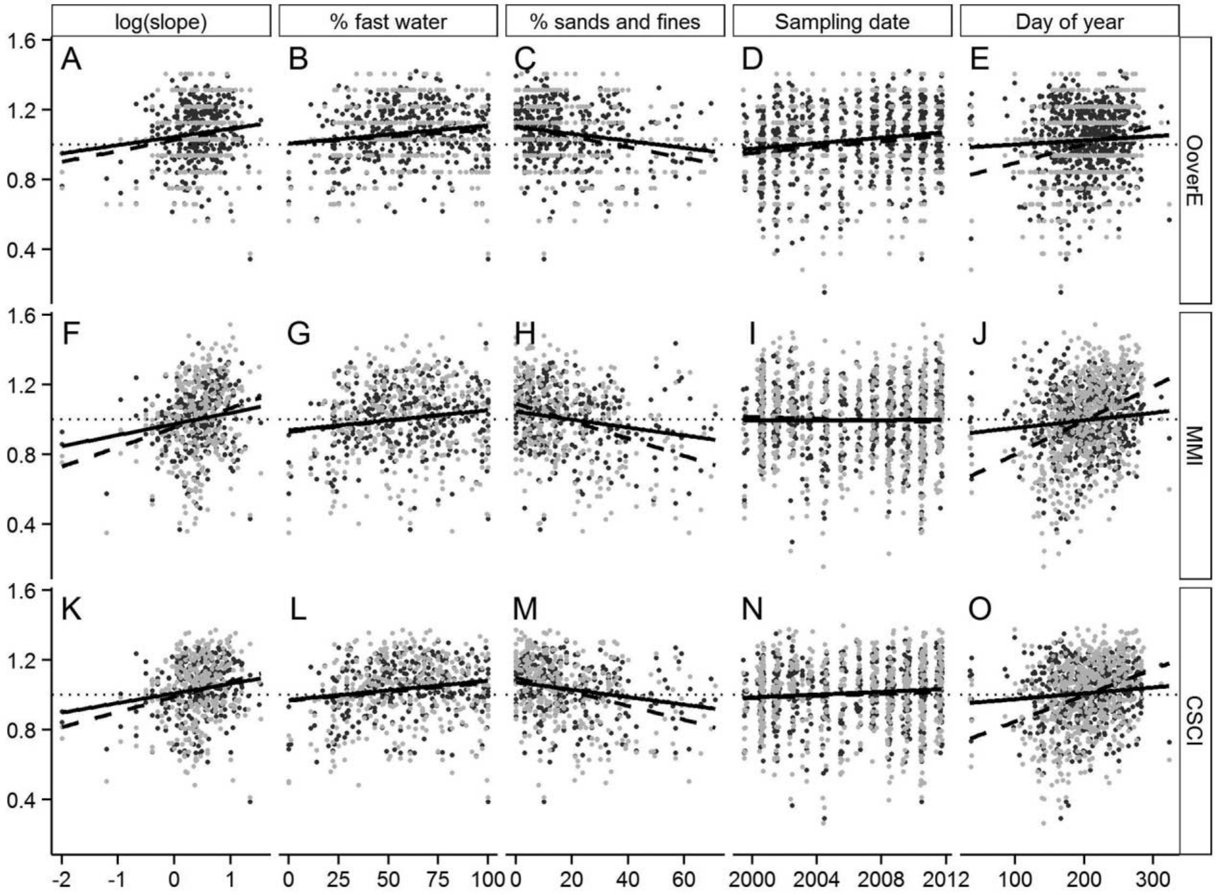

Figure 5. Relationships between observed (O)/expected (E) taxon index (A–E), multimetric index (MMI) (F–J), and the combined index (CSCI) (K–O) scores and slope (A, F, K), % fast water (area of reach with riffle, run, cascade, or rapid microhabitats) (B, G, L), % sand and fines (C, H, M), sampling date (D, I, N), and day of the year (E, J, O) at reference sites for predictive (black symbols, solid lines) and null (gray symbols, dashed lines) indices. The dotted line indicates a perfect relationship without bias.

null than predictive indices. For all analyses of performance relative to E, validation data yielded similar results (not shown).

Establishment of biological condition classes and application to a statewide assessment

We established 4 biological condition classes based on the distribution of CSCI scores at reference calibration sites. Statewide, 52% of streams were likely to be intact (i.e., CSCI ≥ 0.92 [30th percentile of reference calibration sites]). Another 18% were possibly altered (i.e., CSCI ≥ 0.79 [10th percentile]), 11% were likely to be altered (i.e., CSCI ≥ 0.63 [1st percentile]), and 19% were very likely to be altered (i.e., CSCI < 1st percentile) (Table 6). Although many (i.e., 49%) high-activity sites were very likely to be altered, this number varied considerably by region. Few high-activity sites were in this condition class in the more forested regions (e.g., 24% in the North Coast, 15% in the Sierra Nevada), whereas higher numbers were observed in relatively arid regions (e.g., 100% in the Desert/Modoc region and 68% in the Central Valley). In contrast, the percentage of reference sites in the top 2 classes varied much less across regions, from a low of ~85% in the South Coast and Desert/Modoc regions to a high of 98% in the North Coast (Table 6).

DISCUSSION

Our evaluation of index performance across different environmental settings demonstrates that, to the greatest extent possible with existing data, we have designed an index with scores that have comparable meanings for different stream types in an environmentally heterogeneous region of the USA. Each site is benchmarked against appropriate biological expectations anchored by a large and consistently defined reference data set, and deviations from these expectations reflect site condition in a consistent way across environmental settings. Thus, the index can be used to evaluate the condition of nearly all perennial streams in California, despite the region's considerable environmental and biological complexity. Three ele-

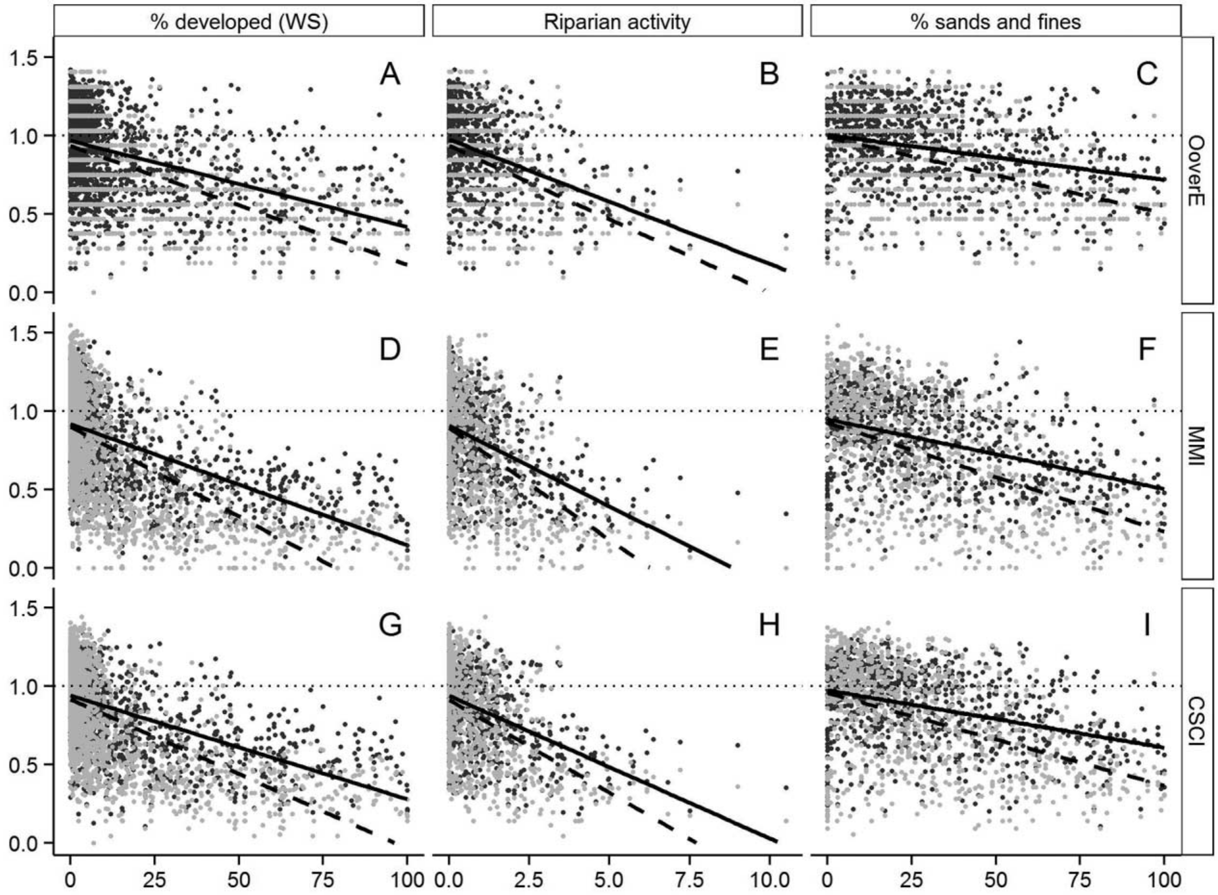

Figure 6. Relationships between observed (O)/expected (E) taxon index (A–C), multimetric index (MMI) (D–F), and the combined index (CSCI) (G–I) scores and % developed area of the watershed (WS) (A, D, G), riparian activity (B, E, H), and % sand and fines (C, F, I) for predictive (black symbols, solid lines) and null indices (gray symbols, dashed lines). The dotted line indicates the reference expectation of 1.

ments of the design process contributed to the utility of this index in an environmentally complex region: a robust reference data set, predictive modeling, and the combination of multiple endpoints into a single index.

Large, representative reference data sets

The 1st element was the large, representative, and rigorously evaluated reference data set (Ode et al. 2016). Natural factors that influence biological assemblages must be adequately accounted for to create an assessment tool that performs well across environmental settings (Cao et al. 2007, Schoolmaster et al. 2013). The strength of relationship between natural factors and biology varies with geographic scale (Mykrä et al. 2008, Ode et al. 2008), and representing locally important factors (such as unusual geology types with limited geographic extent, e.g., Campbell et al. 2009) contributes to the ability of the index to distinguish natural from anthropogenic biological variability in these environmental settings. Our reference data set was spatially representative and encompassed >10 y of sampling. Long-term temporal coverage improves the representation of climatic variability, including El Niño-related storms and droughts. The spatial and temporal breadth of sampling at reference sites provides confidence in the applicability of the CSCI for the vast majority of wadeable perennial streams in California.

Predictive modeling

The 2nd element of the CSCI's design, predictive modeling, enabled the creation of site-specific expectations for 2 indices, and these models created indices superior to those created by null models in nearly every aspect, particularly with respect to bias in certain settings. These results are consistent with a large body of literature showing similar results for indices that measure changes in taxonomic composition (e.g., Reynoldson et al. 1997, Hawkins et al. 2000, Van Sickle et al. 2005, Hawkins 2006, Mazor et al. 2006). However, few studies to date showed that the benefits extend to MMIs (e.g., Bates Prins and Smith 2007, Pont et al. 2009, Hawkins et al. 2010b, Schoolmaster et al. 2013, Vander Laan and Hawkins 2014).

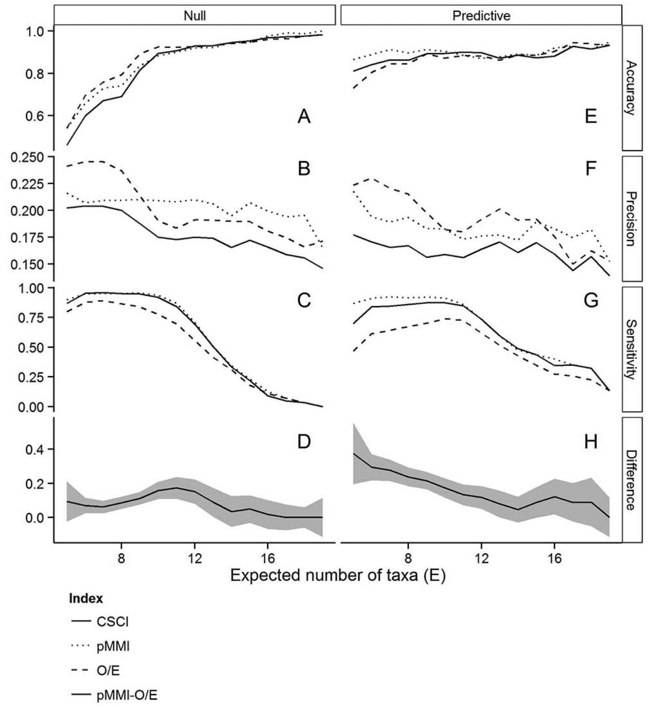

Figure 7. Effect of expected number of taxa (E) on accuracy (A, E), precision (B, F), sensitivity (C, G), and difference in sensitivity between the predictive multimetric index (pMMI) and the observed (O)/expected (E) taxa indices (D, H) for null (A–D) and predictive (E–H) index performance. The gray bands in the bottom panels C and G indicate the 95% confidence interval around the difference. Accuracy = proportion of reference calibration sites in reference condition (i.e., score >10th percentile of reference calibration sites) for each index. Precision = standard deviation of reference calibration sites for each index. Sensitivity = proportion of high-activity sites not in reference condition.

Our preference for predictive over traditional MMIs is not based only on the superior performance the pMMI relative to its null counterpart. The null MMI evaluated in our study was simplistic and did not reflect typical typological approaches to MMI development, which include regionalization in metric selection (e.g., Stoddard et al. 2008), regionalization in scoring (e.g., Ode et al. 2005), or normalization to watershed area (e.g., Klemm et al. 2003) to account for variability across reference sites. However, traditional MMIs based on regionalization usually lack metric and scoring standardization, which complicates interregional comparisons. Even if typological approaches provided equivalent performance to predictive indices, the latter would be preferred because of their ability to set site-specific management goals because predictive indices can better match the true potential of individual sites (Hawkins et al. 2010b). Thus, a watershed manager could take action to maintain a level of diversity a stream can truly support, rather than a level typical of potentially dissimilar reference sites.

Combining multiple indices

The 3rd element of the CSCI's design that contributed to its utility in different stream types was inclusion of both the pMMI and the O/E index. Regulatory agencies expressed a strong preference for a single index to support biocriteria implementation, and we thought that the CSCI was preferable to either the pMMI or O/E index. The different sensitivities of the 2 components should enhance the utility of the CSCI across a broad range of disturbances and settings. Together, they provide multiple lines of evidence about the condition of a stream and provide greater confidence in the results than a single index that might be biased in certain settings. Use of both metric and multivariate indices is widespread in assessments of coastal condition (e.g., the M-AMBI index; Muxika et al. 2007) specifically because the combination takes advantage of the unique sensitivities of each index in different habitat types (Sigovini et al. 2013). Applications of a multiple-index approach in stream assessment programs are uncommon, but the need has been suggested (e.g., Reynoldson et al. 1997, Mykrä et al. 2008, Collier 2009).

The decision to use both the pMMI and O/E index was based, at least partly, on observations that they had different sensitivities in different settings, particularly at low-E sites. The difference between the 2 indices might mean that the O/E index correctly indicates a greater resilience to stress at certain stream types or that the pMMI is more finely tuned to lower levels of stress simply because it was specifically calibrated against high-activity sites in similar settings. Mechanistically, the difference probably occurred because O/E index scores are mainly affected by the loss of common taxa. For example, in low-E sites (which were common in dry, low-elevation environments in southern and central coastal California), the O/E index predicted occurrence of only a small number of highly tolerant taxa (e.g., baetid mayflies) because only these tolerant taxa occur with high probability in these naturally stressful environments. Sensitive taxa also occur at reference sites in drier, low-elevation settings, but they were typically too rare to affect the O/E index (Appendix S2).

The interpretive value of rare, sensitive taxa in estimation of biological integrity of an individual site is unclear, but the ability of a site to support these taxa may be important to the health of a dynamic metacommunity, where rare taxa occupy only a small subset of suitable sites at any one time. Although several investigators have shown that exclusion of rare taxa usually enhances precision of O/E indices (e.g., Ostermiller and Hawkins 2004, Van Sickle et al. 2007), our results suggest that in certain settings, this exclusion may obscure an important response to

Table 6. Percentage of sites in different condition classes by region and site status. Percentiles refer to the distribution of scores at reference calibration (Cal) sites. Overall estimates are based on sites from probabilistic surveys and are not split into Cal or validation (Val) sets. For reference, moderate-, and high-activity sites, numbers in the last 6 columns are percentage of sites. For overall assessments, these numbers are percentage of stream miles. Dashes indicate that no sites were analyzed.

	Total sites		Likely to be intact ≥30th percentile (CSCI ≥ 0.92)		Possibly altered 30th–10th percentile (CSCI ≥ 0.79)		Likely to be altered 1st–10th percentile (CSCI ≥ 0.63)		Very likely to be altered <1st percentile (CSCI < 0.63)	
Region	Cal	Val	Cal	Val	Cal	Val	Cal	Val	Cal	Val
Statewide										
Reference	473	117	75	74	15	16	8	8	1	3
Moderate activity	626	156	53	56	20	20	18	17	8	7
High activity	497	122	13	18	13	14	25	22	49	46
Overall	919		52		18		11		19	
North Coast										
Reference	60	16	85	63	13	31	0	6	2	0
Moderate activity	88	26	58	50	26	15	9	27	7	8
High activity	45	9	29	67	33	33	13	0	24	0
Overall	162		58		23		10		9	
Chaparral										
Reference	74	19	68	63	20	26	9	0	3	11
Moderate activity	146	34	47	65	18	15	29	15	6	6
High activity	126	28	18	21	13	7	18	11	50	61
Overall	147		34		16		17		33	
South Coast										
Reference	96	23	70	70	16	9	14	22	1	0
Moderate activity	202	52	49	52	22	23	19	17	9	8
High activity	241	60	5	10	12	13	32	27	52	50
Overall	387		44		16		16		24	
Sierra Nevada										
Reference	221	55	77	82	14	11	7	5	1	2
Moderate activity	148	35	68	60	20	29	8	9	5	3
High activity	27	8	56	25	11	38	19	13	15	25
Overall	106		70		19		6		5	
Central Valley										
Reference	1	0	100	–	0	–	0	–	0	–
Moderate activity	8	1	0	0	0	0	38	100	63	0
High activity	47	13	0	0	4	8	28	38	68	54
Overall	60		2		8		18		71	
Desert/Modoc										
Reference	21	4	71	75	14	25	14	0	0	0
Moderate activity	34	8	44	63	9	0	29	13	18	25
High activity	5	4	0	50	0	0	0	50	100	0
Overall	57		48		14		9		30	

stress. Including rare taxa in certain environmental settings while excluding them in others may improve the consistency of an O/E index in complex regions, but we did not explore this option. The observation that sensitivity of all indices was lowest where E was highest was unexpected, and may be attributed to several potential causes. Most probably, anthropogenic stress was less severe at high-E than at low-E sites. High-activity sites were identified via indirect measures based on stressor sources (e.g., development in the watershed) rather than direct measures of water or habitat quality, so we could not ensure homogenous levels of disturbance among this set of sites. Alternatively, high-E settings might be more resilient to stress, perhaps because of their greater diversity (Lake 2000). Thus, the indices may have different responses to the same level of stress in different settings, depending on E.

Despite the lower sensitivity of the O/E index at low-E sites, we think that including it in a combined index was preferable to using the more sensitive pMMI by itself. Combining the 2 indices was a simple way to retain high sensitivity at low-E sites, while retaining the advantages of the O/E as a measure of biodiversity (Moss et al. 1987, Hawkins et al. 2000). The ability of the O/E index to measure taxonomic completeness has direct applications to conservation of biodiversity and makes it particularly sensitive to replacement of native fauna by invasive species. Furthermore, because it is calibrated with only reference sites, the O/E index is not influenced by the distribution or quality of high-activity sites. In contrast, we used the pMMI under the assumption that the set of high-activity sites adequately represented the types of stressors that might be encountered in the future. Inclusion of the O/E index in the CSCI provides a degree of insurance against faulty assumptions about the suitability of the high-activity site set for pMMI calibration.

We combined the 2 indices as an unweighted mean for several technical reasons, but primarily because this was the simplest approach to take without stronger support for more complicated methods. As we demonstrated, the CSCI has less variable performance across stream types than its 2 components. Approaches that let the lowest (or highest) score prevail are more appropriate when the components have similar sensitivity, but in our case would be tantamount to using the pMMI alone and muting the influence of the O/E index. Approaches that weight the 2 components based on site-specific factors (e.g., weighting the pMMI more heavily than the O/E index at low-E sites) are worthy of future exploration. Evaluating the pMMI and O/E indices independently to assess biological condition at a site might be useful, particularly at low-E sites, but the combined index is preferred for applications where statewide consistency is important, such as designation of impaired waterbodies.

Unexplained variability

In our study, predictive models were able to explain only a portion of the variability observed at reference sites—sometimes a fairly small portion. For example, the SD of the predictive O/E was only slightly lower than the SD of the null O/E (0.19 vs 0.21) and much larger than that associated with replicate samples (0.13). None of the selected random-forest models explained >39% (for the no. shredder taxa metric) of the variability at reference calibration sites. The unexplained variability may be related to the additional effects of environmental factors that are unsuitable for predicting reference condition (e.g., alterable factors, like substrate composition or canopy cover), environmental factors unrelated to those used for modeling (e.g., temporal gradients, weather antecedent to sampling), field and laboratory sampling error, metacommunity dynamics (Leibold et al. 2004, Heino 2013), or neutral processes in community assembly that are inherently unpredictable (Hubbell 2001, Rader et al. 2012). The relative contribution of these factors is likely to be a fruitful area of bioassessment research. Given the number and breadth of environmental gradients evaluated for modeling, we think it unlikely that additional data or advanced statistical methods will change the performance of these indices.

Setting thresholds

Some investigators have suggested that thresholds for identifying impairment in environmentally complex regions may require different thresholds in different settings based on the variability of reference streams in each setting. For example, Yuan et al. (2008) proposed ecoregional thresholds for an O/E index for the USA based on the observation that index scores at reference sites were twice as variable in some ecoregions as in others. Alternatively, site-specific thresholds could be established based on the variability of a subset of environmentally similar reference sites. We rejected both of these approaches in favor of uniform thresholds based on the variability of all reference calibration sites. We rejected ecoregional thresholds or other typological approaches because the validity of ecoregional classifications may be questionable for sites near boundaries. We rejected site-specific thresholds based on environmentally similar reference sites because they did not improve accuracy or sensitivity relative to a single statewide threshold when predictive indices are used (Appendix S1). These results are consistent with those of Linke et al. (2005), who showed that indices calibrated with environmentally similar reference sites had similar performance to indices based on predictive models that were calibrated with all available reference sites. Other approaches, such as direct modeling of the SD of index scores as a function of natural factors, also might improve comparability of scores across settings (R. Bailey, Cape Breton University, personal communication).

Conclusions and recommended applications

Many recent technical advances in bioassessment have centered on improving the performance of tools used to score the ecological condition of water bodies. Much of the progress in this area has come from regional, national, and international efforts to produce overall condition assessments of streams in particular regions (e.g., Simpson and Norris 2000, Van Sickle et al. 2005, Hawkins 2006, Hering et al. 2006, Stoddard et al. 2006, Paulsen et al. 2008). A key challenge in completing these projects has been incompatibility among scoring tools designed to assess streams in multiple regions, each calibrated for unique and locally important environmental gradients (Cao and Hawkins 2011). This issue has been well documented for large-scale programs in which investigators have attempted to integrate scores from a patchwork of assessment tools built for smaller subregions (Heinz Center 2002, Hawkins 2006, Meador et al. 2008, Pont et al. 2009), but far less attention has been paid to the meaning of index scores at individual stream reaches (Herlihy et al. 2008, Ode et al. 2008). Assessment of CSCI performance across the range of environmental settings in California was essential because the CSCI is intended for use in regulatory applications that affect the management of individual reaches, and consistent meaning of a score was a key requirement of regulatory agencies and stakeholders. We attempted to maximize consistency of the CSCI by using a large and representative reference set and by integrating multiple indices based on predictive models. Consistent accuracy was attained through the use of predictive models, whereas the consistency of precision and sensitivity was improved through the use of multiple endpoints.

The CSCI was designed for condition assessments, but we think it has broad application to many aspects of stream management. For example, it could be used to select comparator sites with similar biological expectations to test sites for use in causal assessments (e.g., CADDIS; USEPA 2010) or to prioritize streams that can support rare or threatened assemblages for restoration or conservation (Linke et al. 2011). The predictions generated by the index can inform management decisions about streams for which no biological data are available. Predictive indices, such as the CSCI, are powerful additions to the stream manager's tool kit, especially in environmentally complex areas. We recognize the challenges in enabling the general public to calculate an index as complex as the one presented here. Fortunately, online automation of many of the steps is possible. For example, much of the GIS analysis can be simplified by using publicly available resources like StreamStats (US Geological Survey 2012). An automated tool is in development, but people who are interested in using the CSCI or examining its component models are encouraged to contact the authors.

ACKNOWLEDGEMENTS

We thank Dave Buchwalter, Rick Hafele, Chris Konrad, LeRoy Poff, John Van Sickle, Lester Yuan, Jason May, Larry Brown, Ian Waite, Kevin Lunde, Marco Sigala, Joseph Furnish, and Betty Fetscher for contributions to the approach described in this paper. Financial support was provided by grants from the US EPA Region IX and the California State Water Resources Control Board. Most of the data used in this analysis were provided by the California Surface Water Ambient Monitoring Program, the California Department of Fish and Wildlife, the Sierra Nevada Aquatic Research Lab, the Stormwater Monitoring Coalition of Southern California, the US Forest Service, and the Geographic Information Center at California State University Chico. We thank Kevin Collier, Bob Bailey, and 1 anonymous referee for their valuable feedback on the manuscript. This paper is dedicated to the memory of Miriam D. Mazor.

LITERATURE CITED

Agresti, A., and Y. Min. 2005. Simple improved confidence intervals for comparing matched proportions. Statistics in Medicine 24:729–740.

Angradi, T. R., M. S. Pearson, D. W. Bolgrein, T. M. Jicha, D. L. Taylor, and B. H. Hill. 2009. Multimetric macroinvertebrate indices for mid-continent US great rivers. Journal of the North American Benthological Society 28:785–804.

Bates Prins, S. C., and E. Smith. 2007. Using biological metrics to score and evaluate sites: a nearest-neighbour reference condition approach. Freshwater Biology 52:98–111.

Blocksom, K. A. 2003. A performance comparison of metric scoring methods for a multimetric index for Mid-Atlantic highland streams. Environmental Management 42:954–965.

California State Water Resources Control Board. 2013. Regional Water Quality Control Board boundaries. California State Water Resources Control Board, Sacramento, California. (Available from: http://www.waterboards.ca.gov/waterboards_map.shtml)

CAMLnet. 2003. List of California macroinvertebrate taxa and standard taxonomic effort. California Department of Fish and Game, Rancho Cordova, California. (Available from: www.safit.org)

Campbell, R. H., T. H. McCulloh, and J. G. Vedder. 2009. The Miocene Topanga group of southern California—a 100-year history of changes in stratigraphic nomenclature. Open-File Report 2007-1285. Version 1.1. US Geological Survey, Reston, Virginia.

Cao, Y., and C. P. Hawkins. 2011. The comparability of bioassessments: a review of conceptual and methodological issues. Journal of the North American Benthological Society 30:680–701.

Cao, Y., C. P. Hawkins, J. Olson, and M. A. 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, K. R., and R. H. Green. 1988. Statistical design and analysis for a "biological effects" study. Marine Ecology Progress Series 46:213–226.

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.

Collier, K. J. 2009. Linking multimetric and multivariate approaches to assess the ecological condition of streams. Environmental Monitoring and Assessment 157:113–124.

Council of European Communities. 2000. Establishing a framework for community action in the field of water policy. Directive 2000/60/EC. Official Journal of European Communities. L327(43):1–72.

Cutler, D. R., T. C. Edwards, K. H. Beard, A. Cutler, and K. T. Hess. 2007. Random forests for classification in ecology. Ecology 88:2783–2792.

Davis, W. S., and T. P. Simon. 1995. Biological assessment and criteria: tools for water resource planning and decision making. Lewis Press, Boca Raton, Florida.

Fore, L. S., and C. Grafe. 2002. Using diatoms to assess the biological condition of large rivers in Idaho (U.S.A.). Freshwater Biology 47:2014–2037.

Gauch, H. G. 1982. Multivariate analysis in community ecology. Cambridge University Press, Cambridge, UK.

Gerritsen, J. 1995. Additive biological indices for resource management. Journal of the North American Benthological Society 14:451–457.

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: its utility in regional and global assessments. Ecological Applications 16:1277–1294.

Hawkins, C. P., Y. Cao, and B. Roper. 2010a. Method of predicting reference condition biota affects the performance and interpretation of ecological indices. Freshwater Biology 55:1066–1085.

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

Hawkins, C. P., J. R. Olson, and R. A. Hill. 2010b. The reference condition: predicting benchmarks for ecological and water-quality assessments. Journal of the North American Benthological Society 29:312–343.

Heino, J. 2013. The importance of metacommunity ecology for environmental assessment research in the freshwater realm. Biological Reviews 88:166–178.

Heinz Center (H. John Heinz III Center for Science and the Environment). 2002. The state of the nation’s ecosystems: measuring the lands, waters, and living resources of the United States. Cambridge University Press, New York.

Herbst, D. B., and E. L. Silldorff. 2006. Comparison of the performance of different bioassessment methods: similar evaluations of biotic integrity from separate programs and procedures. Journal of the North American Benthological Society 25:513–530.

Herbst, D. B., and E. L. Silldorff. 2009. Development of a benthic macroinvertebrate index of biological integrity (IBI) for stream assessments in the Eastern Sierra Nevada of California. Sierra Nevada Aquatic Research Lab, Mammoth Lakes, California. (Available from: http://www.waterboards.ca.gov/lahontan/water_issues/programs/swamp/docs/east_sierra_rpt.pdf)

Hering, D., R. K. Johnson, S. Kramm, S. Schmutz, K. Szoszkiewicz, and P. F. Verdonschot. 2006. Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: a comparative metric-based analysis of organism response to stress. Freshwater Biology 51:1757–1785.

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

Horvitz, D. G., and D. J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47:663–685.

Hubbell, S. P. 2001. The unified neutral theory of biodiversity. Monographs in Population Biology 32:1–392.

Hughes, R. M., D. P. Larsen, and J. M. Omernik. 1986. Regional reference sites: a method for assessing stream potentials. Environmental Management 10:629–635.

Hughes, R. M., P. R. Kauffmann, 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.

Kaufmann, P. R., P. Levine, E. G. Robinson, C. Seeliger, and D. V. Peck. 1999. Surface waters: quantifying physical habitat in wadeable streams. EPA/620/R-99/003. Office of Research and Development, US Environmental Protection Agency, Washington, DC.

Kincaid, T., and A. Olsen. 2013. Spsurvey: spatial survey design and analysis. R package version 105. R Project for Statistical Computing, Vienna, Austria.

Klemm, D. J., K. A. Blocksom, F. A. Fulk, A. T. Herlihy, R. M. Hughes, P. R. Kaufmann, D. V. Peck, J. L. Stoddard, and W. T. Thoeny. 2003. Development and evaluation of a macroinvertebrate biotic integrity index (MBII) for regionally assessing Mid-Atlantic Highlands streams. Environmental Management 31:656–669.

Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, and A. Engelhardt. 2012. caret: classification and regression training. R package, version 5.15-045. R Project for Statistical Computing, Vienna, Austria. (Available from: http://www.epa.gov/nheerl/arm)

Lake, P. S. 2000. Disturbance, patchiness, and diversity in streams. Journal of the North American Benthological Society 19:573–592.

Leibold, M. A., M. Holyoak, N. Mouquet, P. Amarasekare, J. M. Chase, M. F. Hoopes, R. D. Holt, J. B. Shurin, R. Law, D. Tilman, M. Loreau, and A. Gonzalez. 2004. The metacommunity concept: a framework for multi-scale community ecology. Ecology Letters 7:601–613.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R News 2:18–22.

Linke, S., R. H. Norris, D. P. Faith, and D. Stockwell. 2005. ANNA: a new prediction method for bioassessment programs. Freshwater Biology 50:147–158.

Linke, S., E. Turak, and J. Nel. 2011. Freshwater conservation planning: the case for systematic approaches. Freshwater Biology 56:6–20.

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. 2012. cluster: cluster analysis basics and extensions. R package version 1.14.3. R Project for Statistical Computing, Vienna, Austria.

Mazor, R. D., T. B. Reynoldson, D. M. Rosenberg, and V. H. Resh. 2006. Effects of biotic assemblage, classification, and assessment method on bioassessment performance. Canadian Journal of Fisheries and Aquatic Sciences 63:394–411.

Meador, M. R., T. R. Whittier, R. M. Goldstein, R. M. Hughes, and D. V. Peck. 2008. Evaluation of an index of biotic integrity approach used to assess biological condition in western US streams and rivers at varying spatial scales. Transactions of the American Fisheries Society 137:13–22.

Moss, D., T. Furse, J. F. Wright, and P. D. Armitage. 1987. The prediction of macro-invertebrate fauna of unpolluted runningwater sites in Great Britain using environmental data. Freshwater Biology 17:41–52.

Muxika, I., Á. Borja, and J. Bald. 2007. Using historical data, expert judgment and multivariate analysis in assessing reference conditions and benthic ecological status, according to the European Water Framework Directive. Marine Pollution Bulletin 55:16–29.

Mykrä, H., J. Aroviita, J. Kotanen, H. Hämäläinen, and T. Muotka. 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.

Ode, P. R., C. P. Hawkins, and R. D. Mazor. 2008. Comparability of biological assessments derived from predictive models and multimetric indices of increasing geographic scope. Journal of the North American Benthological Society 27:967–985.

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.

Ode, P. R., A. C. Rehn, R. D. Mazor, K. C. Schiff, E. D. Stein, J. T. May, L. R. Brown, D. B. Herbst, D. Gillett, K. Lunde, and C. P. Hawkins. 2016. Evaluating the adequacy of a reference-pool site for ecological assessments in environmentally complex regions. Freshwater Science 35:237–248.

Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, P. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, and H. Wagner. 2013. vegan: community ecology package. R package version 2.0-6. R Project for Statistical Computing, Vienna, Austria.

Olson, J. R., and C. P. Hawkins. 2012. Predicting natural baseflow stream water chemistry in the western United States. Water Resources Research 48:W02504.

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.

Paulsen, S. G., A. Mayio, D. V. Peck, J. L. Stoddard, E. Tarquinio, S. M. Holdsworth, J. Van Sickle, L. L. Yuan, C. P. Hawkins, A. T. Herlihy, P. R. Kaufmann, M. T. Barbour, D. P. Larsen, and A. R. Olsen. 2008. Condition of stream ecosystems in the US: an overview of the first national assessment. Journal of the North American Benthological Society 27:812–821.

Peck, D. V., A. T. Herlihy, B. H. Hill, R. M. Hughes, P. R. Kaufmann, D. J. Klemm, J. M. Lazorchak, F. H. McCormick, S. A. Peterson, S. A. 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. Office of Research and Development, US Environmental Protection Agency, Corvallis, Oregon.

Pont, D., R. M. Hughes, T. R. Whittier, and S. Schmutz. 2009. A predictive index of biotic integrity model for aquatic–vertebrate assemblages of western U.S. streams. Transactions of the American Fisheries Society 138:292–305.

Rader, R. B., M. J. Keleher, E. Billman, and R. Larsen. 2012. History, rather than contemporary processes, determines variation in macroinvertebrate diversity in artesian springs: the expansion hypothesis. Freshwater Biology 57:2475–2486.

Rehn, A. C. 2009. Benthic macroinvertebrates as indicators of biological condition below hydropower dams on West Slope Sierra Nevada streams, California, USA. River Research and Applications 25:208–228.

Rehn, A. C., P. R. Ode, and C. P. Hawkins. 2007. Comparison 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, Rancho Cordova, California. (Available from: http://www.waterboards.ca.gov/water_issues/programs/swamp/docs/reports/assess_nocal2005.pdf)

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

Reynoldson, T. B., R. H. Norris, V. H. Resh, K. E. Day, and D. M. Rosenberg. 1997. The reference condition: a comparison of multimetric and multivariate approaches to assess water-quality impairment using benthic macroinvertebrates. Journal of the North American Benthological Society 16:833–852.

Richards, A. B., and D. C. Rogers. 2011. List of freshwater macroinvertebrate taxa from California and adjacent states including standard taxonomic effort levels. Southwest Association of Freshwater Invertebrate Taxonomists, Chico, California. (Available from: www.safit.org)

Roth, N., M. Southerland, J. Chaillou, R. Klauda, P. Kayzak, S. Stranko, S. Weisberg, L. Hall, and R. Morgan. 1998. Maryland biological stream survey: development of a fish index of biotic integrity. Environmental Monitoring and Assessment 51:89–106.

Royer, T. V., C. T. Robinson, and G. W. Minshall. 2001. Development of macroinvertebrate-based index for bioassessment of Idaho rivers. Environmental Management 27:627–636.

Scherer, R. 2013. PropCIs: various confidence interval methods for proportions. R package, version 0.2-4. R Project for Statistical Computing, Vienna, Austria.

Schoolmaster, D. R., J. B. Grace, E. W. Schweiger, B. R. Mitchell, and G. R. Guntenspergen. 2013. A causal examination of the effects of confounding factors on multimetric indices. Ecological Indicators 29:411–419.

Sigovini, M., E. Keppel, and D. Tagliapietra. 2013. M-AMBI-revisited: looking inside a widely-used benthic index. Hydrobiologia 717:41–50.

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 freshwaters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, Cumbria, UK.

Sleeter, B. M., T. S. Wilson, C. E. Soulard, and J. Liu. 2011. Estimation of late twentieth century land-cover change in

California. Environmental Monitoring and Assessment 173: 251–266.

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., A. T. Herlihy, D. V. Peck, R. M. Hughes, T. R. Whittier, and E. Tarquinio. 2008. A process for creating multimetric indices for large-scale aquatic surveys. Journal of the North American Benthological Society 27:878–891.

Stoddard, J. L., D. P. Larsen, C. P., Hawkins, R. K. Johnson, and R. H. Norris. 2006. Setting expectations for the ecological condition of streams: the concept of reference condition. Ecological Applications 16:1267–1276.

Stribling, J. B., B. K. Jessup, and D. L. Feldman. 2008. Precision of benthic macroinvertebrate indicators of stream condition in Montana. Journal of the North American Benthological Society 27:58–67.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn. 2007. Bias in random forest variable importance measures: illustrations, sources, and a solution. BMC Bioinformatics 8:25.

USEPA (US Environmental Protection Agency). 2002. Summary of biological assessment programs and biocriteria development for states, tribes, territories, and interstate commissions: streams and wadeable rivers. EPA-822-R-02-048. Office of Environmental Information and Office of Water, US Environmental Protection Agency, Washington, DC.

USEPA (US Environmental Protection Agency). 2010. Causal Analysis/Diagnosis Decision Information System (CADDIS). Office of Research and Development, US Environmental Protection Agency, Washington, DC. (Available from: http://www.epa.gov/caddis)

US Geological Survey. 2012. The StreamStats program. US Geological Survey, Reston, Virginia. (Available from: http://streamstats.usgs.gov)

Vander Laan, J. J., and C. P. Hawkins. 2014. Enhancing the performance and interpretation of freshwater biological indices: an application in arid zone streams. Ecological Indicators 36:470–482.

Van Sickle, J. 2010. Correlated metrics yield multimetric indices with inferior performance. Transactions of the American Fisheries Society 139:1802–1817.

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.

Wright, J. F., D. W. Sutcliffe, and M. T. Furse. 2000. Assessing the biological quality of freshwaters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, Cumbria, UK.

Yates, A. G., and R. C. Bailey. 2010. Selecting objectively defined reference stream sites for stream bioassessment programs. Environmental Monitoring and Assessment 170:129–140.

Yoder, C. O., and M. T. Barbour. 2009. Critical elements of state bioassessment programs: a process to evaluate program rigor and comparability. Environmental Monitoring and Assessment 150:31–42.

Yuan, L. L., C. P. Hawkins, and J. Van Sickle. 2008. Effects of regionalization decisions on an O/E index for the US national assessment. Journal of the North American Benthological Society 27:892–905.

Appendix S1. Nearest-neighbor thresholds do not improve performance of predictive indices.

Variable impairment thresholds may be useful when the precision of an index varies greatly across settings (Death and Winterbourn 1994). For example, Yuan et al. (2008) observed 2-fold differences in variability at reference sites across ecoregions in an observed (O)/expected (E) taxa index for the USA, results that justified different thresholds for each region. In such circumstances, a uniform threshold may increase the frequency of errors in the more variable settings. Reference sites with scores below a uniform threshold may be disproportionately common in settings where the index is less precise. A variable threshold that is lower in more variable settings may reduce this error rate (i.e., the reference error rate).

To determine if variable impairment thresholds based on site-specific characteristics could lead to an unbiased distribution of errors across regions, we evaluated 2 approaches to establishing thresholds: 1) a traditional approach, where a single number (based on variability in scores at all reference calibration sites) was used as a threshold, and 2) a site-specific approach, where thresholds were based on only a subset of the most environmentally similar reference calibration sites. In both cases, we considered sites to be in reference condition if their index score was >10th percentile of the relevant set of reference calibration site values. We measured environmental similarity as standard Euclidean distances along all environmental gradients used in predictive models (Table 1). We evaluated several different sizes of reference-site subsets (25, 50, 75, 100, and 200, and the full set of 473). We calculated the error rate for all regions (except for the Central Valley, which had only 1 reference site) as the proportion of sites with scores below the threshold. We plotted these regional error rates against the number of reference sites used to calculate the threshold (Fig. S1) and transformed scores at test sites into percentiles relative to each of these distributions. We used the predictive California Stream Condition Index

(CSCI) and its null equivalent in this analysis.

Variable thresholds greatly reduced the regional bias of the error rate of the null index, but had a negligible effect on the predictive index. For example, the null index had a very high error rate (0.30) in the South Coast when a uniform threshold was used, but this error rate dropped to 0.10 when variable thresholds based on 25 or 50 reference sites were used. In contrast, the regional error rate of the predictive index was always <0.15 and was not highly influenced by the number of reference sites used to establish thresholds.

We recommend a uniform threshold used in conjunction with a predictive index because of the added complexity and minimal benefits provided by the variable, site-specific thresholds.

Literature cited

Death, R. G., and M. J. Winterbourn. 1994. Environmental stability and community persistence: a multivariate perspective. Journal of the North American Benthological Society 13:125–139.

Yuan, L. L., C. P. Hawkins, and J. Van Sickle. 2008. Effects of regionalization decisions on an O/E index for the US national assessment. Journal of the North American Benthological Society 27:892–905.

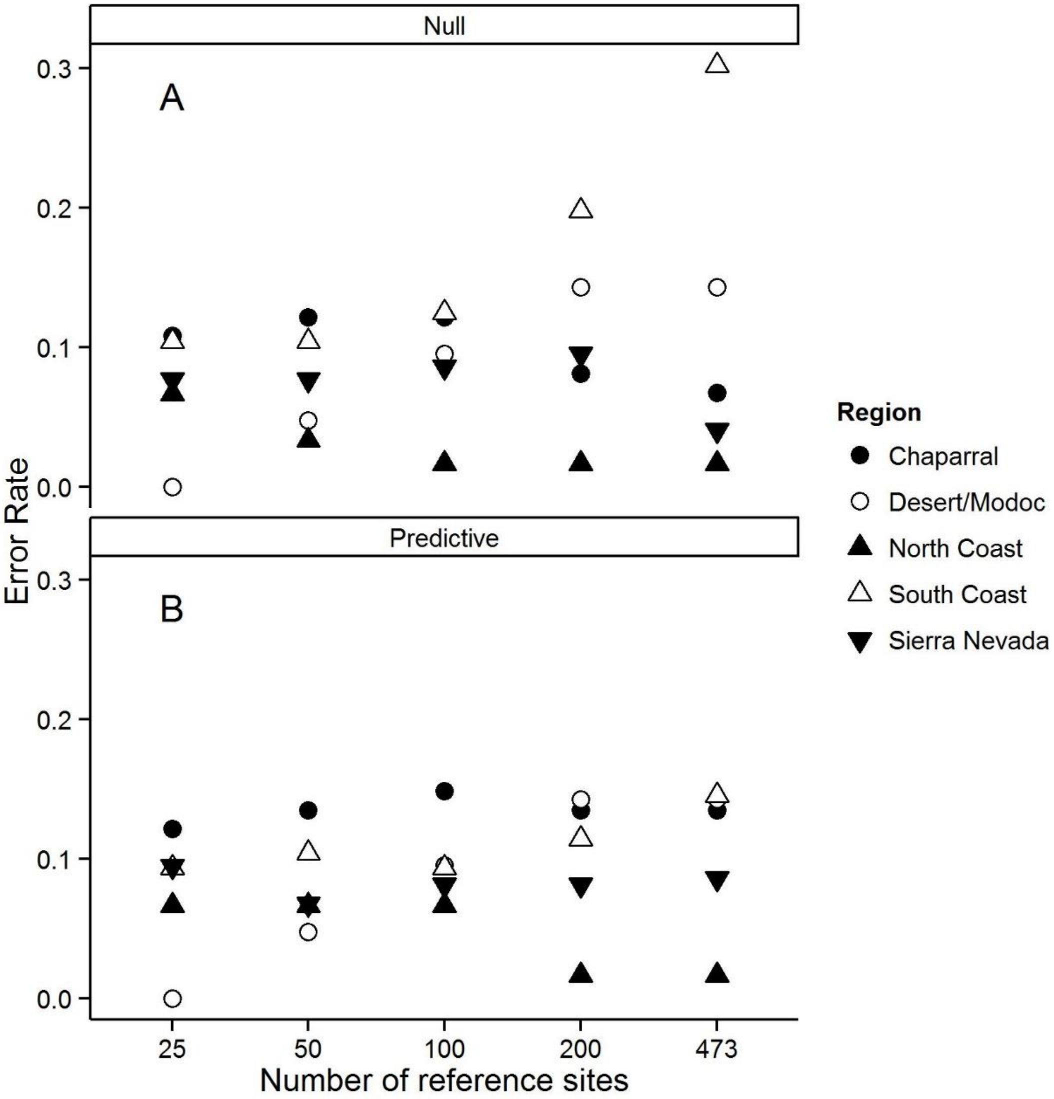

Fig. S1. Effects of nearest neighbor thresholds on error rates, calculated as the proportion of reference calibration sites below the threshold for null (A) and predictive (B) indices. Each point represents a different region. The highest number of reference sites is equivalent to the uniform threshold used in the main study.

Appendix S2. Index responsiveness as a function of predicted % sensitive taxa: a comparison of a predictive metric approach and the observed (O)/expected (E) taxa index.

The responsiveness of a bioassessment index depends on its ability to change in response to stress, and the loss of sensitive taxa is typically one of the strongest responses to stress (Rosenberg and Resh 1993, Statzner et al. 2004). To see if the ability to detect the loss of sensitive taxa depends on number of common taxa (E), we compared the proportion of sensitive taxa expected by an O/E index and a predictive multimetric index (pMMI) under different values of E. For the pMMI, this proportion was calculated as the predicted % intolerant taxa metric, as described in the accompanying manuscript. For the O/E, this proportion was calculated as the % of expected operational taxonomic units (OTUs) that are sensitive (OTUs with tolerance value < 3. For OTUs consisting of multiple taxa with different tolerance values, we used the median tolerance value). CAMLnet (2003) was the source of tolerance values. Estimates from both the O/E and pMMI were plotted against E to see whether the 2 indices allowed consistent ranges of response across values of E. These predictions were compared with the observed % intolerant taxa at reference sites to confirm the validity of these estimates.

At high-E sites (E > 14), both the pMMI and O/E had a consistent capacity to detect loss of sensitive taxa (Fig. S2A, C). Furthermore, both indices estimated similar proportions of sensitive taxa (~40%), suggesting that the 2 indices have similar sensitivity in these settings. Both indices also predicted a decline in the proportion of sensitive taxa at low-E sites, indicating that E affects the sensitivity of the pMMI and O/E. However, at the lowest levels of E, the O/E had no capacity to detect loss of sensitive taxa, whereas the pMMI predicted ~20% sensitive taxa at these sites, preserving a limited capacity to respond to loss of sensitive taxa. This capacity explains why the pMMI was more sensitive than the O/E at low-E sites.

Inspection of the data at reference sites indicates that sensitive taxa were truly present at these low-E sites (Fig. S2B, D) and that modeling the metric directly sets more accurate expectations for sensitive taxa in these settings (metric prediction vs observed $R^2 = 0.80$; O/E prediction vs observed $R^2 = 0.55$). However, these taxa were excluded from the index because of the minimum capture probability (i.e., 50%). Therefore, the predictive metric and not the O/E will be able respond to the loss of sensitive taxa at low-E sites.

Literature Cited

CAMLnet. 2003. List of California macroinvertebrate taxa and standard taxonomic effort. California Department of Fish and Game, Rancho Cordova, California. (Available from: www.safit.org)

Rosenberg, D. M., and V. H. Resh. 1993. Freshwater biomonitoring and benthic macroinvertebrates. Chapman and Hall, New York.

Statzner, B., S. Dolédec, and B. Hugueny. 2004. Biological trait composition of European stream invertebrate communities: assessing the effects of various trait filter types. Ecography 27:470–788.

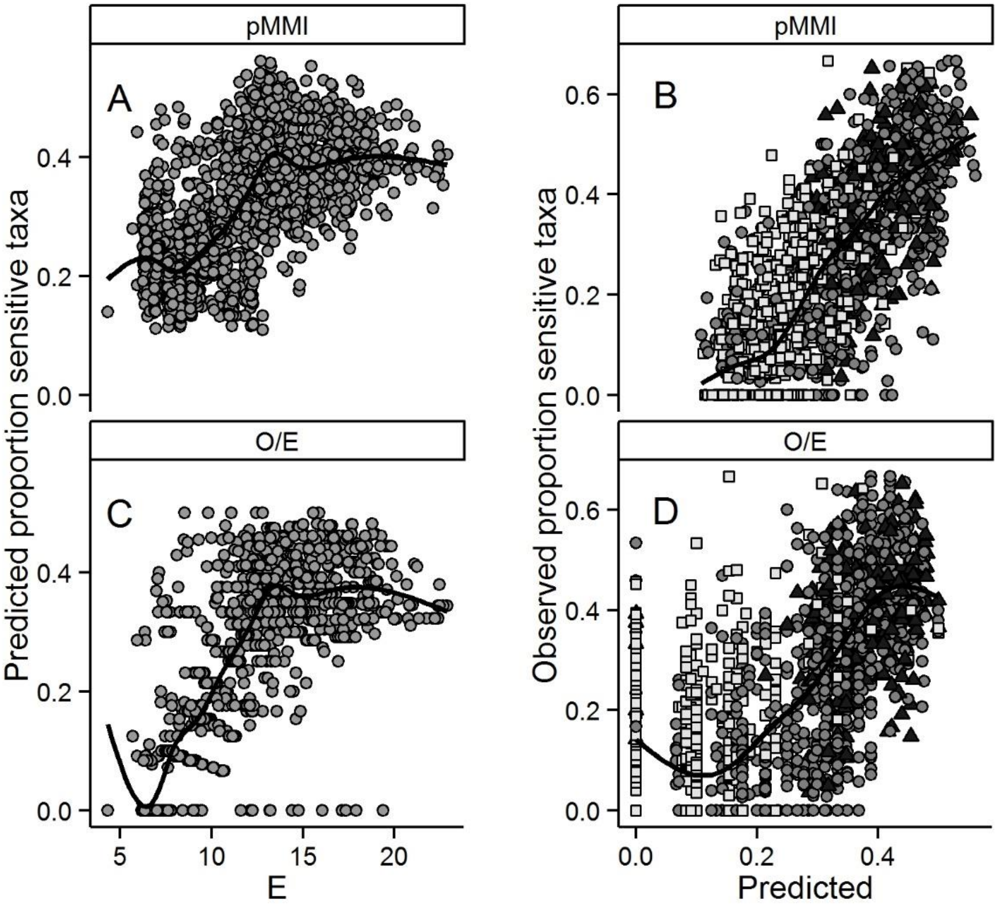

Fig. S2. Proportion of sensitive taxa predicted by a predictive multimetric index (pMMI) (A, B) and an observed (O)/expected (E) taxa index (C, D) at all sites (A, C), or observed at reference calibration (B, D) sites. Dark triangles represent sites with high (>15) numbers of expected taxa, gray circles represent sites with moderate (10–15) numbers of expected taxa, and white squares represent sites with low (<10) numbers of expected taxa. The solid line represents a smoothed fit from a generalized additive model.