Running Head

Bioassessment in complex environments

Title

Maximizing the applicability of bioassessment scoring tools in environmentally complex regions

Abstract

Regions with great natural environmental complexity present a challenge for the development of stream bioassessment indices that are effective for both regional condition assessments and site scale assessments, such as those used for regulatory applications. Tools intended for site scale applications must provide accurate biological interpretations on a site-by-site basis in all environmental settings, but also must produce consistent interpretation across the entire region of interest. In this study, we develop an assessment tool based on benthic macroinvertebrates with sufficient regional consistency and site specificity to support a wide range of applications in perennial wadeable streams in environmentally complex California, USA. The achievement of both consistency and site-specificity was supported by two key elements of our approach: 1) use of a large reference data set that represents the full range of natural settings in California; 2) development of site-specific models for predicting biological communities expected at novel test sites using an index based on the ratio of observed-to-expected taxa (O/E) and a predictive multimetric index (pMMI). We further improved performance by combinating the pMMI and O/E endpoints into a single index. Predictive models with site-specific expectations had less bias and more sensitivity than non-site-specific approaches. The sensitivity of the modeled endpoints (O/E and pMMI) depended upon environmental setting, and combining the two endpoints into a single index yielded strong consistency across settings.

Key Words

Bioassessment, Predictive modeling, Streams, Site-specificity, Predictive multimetric index, Reference Condition, Biocriteria,

Authors

R.D. Mazor, A. Rehn, P. R. Ode, M. Engeln, K. Schiff

Acknowledgements

To be completed.

Introduction

As bioassessment data become more widely used in regulatory applications such as biocriteria (Davis and Simon 1995, Council for European Communities 2000, USEPA 2002, Yoder and Barbour 2009), the need for both regional consistency and site-specific accuracy becomes increasingly important. Many bioassessment indices developed to support stream condition assessments have properly emphasized regional accuracy (e.g., Ode et al. 2005, Stoddard et al. 2008). In surveys of large areas, site-specific errors do not affect regional condition estimates as long as errors are unbiased (Herbst and Silldorff 2006, Ode et al. 2008, Yuan et al. 2008). However, when bioassessment data are used in site-specific applications like regulatory assessments, site-specific errors are more consequential and may lead to inappropriate, expensive, and ultimately unsuccessful management decisions. Therefore, bioassessment indices must accommodate both regional consistency and site specificity if used for regulatory applications (Herlihy et al. 2008).

Achieving both consistency and site specificity is particularly challenging in environmentally complex regions (Hughes 1995, Yuan et al. 2008, Pont et al. 2009). Large natural gradients create unique environmental settings that support distinct biological communities in unaltered streams (Townsend and Hildrew 1994, Statzner et al. 2004, Poff et al. 2006). Environmental diversity can complicate interpretation of indices that ignore this source of variability. Bioassessment indices should account for different aquatic assemblages expected under undisturbed conditions so that deviations from reference conditions due to anthropogenic disturbance are not confounded by natural variability, and they should do so equally well in all settings (Schoolmaster et al. 2013).

Several multi-metric indices (MMIs) based on benthic macroinvertebrates (BMIs) have been developed for California streams, each built for different regions of the state and each differing in accuracy, precision, and sensitivity (e.g., Ode et al. 2005, Rehn 2009, Herbst and Silldorff 2009). Such regionalizations (or other typological ways of matching assessed sites to appropriate reference sites) are common to MMI development but may not adequately partition systematic and continuous variation across sites (Cao and Hawkins 2011). A regional approach, such as the one used so far in California, may lead to inconsistent tools that confound inter-regional interpretability and preclude unbiased statewide assessments (Pont 2009; Hawkins et al. 2010, Cao and Hawkins 2011). Inconsistency may lead to more regulatory actions in some regions than others because of the different assessment tools used, not because of conditions of the sites. Despite their ability to link the composition of a biological assemblage with functional measures of ecosystem health (Barbour et al. 1995, Gerritsen 1995, Collier 2009), the typological nature of typical MMIs has often been cited as one of their shortcomings (Hannaford and Resh 1995, Norris 1995, Reynoldson et al. 1997, Norris and Hawkins 2000).

Predictive modeling of the reference condition is an increasingly common way to obtain site-specific expectations for diverse environmental settings (Hawkins et al. 2010). Thus far, predictive modeling has almost exclusively been applied to multivariate endpoints that focus on taxonomic completeness of a sample , such as the ratio of observed-to-expected taxa (O/E, Wright et al. 2000), or location of sites in ordination-space (e.g., BEAST, Reynoldson et al. 1995). An O/E index for BMI assemblages was developed for mountainous regions of California in 2005 (unpublished, but see Ode et al. 2008) but was based on a limited number (209) of reference sites that poorly represented several regions of the state. Applications of predictive models to multimetric indices (MMIs) are relatively new (e.g., Cao et al. 2007, Pont et al. 2009, Vander Laan et al. in press). The ecological comprehensiveness of MMIs, which include metrics related to taxonomic diversity, life history traits, trophic groups, habits, and pollution tolerance, provides useful information about biological condition that may not be incorporated in an index based strictly on loss of taxa (Gerritsen et al. 1995).

In addition to predictive modeling, the use of multiple endpoints or site-specific thresholds may also improve site specificity in bioassessment. Not all endpoints work equally well under all settings or respond similarly to all disturbances, a primary justification for integrating metrics into multimetric indices (Karr et al. 1981, Stoddard et al. 2008, Collier 2009, Schoolmaster et al. 2012) or using multiple biological assemblages in stream assessments (Council of European Communities 2000, Hering et al. 2006, Resh 2008). Site-specific thresholds also may be an appropriate approach where large differences in variability among settings exist (Death and Winterbourne 1994). For example, Yuan et al. (2008) observed reference site standard deviations in an O/E index for the United States ranged from a low of 0.17 to a high of 0.34 by ecoregion, justifying different thresholds for each region. In such circumstances, biological expectations may be similar among settings, but variable thresholds allow larger deviations from expectations in some settings than others

Our goal in this study was to construct a scoring tool for perennial wadeable streams that achieves consistency with site-specificity across environmentally complex California, a region with nine Level III ecoregions (Omernik 1987). We accomplished this goal in a three step process. First, we constructed BMI –based predictive models for taxa loss (O/E) and multi-metric (pMMI) endpoints. Second, we compared the performance of O/E, pMMI, and a combined O/E + pMMI endpoint for accuracy, precision, and sensitivity among the variety of environmental settings in California. Third, we evaluated the utility of incorporating site-specific thresholds for interpreting the assessment index. The primary motivation for model development, endpoint comparison, and threshold evaluation was to support upcoming regulatory application of biological condition for the State of California. However, our broader goal was to produce a robust assessment tool that would support a wide variety of bioassessment applications.

Methods

California contains continental-scale environmental diversity within 424,000 km^2, encompassing some of the highest, lowest, hottest, coldest, wettest, and driest portions of the United States. It supports temperate rainforests in the North Coast, deserts in the East, and chaparral, oak woodlands, and grasslands with a Mediterranean-climate in central parts of the state (Omernik 1987). Although much of the state is protected open space, vast regions of the state has 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, recreation, and other resource extraction activities occur throughout less populated regions of the state. To facilitate data evaluation, the state was divided into six regions and ten subregions based on ecoregional (Omernik 1987) and hydrologic boundaries (Figure 1).

Aggregation of dataset

More than 20 federal, state, and regional monitoring programs were inventoried to assemble data sets for index development. A total of 4457 samples from 2352 unique sites between 1999 and 2010 were aggregated into a single database. When multiple programs sampled identical candidate sites or sites in close proximity (within 300 m), data were treated as a single site to minimize redundancy.

Biological data

Slightly more than half (55%) of BMI data were collected using the reachwide protocol of the US EPA's Environmental Monitoring and Assessment Program (EMAP, Peck et al. 2006), but the rest were collected with targeted riffle protocols (Herbst and Silldorff 2006, Rehn et al. 2007). Previous studies have documented the comparability of these protocols (Gerth and Herlihy 2006, Herbst and Silldorff 2006, Rehn et al. 2007). Approximately half of the samples were identified to Level 2 (i.e., most taxa to species, with Chironomidae to genus) in the Standardized Taxonomic Effort of the Southwest Association of Freshwater Invertebrate Taxonomists (SAFIT, Richards and Rogers 2011), and half were identified to Level 1 (most taxa to genus, with Chironomidae to family).

Because sample size and taxonomic effort varied widely within the dataset, samples were screened for modeling adequacy according to three criteria, based on the intended requirements of the two endpoints. For the MMI, 500-count samples were desired, so samples with fewer than 450 (i.e., within 10% of target) individuals were excluded. For the O/E, only unambiguously distinct taxa (i.e., taxa identified to a standard level) may be used. Therefore, operational taxonomic units (OTUs) were defined (generally, most taxa to genus, with Chironomidae to subfamily), and ambiguous taxa (e.g., an insect identified to family when the OUT specifies genus) were excluded from analyses in the development of the O/E. Because at least 400 organisms were desired for calculation of the O/E, samples with fewer than 360 individuals and fewer than 50% ambiguous taxa were excluded from analysis. Screening for requirements of both models yielded a data set of 3518 samples from 1985 sites.

Geographic data

A large number of spatial data sources were aggregated to characterize natural and anthropogenic gradients known to affect benthic communities, such as land cover, road density, hydrologic alteration, mining, geology, elevation and climate (Table 1 and 2, described in Ode et al., in review). Land cover and other measures of human activity were quantified into metrics that were calculated at three spatial scales: within the entire upstream drainage area (watershed), within 5 km upstream and within 1 km upstream. Polygons defining these spatial analysis units were created using ArcGIS tools (ESRI 2009).

Designation of reference sites and creation of data subsets

For development and evaluation of assessment tools, the data set was divided into three subsets: Reference sites, stressed sites, and intermediate sites. A "minimally disturbed" definition was used to identify reference sites (Stoddard et al. 2006) using objective criteria based primarily on landcover parameters as described in Ode et al. (in review); screening criteria are provided in Table 2. Reference sites were identified as those that had low levels of urban or agricultural land use or road density at the catchment scale, as well as within 5 km and 1 km upstream of the catchment. Sites were also excluded if there was moderate human activity in the riparian zone (measured as W1_Hall, Kaufmann et al. 1999), mining activity upstream, dams within 10 km, altered conductivity, or invasive invertebrates known at the sites.

Designation of highly stressed sites was necessary for calibration of the MMI (described below), and used for evaluation of both endpoints. Highly stressed sites were identified as those that met any of the following criteria: Developed land at the watershed, 1 km- or 5 km-scales ≥ 50%; road density ≥ 5 km/km^2; or W1_Hall ≥ 5. The reference data set was evaluated to ensure good representation of as many environmental settings as possible. Sites not identified as reference or stressed were designated as intermediate sites.

Reference and stressed sites were further divided into calibration (80%) and validation (20%) sets. Assignment to these sets was stratified to ensure representation of various subregions in both calibration and validation data sets. Because reference sites were found in nearly all regions of the state, 9 of the 10 subregions were used for stratification; the one reference site found in the Central Valley was combined into a stratum with sites from the Interior Chaparral.

In contrast to reference sites, stressed sites were scarce in mountainous regions, requiring a different stratification scheme for creation of calibration and validation site sets. Stressed sites were aggregated into five strata: North Coast, Chaparral, Central Valley, South Coast Xeric + Deserts, and Other Mountains (including the Modoc, Sierras, and South Coast Mountains). To avoid overrepresentation of highly stressed regions within the development data set, only 40 stressed sites from each region were assigned to the calibration sets. For sites with data from multiple sampling events, one sample was randomly designated for use in index development or performance evaluation. The distribution of sites used for development is summarized in Table 3.

Predicting number of taxa and development of the O/E

Taxonomic completeness, as measured by the ratio of observed-to-expected taxa (O/E), quantifies degradation as loss of taxa. To measure taxonomic completeness of bioassessment samples, a RIVPACS-type O/E index was developed to calculate the ratio of observed to expected taxa following Wright et al. (2000). RIVPACS models identify biologically homogeneous clusters of reference sites, then uses predictive models to determine probabilities that a test site is similar to each cluster. The probability of observing a taxon at a given test site (i.e., the capture probability) is then calculated as the probability of being similar to a reference cluster, multiplied by the frequency of the taxon in that cluster, summed across all reference clusters.

In order to identify biologically homogenous clusters of reference sites, samples were aggregated to OTUs and standardized to a count of 400 through random subsampling without replacement, after excluding ambiguous taxa. Standard samples were then transformed into presence/absence data and rare OTUs (i.e., occurring in fewer than 5% of reference calibration samples) were removed. A dendrogram was created using Sørensen as a distance measure and flexible beta (beta = -0.25) for linkage. Clusters containing at least 10 sites and subtended by relatively long branches (to maximize variance in taxonomic composition among groups) were manually identified through visual inspection of the dendrogram. Cluster analyses were performed in R version 2.15.2 using scripts written by J. Van Sickle and the cluster package (Maechler et al. 2012).

In order to predict group membership for novel sites, a 10,000-tree random forest model was constructed using the randomForest package in R (Liaw and Wiener 2002). First, candidate predictors that were minimally correlated with each other (Pearson's $R^2 \leq 0.5$) were identified (Table 1); within sets of correlated variables, the one that was simplest to calculate (e.g., calculated from point data, rather than delineated catchments) was selected as a candidate predictor. All candidate predictors were derived from GIS data and reflected environmental gradients that are minimally affected by human activity (e.g., elevation, geology). An initial model using all candidate predictors was then refined by hand using subsets of predictors with high observed Gini importance. A final model was selected that minimized standard deviation of O/E scores at calibration reference sites with the fewest number of predictors. Null models in which all reference calibration sites are treated as a single group were also produced (Van Sickle et al. 2005). Because previous studies have shown that exclusion of species with low capture probabilities improves model performance (e.g., Hawkins et al. 2000, Ostermiller and Hawkins 2004), the O/E model based on a capture probability ≥ 0.5 was used.

Predicting metric values and development of the pMMI

Alterations to the ecological structure of samples may be measured as changes in multimetric index (MMI) scores from reference expectations. To develop a pMMI, we followed the approach of Vander Laan et al. (in press). In contrast to traditional MMIs, a pMMI reduces the effects of natural gradients on metric values by predicting the expected value under a given environmental setting and using the residual instead of the raw metric for scoring. Although traditional MMIs may reduce the effects of natural gradients through typological approaches (e.g., ecoregions, as in Ode et al. 2005), they do not provide site-specific expectations for different environmental settings within each stream type (Hawkins et al. 2010).

To construct the pMMI, metric values at reference sites were predicted using random forest models from GIS-based predictors minimally affected by human activity. These models were developed using only the reference calibration data set. Deviation from predicted was measured as the metric residual, and scored on a scale from 0 to 1. Unlike O/E approach, the pMMI was calibrated not just by reference condition, but also by metric values at stressed sites. Specifically, the range of values from reference to stressed sites is used for metric scoring.

The pMMI was developed following 6 steps: 1) Metric calculation; 2) Initial model development; 3) Metric screening; 4) Model refinement; 5) Metric scoring; 6) Index aggregation and standardization. Apart from steps 2 and 4, this process is comparable to the steps required to develop a traditional MMI (e.g., Stoddard et al.2008). To contrast with a non-site-specific approach "null" MMI was also developed by following the same process, but substituting the mean metric value at reference calibration sites for site-specific predictions produced by the models.

Metric calculation

Biological metrics that characterize the ecological structure of the benthic community were calculated for each sample in the data set. In order to calculate metrics, biological data were first standardized with respect to taxonomic effort and sample size. The Standard Effort Level 1 (i.e., most taxa to genus, with midges left at family) defined by the Southwest Association of Freshwater Invertebrate Taxonomists (SAFIT, Richards and Rogers 2011) was used to standardize taxonomy. Samples with more than 500 individuals were then standardized to a count of 500 using random sampling without replacement.

A suite of 51 widely used bioassessment metrics (Table 4) was then calculated using custom scripts in R, as well as the Vegan package (Oksanen et al. 2013) for diversity indices. Conceptually related metrics were assigned to metric groups (e.g., % EPT and EPT taxa were assigned to the group "EPT metrics"). These metrics and groups are summarized in Table 4.

Initial model development

Preliminary predictive models were developed for all 51 metrics so that residuals from predicted values could be screened for inclusion in the pMMI. Because of the large number of models evaluated, a quick-and-dirty approach to model development was used prior to metric selection. Therefore, refinement of the models (specifically, reducing the number of predictors) would be necessary for only selected metrics.

Initial random forest models were built with 1000 trees for each metric using all 18 candidate predictors. Models that explained more than 10% of the variance in the metric were used to predict metric values for the full data set, and then calculate metric residuals. Otherwise, raw metric values were used for further analysis.

Metric selection

Metrics were selected in a stepwise fashion that maximized responsiveness to stress and minimized redundancy. Responsiveness was quantified as the absolute value of the t-statistic between the reference and stressed subsets of the calibration data set (using raw metrics or metric residuals, as determined above). The metric with the greatest responsiveness (i.e., highest absolute t-statistic) was selected first. Conceptually redundant metrics were excluded from further consideration if they belonged to the same metric group as the selected metric (e.g., Coleoptera taxa and percent Coleoptera), and statistically redundant metrics were excluded if they had a Person's $R^2 \geq 0.5$ with the selected metric. The next most responsive metric was selected, and the process repeated until there were no candidate metrics with an absolute t-statistic greater than 10.

Model refinement

Because the initial random forest models used many candidate predictors, models were refined to reduce the number of predictors. In contrast to the manual refinement process used for the O/E, an automated approach was used for the pMMI. This automation was useful because of the large number of models requiring refinement, even after metric selection. To refine the random forest models for the selected metrics, the caret package was used in R to implement recursive feature elimination (Kuhn et al. 2012). With recursive feature elimination, predictors are iteratively excluded from the model to identify the subset that maximizes the percent variance explained. Subsequently, the selected predictors were used to build a new model using the randomForest package using 1000 trees. If the final model explained more than 10% of the variance in the metric, metric residuals were used for subsequent analysis; otherwise the raw metrics were used.

Metric scoring

Metrics or metric residuals were scored following Cao et al. (2007). Scoring transforms metrics or residuals to a standard scale ranging from 0 (i.e., similar to stressed sites) to 1 (i.e., similar to reference sites). Metrics that decrease with stress were scored as follows:

$$(\text{Observed metric} - \text{Min}_d)/(\text{Max}_d - \text{Min}_d)$$

where Min_d is the fifth percentile of stressed calibration sites and Max_d is the 95th percentile of reference calibration sites; the fifth and 95th percentiles were used instead of minimum or maximum values because these distribution points are more robust to outliers, and yield more responsive and less variable metrics (Blocksom et al. 2003, Stoddard et al. 2008). Metrics that increase with stress were scored as follows:

$$(\text{Observed metric} - \text{Max}_i)/(\text{Min}_i - \text{Max}_i)$$

where Min_i is the 5th percentile of reference calibration sites, and Max_i is the 95th percentile of stressed sites. Scores were then trimmed to 0 or 1.

MMI aggregation and standardization

The raw index was calculated as the mean score for all selected metrics. This raw index was then divided by the mean of reference calibration sites so that the pMMI had a reference expectation of 1 and was on the same scale as the O/E.

Performance Evaluation

Evaluation of index performance focused on accuracy, precision, and sensitivity. Performance of each index compared to its null counterpart. Many of the approaches to measuring performance have been widely used in index development literature (e.g., Hawkins et al. 2000, Clarke et al. 2003, Ode et al. 2008). Because all indices were scored on similar scales (i.e., a minimum of zero, with a reference expectation of 1), no adjustments were required to make comparisons (Herbst and Silldorff 2006, Cao and Hawkins 2011).

Accuracy

Accuracy was defined as the ability of an index to provide high scores at reference sites, regardless of environmental setting. Operationally, accuracy was evaluated as the closeness of the mean of reference scores to 1. However, because the means of reference scores for both MMIs and the null O/E were mathematically fixed to 1, this evaluation is only meaningful for validation data.

Bias, which is related to accuracy, was evaluated against both categorical gradients and continuous gradients. For categorical gradients (e.g., regions), ANOVA was used to evaluate the consistency of scores at reference sites among categories. For both calibration and validation data sets, scores were compared across regions of the state. A high F-statistics indicates bias.

Bias was also evaluated as the percent of variance explained in the indices by a 1000-tree random forest model based on several natural gradients (Table 1). Percent variance explained was expected to be low (or even negative) if the index was not biased by natural gradients. For this analysis, both validation and calibration reference sites were used in a single model.

Because random forest cannot accommodate missing data, bias related to field-measured values was assessed by regressing index scores against single variables (specifically, % fast-water habitat, % sands and fines, and slope). Again, both validation and calibration reference sites were combined for these analyses.

Finally, to see if there was an effect of temporal variability on index, scores from reference sites (both calibration and validation), we tested scores against year using ANOVA. Seasonal effects were assessed by regressing reference site scores against the cosine of proportion of year.

Precision

Precision was defined as the consistency of scores at among reference sites or within replicate samples. It is a product of the variability of the index. Precision was first evaluated as the standard deviation of indices at reference sites, for both calibration and validation sets separately. Additionally,

precision was evaluated as the mean within-site standard deviation at sites where replicate samples were calculated. For both measures, smaller standard deviations indicate better precision.

Sensitivity

Sensitivity was defined as the strength and direction of response of an index to stress. Sensitivity is an aggregate property that incorporates both accuracy and precision; that is, sensitivity in an index arises when it can accurately detect large differences between stressed and unstressed sites, relative to the inherent variability of the index. As with accuracy, sensitivity was evaluated for both categorical and continuous gradients.

First, sensitivity was evaluated comparing reference and stressed sites. A large t-statistic indicates a sensitive index. This analysis was conducted for both the calibration and validation data sets separately.

Because the t-test examines differences in scores only at the extreme ends of the stressor gradient (i.e., reference vs. highly stressed sites), additional analyses were conducted to assess response across a broader range of this gradient. For example, sensitivity was also evaluated as the percent of variance explained in the indices by a 1000-tree random forest model based on stressor gradients (Table 2). For this analysis, a sample from every site in the data set (i.e., reference, stressed, and intermediate sites) was used in a single model. A large percent of variance explained indicates good sensitivity of the index to stressor gradients. Because random forest cannot accommodate missing data, field-measured variables were excluded from this analysis.

Finally, sensitivity to selected stressors was examined using bivariate plots to examine the range and shape of the response. Because this approach does not have the data requirements of the random forest approach, field-measured stressors (e.g., W1_Hall, % Sands and Fines) were included in the analysis.

Comparing the endpoints in different settings

To see if environmental setting affected agreement between the two endpoints, pMMI scores were regressed against the O/E. Levels of agreement were determined by setting benchmarks at the 1st and 10th percentiles of the reference calibration distribution: Samples with scores above the 10th percentile for an endpoint were considered to be in reference condition for that endpoint, and samples with scores below the 1st percentile were considered to be in non-reference condition for that endpoint; sites with scores between these two benchmarks for either index were considered to be ambiguous and excluded from further analysis of agreement. Logistic regression was then used to see if probability of disagreement between the two endpoints depended on environmental setting defined by E.

To see if the O/E was less sensitive than the pMMI to loss of taxa in “low E” settings (e.g., E < 10), we compared the proportion of sensitive taxa expected by both endpoints under different settings defined by E. This analysis assumes that loss of sensitive taxa is the major component of the response to disturbance across settings. For the pMMI, this proportion was calculated as the predicted percent

intolerant taxa metric, as described above. For the O/E, this proportion was calculated as the percent of OTUs expected that are sensitive (specifically, OTUs with a median tolerance value < 3). Estimates from both the O/E and pMMI were plotted against E to see if they allowed consistent ranges of response across environmental settings. These predictions were also compared to the observed % intolerant taxa at reference sites to confirm the validity of these estimates.

Combining the endpoints

We explored combining the two endpoints into a single index because of the potential advantages of using multiple ways of characterizing benthic macroinvertebrate community structure. Furthermore, a combined index may have better or more consistent performance than the individual endpoints, given that each was suspected to lose sensitivity in certain settings or under certain types of disturbance. Therefore, a combined index (the California Stream Condition Index, CSCI) was calculated by averaging the pMMI and O/E. A null equivalent was calculated by averaging the null MMI and O/E.

Establishing and evaluating site-specific thresholds

To see if site-specific thresholds improved the accuracy of assessment indices, two approaches to establishing impairment thresholds were evaluated: A traditional, non-site-specific approach based on the variability of all reference calibration sites, and a site-specific approach based on a subset of the most environmentally similar reference calibration sites. In both cases, sites were considered to be in reference condition if the score was greater than the 10th percentile of the relevant set of reference sites. Only the combined index was used in this analysis.

In order to establish site-specific thresholds, pairwise environmental distances along four gradients (elevation, precipitation, temperature, and watershed area) were measured using Euclidean distance on range-standardized variables (i.e., observed value minus the minimum value, divided by the maximum minus the minimum values). For each test site, a set number of nearest reference calibration sites (25, 50, 75, 100, and 200, as well as the full set of 473) were identified. Scores at test sites were transformed into percentiles relative to each of these distributions.

Performance of thresholds

Performance of the different approaches to establishing thresholds was evaluated in multiple ways. First, ANOVAs were performed on percentile-transformed scores to determine if there was a bias among regions within reference site scores. Second, error rates were calculated as the portion of reference sites with scores in the lower 10^{th} percentile, and as the portion of stressed sites with scores higher than the 10^{th} percentile. These tests were performed on calibration and validation data sets separately, and for both the null and predictive index.

Results

Biological diversity reflects environmental diversity

Biological community structure varied strongly across natural gradients within the state, as indicated by multivariate analysis and evaluation of bioassessment metrics. For example, visual examination of the dendrogram produced by cluster analysis yielded 11 groups, ranging in size from 13 to 61 members (Figure 2). Although a few of these groups were geographically restricted, most were distributed across many regions of the state. For example, group 10 was concentrated in the Transverse Ranges of Southern California, and group 7 was entirely within the Sierra Nevada. In contrast, groups 1 and 4 were broadly distributed across the northern two-thirds of California. Environmental gradients associated with this geographic distribution were different among several groups. For example, groups 8 through 11, all located in the southern portions of the state, were generally drier and hotter than other groups. Biotic groups varied strongly by richness, affecting the expected number of taxa in each group. For example, the median number of expected taxa (i.e., sum of capture probabilities > 0.5) in group 3 was 17.2, but only 7.8 in group 11. The median number of expected taxa was below 10 for three of the eleven groups. These low-E groups were preponderantly located in the southern portions of the state. Similarly, metric values also varied strongly by natural gradients at reference sites (slope of regression: 1.4 ± 0.07 standard error. R^2: 0.33). For example, the number of EPT taxa observed at reference sites ranged from ~10 at the southern part of the state to more than 20 in the northern reaches.

Predictive models can create useful site-specific expectations

Predictive models successfully created site-specific expectations for both the multivariate endpoint (i.e., expected number of taxa) and the majority of metrics that were evaluated. Model and index development are discussed in each section below.

Predicting number of taxa

A random forest model based on five predictors (i.e., latitude, elevation, watershed area, rainfall, and temperature, Table 1) successfully predicted the number of taxa at reference sites, despite a fairly high out-of-bag classification error rate (i.e., 58%, ranging from 36% to 92% by group). For example there was a strong relationship between expected and observed taxa at reference sites, with an adjusted r^2 of 0.74 (calibration) and 0.64 (validation) at reference sites; the slope (i.e., 1.05 and 0.99 at calibration and validation reference sites, respectively) and intercept (i.e., -0.36 and 0.52) were both similar to what would be expected from a perfect prediction (i.e., slope of 1 and intercept of 0) (Figure 3).

Predicting metric values (and developing the pMMI)

All metrics

Predictive models successfully the reduced confounding influence of natural gradients in most metrics, yielding site-specific reference expectations (Table 4). For example, of the 51 metrics screened

for inclusion, initial random forest models (i.e., those based on all 18 candidate predictors) explained more than 10% of the variance for 36 of them; for 6 metrics, more than 40% of the variance was explained. Models that explained large amounts of variance were observed for all metric groups, apart from invasive species metrics. In general, models explained the most variance for percent-taxa metrics, and the least for percent-abundance metrics, although this pattern was not consistent for all metric types.

Metrics selected for pMMI

The metric selection procedure yielded a pMMI comprised of eight metrics, all of which were based on residual predictive models (Table 5). The variance explained by the models was high for several metrics. For example, the random forest model explained more than half the variance in the % intolerant taxa metric, and more than 40% of the shredder taxa and clinger taxa metrics. However, only 12% of variance was explained for collector taxa and % non-insect taxa metrics.

The selected metrics represented a variety of broad metric classes (Table 5). Three of the metrics were based on taxonomic characteristics, two on functional feeding groups or tolerance values, and one on habit. Some metrics (e.g., Coleoptera Taxa, % Non-Insect Taxa) were similar to those used in regional indices previously developed in California (e.g., Ode et al. 2005). However, some widely used metrics (e.g., EPT metrics) were not selected because they were highly correlated with other metrics that had greater responsiveness to stress in calibration data (Table 4).

Final random forest models for these 8 metrics varied in their ability to predict values of selected metrics at reference sites (Table 5, Figure 3). The percent variance explained by each random forest model ranged from a low of 12 (for % Collector taxa and % Noninsect taxa) to a high of 53 (for % Intolerant taxa). For calibration data, intercepts of regressions of observed versus expected values were significantly lower ($p < 0.05$) than zero for every metric, suggesting that models systematically under-predicted metric values at reference sites; however, p-values were higher for validation data, and were under 0.05 for only two metrics (i.e., Shannon diversity and Collector taxa). Slopes for calibration data were not significantly different from 1 for any metric; for validation data, significantly smaller slopes were observed for three metrics (i.e., Shannon diversity, Collector taxa, and % Noninsect taxa). Although correlation coefficients between observed and expected values at reference sites were very high (adjusted $R^2 > 0.9$) for all metrics), relationships were weaker for validation data, with R^2 ranging from 0.07 (for Collector taxa) to 0.60 (for % Intolerant taxa).

The number of predictors used in these eight models ranged from eight (for Clinger Taxa) to all 18 (for Coleoptera Taxa, Tolerance Value, and % Collector Taxa) (Table 1). Predictors related to location (e.g., latitude, elevation) were used in all models, whereas 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 metrics also had the highest importance, as measured by % increase in mean-squared error. The least frequently used predictor (i.e., % nitrogenous geology) was used in three models.

Indices based on predictive models have superior performance to null indices

Effects of predictive modeling on bioassessment metrics

For most metrics, reducing the influence of natural gradients through predictive modeling improved discrimination ability (Table 4). For example, in 37 of the 51 metrics evaluated, the t-statistic was greater for the residuals than the raw metric; only 8 metrics had greater discrimination ability (difference in t-statistic > 0.1) for the raw metrics than the residuals, and for 6, the difference in the t-statistic was less than 0.1.

Performance evaluation of the O/E, pMMI, and the combined index

By all measures, the predictive indices (whether used alone or combined) performed much better than their null counterparts, particularly with respect to bias (Table 6). For example, regional differences in scores at reference sites were large and significant for all null indices (Table 6 Part B, Figure 4), and responses to natural gradients were strong (Figure 5). In contrast, these biases were greatly diminished for predictive indices. Biases were reduced even for gradients unrelated to predictors used in predictive models, such as date of sampling.

Predictive modeling also improved several aspects of precision. Variability of scores at reference sites was lower for all predictive indices than for their null counterparts, particularly for the pMMI (Table 6). Regional differences in precision were larger for the MMIs than the O/Es (even with predictive models), and combining these two endpoints into a single index appeared to improve regional consistency in variability of the CSCI (Figure 4). Predictive modeling had a negligible effect on within-site replicability (Table 6 Part B).

In contrast to precision and accuracy, sensitivity was more affected by endpoint than index type. Specifically, the MMIs (both predictive and null) were slightly more sensitive than the combined indices, which in turn were more sensitive than the O/Es. This pattern was evident in all measures of sensitivity, such as the magnitude of t-statistics, variance explained by multiple stressors in a random forest model, or steepness of slopes against individual stressor gradients (Tables 6 Part B, Figure 6).

Setting may affect index sensitivity

Overall agreement between the two indices was high, but with a relative positive bias for the O/E, consistent with its lower sensitivity (Adjusted R^2 = 0.54, slope = 0.76 ± 0.02, intercept = 0.11 ± 0.01) (Figure 7). Consequently, the overwhelming majority of sites (98%) where the indices unambiguously disagreed (n=373) were in better condition for the O/E than the pMMI. These disagreements were preponderant in low-E settings (e.g., <10). Logistic regression showed that the probability of disagreement was highest when E was low (probability of disagreement = 0.84 - 0.17 E, $p < 0.01$, n = 1448).

Sensitivity of the MMI was greater in certain settings because the O/E had limited range of response in low-E settings, where few sensitive taxa are expected (Figure 8). For settings with more than 14 expected taxa, the proportion of sensitive taxa expected was consistently ~0.35. However, where E was lower than 14, the proportion of sensitive taxa expected decreased linearly, reaching zero at E of

~6. In contrast, when percent sensitive taxa was directly modeled (as in the pMMI), the decline is less severe (i.e., from ~0.40 to 0.20). Inspection of the data at reference sites indicates that sensitive taxa were truly present at these low-E settings (right panels in Figure 8) and that directly modeling the metric sets more accurate expectations for sensitive taxa in these settings (metric prediction vs. observed R^2 = 0.80; O/E prediction versus 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 in low-E settings.

Site-specific thresholds do not improve predictive indices

Establishing site-specific expectations obviated the need for site-specific thresholds. Site-specific thresholds had little impact on the performance of predictive models, but greatly reduced the bias of null models (Figure 9). In other words, once site-specificity was accommodated by predictive modeling, further accommodations were unfruitful. For example, the large regional differences evident in the null index scores decreased as the number of neighbors increased, approaching values of the ANOVA F-statistic observed for predictive models when 25 neighbors were used. In contrast, bias was low for predictive models, regardless of how many neighbors were used for comparison. Nonetheless, error rates were minimally affected by number of neighbors for either null or predictive models. In general, null models had lower reference error rates, and predictive models had lower stressed error rates, but neither index type responded strongly to changes in numbers of neighbors (data not shown).

Although using nearest neighbors improved the accuracy of null models, there was a tradeoff with sensitivity, as shown by the increased error rate in scoring stressed sites. For both stressed calibration and validation sets, the lowest error rate for the null models using all neighbors (data not shown).

Discussion

Many of the recent technical advances in bioassessment have centered on improving the performance of tools used to score the ecological condition of waterbodies. Much of the progress in this area has come from regional, national and international efforts to produce overall condition assessments of streams in their regions (Simpson and Norris 2000, EMAP papers, Hawkins 2006, Hering et al. 2006, Stoddard et al. 2006, Van Sickle et al. 2005, EU papers, Environment Canada, USEPA-NARS). A key challenge in completing these projects has been incompatibility among scoring tools designed to assess regions of different spatial scales. This issue has been well-documented for large scale programs attempting to integrate data from a patchwork of scoring tools built for smaller regions (Heinz Center 2002, Hawkins 2006, Meador et al. 2008, Pont et al. 2009, EU papers), but far less attention has been paid to the limitations of applying large regional tools to local scale assessments, particularly at the scale of individual stream reaches (Herlihy et al. 2008, Ode et al. 2008).

Ultimately, successful implementation of bioassessment techniques in a variety of applications requires that scoring tools perform well at both objectives: consistency throughout the region and accuracy at the site. Most scoring tools built to support large area condition assessments explicitly consider site-specific accuracy during model building and evaluation (see Hawkins 2006, REFs). However, these condition assessments don't require as much site-specific accuracy as site-specific applications. As long as over- and under-estimates of site condition balance each other, the resulting overall condition assessment will be unaffected (Ode et al. 2008, Yuan et al. 2008). In contrast, both consistency and site-specific accuracy are critical for assessments at local scales, particularly for regulatory applications. If an assessment tool sets biological expectations that are inappropriate for a site, inaccurate assessments may trigger costly and unnecessary remediation.

Our success in achieving high levels of both statewide consistency and site specificity in the CSCI was the result of three elements of its development. The first element was the large, representative, and rigorously evaluated reference data set (Ode et al. in review). To create an assessment tool that incorporates site-specificity, natural gradients that influence biological assemblages need to be fully accounted for (Stoddard et al. 2006). Even gradients that are broadly unimportant may be locally influential (Ode et al. 2008), and these gradients should be adequately represented in the reference data set. The data set we used required more than 10 years to collect the hundreds of samples necessary to capture the diversity of natural gradients throughout the state, is described by Ode et al. (in review). The breadth of sampling for reference sites across both space and time provides confidence in the applicability of the CSCI for the vast majority of wadeable perennial streams in California.

The second component that enabled the CSCI to achieve both consistency and site specificity was predictive modeling. Predictive modeling enabled the creation of site specific expectations for a variety of endpoints, and these models created superior indices 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 multivariate-based species loss indices (e.g., Reynoldson et al. 1997, Hawkins and Norris 2000, Van Sickle et al. 2005, Hawkins 2006, Mazor et al. 2006), but one of the first to show that the benefits are even greater for multi-metric indices.

The null multi-metric indices evaluated in this study were intentionally simplistic and do not reflect the more typical typological approaches to multi-metric index development, such as 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 in reference sites. However, regionalizations that lack standardization complicate inter-regional comparisons, especially for sites located near regional boundaries, inviting contentious arguments regarding applicability. Even if typological approaches provided equivalent performance to predictive indices, the latter would be preferred because of their improved interpretability.

The third component that enabled the CSCI to achieve both consistency and site specificity was the inclusion of multiple endpoints. The decision to combine endpoints for the CSCI was based, at least partly, on observations that the two endpoints had different sensitivities in different settings. For example, in drier, low elevation settings, the taxonomic completeness endpoint only predicted a small number of highly tolerant taxa (e.g., baetid mayflies) to occur since these tolerant taxa normally occur 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. The O/E did have some advantages over the pMMI. Notably, validation was better for the O/E than the pMMI. This result may be expected considering the more extensive calibration required for the pMMI. Although we were satisfied with the validation of both endpoints, the superior validation of the O/E should be considered its strength. Combining the O/E with the pMMI was an effective way to retain high sensitivity across these environmental settings.

In addition to the respective technical strengths of the two endpoints, there are several philosophical reasons why the combination of endpoints into a single CSCI results in a tool with more robust and defensible applicability (see Collier 2009). Whereas the O/E is sensitive to species loss and has clear application to biodiversity conservation, MMIs provide an explicit measure of other aspects of ecosystem function, like changes to trophic structure. MMIs are calibrated specifically designed to respond to stress, while O/Es only measure deviation from a reference state, independent of stressor gradients. Their different sensitivities enhance the utility of the combined index across a broader range of disturbances and function as multiple lines of evidence, providing greater balance and confidence in the results than a single index.

This study may reflect the limits of how much site specificity can be incorporated in an assessment tool intended for use in an environmentally complex region. Predictive models were only able to explain a portion of the variability observed at reference sites—sometimes a fairly small portion. For example, the standard deviation in the predictive O/E was only slightly lower than the null O/E (i.e., 0.19 vs. 0.21). Similarly, random forest models explained as little as 12% of the variability for certain metrics (specifically, % collector taxa and % non-insect taxa). The unexplained variance may be related to environmental gradients that are unsuitable for setting biological expectations (e.g., alterable gradients, like substrate composition or canopy cover), to gradients unrelated to those used for modeling (e.g., temporal gradients, like weather antecedent to sampling), or field and laboratory sampling error. Given the number and breadth of gradients evaluated for modeling, as well as the apparent lack of bias to gradients not used in modeling (e.g., date of sampling, substrate, slope), it is

unlikely that additional data or advanced statistical methods will greatly change the performance of these indices.

Our evaluation of site-specific thresholds reinforces the conclusion that predictive modeling alone achieved as much consistency as practical for this data set. Unlike Yuan et al. (2008), we did not see improvements to predictive indices when variable thresholds were used, perhaps because of the different scales of the two studies (i.e., nationwide vs. statewide). For the null index, where environmental gradients are not accounted for, there were large differences in site-specific thresholds between sub-regions of California. These differences increased as additional reference sites were added. However, the site-specific thresholds for the predictive model were much more similar between regions, and these differences did not change regardless of sample size. This pattern suggests that predictive modeling used for the CSCI is sufficient to account for natural variability at the range of reference sites in California. In light of the potential complications site-specific thresholds may impose on regulatory applications of the index, predictive indices were clearly preferable.

Conclusions

Many applications of bioassessment require that scoring tools perform well at the two objectives we addressed in this study: regional consistency and site-specific accuracy. Whereas condition assessments of large regions provide valuable context for interpreting local data, application to local management questions requires that tools give accurate assessments on every site where they are used. The proliferation of bioassessment literature in the past two decades documents a global shift toward the emphasis of ecological assessment techniques in aquatic resource management. However, there is still a long way to go before bioassessment realize its full potential for improving the management of stream health. In the USA and elsewhere, improving the technical capabilities of regional governments to produce accurate assessments is a key emphasis of national programs (USEPA 2013). Scoring tool performance should be viewed an important component of these capabilities

Acknowledgments

[Please let us know if you'd like to be acknowledged, or would prefer to be a co-author]

Cited Literature

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.

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

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.

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.

Clarke, R.T., J.F. Wright, 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 Asseessment 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.

Death R.G. and M.J. Winterbourne. 1994. Environmental stability and community persistence: A multivariate perspective. Journal of the North American Benthological Society. 13: 125-139.

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.

Hannaford, M.J. and V.H. Resh. 1995. Variability in macroinvertebrate raid-assessment surveys and habitat assessments in a northern California stream. Journal of the North American Benthological Society 14: 430-439.

Hawkins, C.P., R.H. Norris, J.N. Hogue, and J.W. Feminella. 2000. Developent 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. 2010. The reference condition: Predicting benchmarks for ecological and water-quality assessments. Journal of the North American Benthological Society 29: 312-343.

The Heinz Center (The 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, NY

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, CA

Hering, D., R.K. Johnson, S. Kramm, S. Schmutz, K. Szoszkiewicz, and P.F. Verdonschot. 2006. Assessment of European streams with diatoms, macrophyts, 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.

Hughes, R.M. 1995. Defining acceptable biological status by comparing with reference conditions. Pages 31 – 47 *in* W.S. Davis and T. P. Simon, editors. Biological assessment and criteria: tools for water resource planning and decision making. Lewis Publishers. Chelsea, MI.

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

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. US EPA. Office of Research and Development. Washington, DC.

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.

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.

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

Maechler, M. P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. 2012. cluster: Cluster Analysis Basics and Extensions. R package version 1.14.3

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 Science 63: 394-411.

Mazor, R.D., K. Schiff, P. Ode, E.D. Stein. 2012. Bioassessment in nonperennial streams. Technical Report 695. Southern California Coastal Water Research Project. Costa Mesa, CA. Available from ftp://ftp.sccwrp.org/pub/download/DOCUMENTS/TechnicalReports/695_NonperennialStreamsSanDiego.pdf

Meador, M.R., T.R. Whittier, R.M. Goldstein, R.M. Hughes, and D.V. Peck. 2008. Evaluatoin 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.

Norris, R.H. 1995. Biological monitoring: The dilemma of data analysis. Journal of the North American Benthological Society 14: 440-450.

Norris, R.H. and C.P. Hawkins. 2000. Monitoring river health. Hydrobiologia 435: 5-17.

Ode, P.R., A.C. Rehn, and J.T. May. 2005. A quantitative tool for assessing the integrity of southern coastal California streams. Environmental Management 35: 493-504.

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, R.D. Mazor, K. Schiff, J. May, L. Brown, D. Gillett, and D. Herbst. In review. An approach for evaluating the suitability of a reference site network for the ecological assessment of streams in environmentally complex regions.

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.

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

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

Peck, D.V., A.T. Herlihy, B.H. Hill, R.M. Hughes, P.R. Kaufmann, D.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, OR.

Poff, N.L., J.D.. Olden, N.K.M. Vieira, D.S. Finn, M.P. Simmons, and B.C. Kondratieff. 2006. Functional trait niches of North American lotic insects: traits-based ecological applications in light of phylogenetic relationships. Journal of the North American Benthological Society 25: 730-755.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 Soceity. 138: 292-305.

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 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, CA.

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.

Resh, V.H. 2008. Which group is best: Attributes of different biological assemblages used in freshwater biomonitoring programs. Environmental Monitoring and Assessment 138: 131-138

Reynoldson, T.B., R.C. Bailey, K.E. Day, and R.H. Norris. 1995. Biological guidelines for freshwater sediment based on BEnthic Asessment of SedimenT (the BEAST) using a multivariate approach for predicting biological state. Australian Journal of Ecology 20: 198-219.

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

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, CA. Available from www.safit.org.

R Core Team. 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Schoolmaster Jr., D.R, J.B. Grace, and E.W. Schweiger. A general theory of multimetric indices and their properties. Methods in Ecology and Ecolution 3: 773-781.

Schoolmaster Jr., 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.

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. Sutciffe, 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.

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.

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

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

Townsend, C.R. and A.G. Hildrew. 1994. Species traits in relation to a habitat templet for river systems. Freshwater Biology. 31: 265-275.

USEPA (U.S. 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.U.S. Environmental Protection Agency, Office of Environmental Information and Office of Water, Washington, DC.

USEPA (U.S. Environmental Protection Agency). 2013. Biological Program Review: Assessing Level of Technical Rigor to Support Water Quality Management. EPA 820-R-13-001. USEPA Office of Science and Technology. Washington, DC

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

Vander Laan, J.J. and C.P. Hawkins. In press. Effects of spatial isolation and sample evenness on bioassessment indices of arid zone streams.

Wright, J.F, D.W. Sutcliffe, M.T. Furse. 2000. Assessing the Biological Quality of Freshwaters: RIVPACS and Other Techniques. Freshwater Biological Association. Ambleside, UK. 373 pages.

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(1):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.

Tables

Table 3. Number of sites used to develop the index. Cal: Calibration data set. Val: Validation data set.

	Reference		Stressed		Intermediate
Region	Cal	Val	Cal	Val	
North Coast	60	16	40	13	115
Chaparral	74	19	40	102	192
--Coastal Chaparral	48	13	36	91	148
--Interior Chaparral	26	6	4	11	44
South Coast	96	23	48	220	287
--South Coast Mountains	69	17	8	2	111
--South Coast Xeric	27	6	40	218	176
Central Valley	1	0	40	18	11
Sierra Nevada	221	55	26	6	186
--Western Sierra Nevada	105	26	18	4	96
--Central Lahontan	116	29	8	2	90
Deserts / Modoc	21	4	3	2	46
Total	473	117	197	361	837

Table 1. Predictors and their importance for random forest models of each endpoint and metric. MSE: Mean-squared error. Dashes indicate that the predictors were not used to model the metric. Sources: A. National Elevation Dataset (http://ned.usgs.gov/). B. PRISM climate mapping system (http://www.prism.oregonstate.edu). C: Generalized geology, minerology, and climate data from conductivity prediction model (Olson and Hawkins 2012).

		Predictor importance (% increase MSE)									
Predictor	Description	O/E	Shannon diversity	% intolerant taxa	Tolerance value	% collector taxa	Shredder taxa	Clinger taxa	Coleoptera taxa	% non-insect taxa	Source
Location											
New_Long	Longitude	--	0.08	0.002	0.11	1.3	0.9	10.7	0.8	0.0006	
New_Lat	Latitude	0.09	0.06	0.004	0.17	1.4	1.2	10.5	0.6	0.0006	
SITE_ELEV	Elevation	0.11	0.05	0.004	0.05	0.8	0.4	6.3	1.2	0.0012	A
Catchment											
LogWSA	Log watershed area	0.06	--	--	0.02	0.3	1.3	--	0.1	--	A
ELEV_RANGE	Difference in elevation between sample point and highest point in catchment	--	0.01	--	0.03	0.4	0.2	3.2	--	--	A
Climate											
TEMP_00_09	10-y (2000-2009) average tempearture	0.09	0.04	0.005	0.09	0.8	0.6	6.2	0.4	0.0008	B
PPT_00_09	10-y (2000-2009) average precipitation	0.07	0.02	0.003	0.12	0.6	0.9	4.5	0.3	0.0006	B
SumAve_P	Average of mean June to Sep 1971 to 2000 monthly ppt	--	0.01	0.003	0.07	0.4	0.2	4.6	0.3	0.0005	B
Geology											

KFCT_AVE	Average soil erodibility factor (K)	--	0.02	0.003	0.05	0.7	0.2	--	0.4	--	C
BDH_AVE	Average bulk density	--	0.02	--	0.07	0.4	0.3	4.2	0.3	0.0004	C
MgO_Mean	% MgO geology	--	0.01	--	0.04	0.4	0.2	--	0.2	0.0002	C
Log_P_MEAN	Log % P geology	--	0.01	--	0.05	0.3	0.2	--	0.2	0.0003	C
CaO_Mean	% CaO geology	--	0.01	--	0.03	0.3	0.2	--	0.1	0.0002	C
PRMH_AVE	Average soil permeability	--	0.01	0.002	0.02	0.9	--	--	0.4	0.0002	C
S_Mean	% S geology	--	0.01	--	0.02	0.3	0.2	--	0.1	--	C
PCT_SEDIM	% Sedimentary geology	--	0.01	--	--	0.3	--	--	0.1	0.0001	C
LPREM_mean	Average log geometric mean hydraulic conductivity	--	--	0.002	0.04	--	0.2	--	0.2	0.0002	C
Log_N_MEAN	Log % N geology	--	--	--	0.02	0.2	--	--	0.1	--	C

Table 2. Stressors and other environmental gradients used to evaluate index performance. WS: Watershed. 5K: Watershed clipped to a 5-km buffer of the sample point. 1K: Watershed clipped to a 1-km buffer of the sample point. Variables marked with an asterisk (*) indicate those used in the random forest evaluation of index sensitivity. W1_HALL: proximity-weighted human activity index (Kaufmann et al. 1999). Sources are as follows: A: National Landcover Data Set. B: Custom roads layer. C: National Hydrography Dataset Plus. D: National Inventory of Dams. E: Mineral Resource Data System. F: Predicted specific conductance (Olson and Hawkins 2012). G: Field-measured variables.

	Variable	Scale	Threshold	Unit	Source
*	% Agriculture	1k, 5k, WS	3	%	A
*	% Urban	1k, 5k, WS	3	%	A
*	% Ag + % Urban	1k and 5k	5	%	A
*	% Code 21	1k and 5k	7	%	A
*		WS	10	%	A
*	Road density	1k, 5k, WS	2	km/km^2	B
*	Road crossings	1k	5	crossings/ km^2	B, C
*		5k	10	crossings/ km^2	B, C
*		WS	50	crossings/ km^2	B, C
*	Dam distance	WS	10	km	D
*	% canals and pipelines	WS	10	%	C
*	Instream gravel mines	5k	0.1	mines/km	C, E
*	Producer mines	5k	0	mines	E
	Specific conductance	site	99/1**	prediction interval	F
	W1_HALL	reach	1.5	NA	G
	% Sands and Fines	Reach		%	G
	Slope	Reach		%	G

** The 99^{th} and 1^{st} percentiles of predictions were used to generate site-specific thresholds for specific conductance. Because the model was observed to under-predict at higher levels of specific conductance (data not shown), a threshold of 2000 µS/cm was used as an upper bound if the prediction interval included 1000 µS/cm.

Table 4. Metrics evaluated for inclusion in the pMMI. Subheaders identify metric groups. EPT: Ephemeroptera, Plecoptera, and Trichoptera. Index: Indicates whether metric was selected for inclusion in the predictive (p) or null (n) MMI. Variance explained: Percent variance explained by the initial random forest model based on all candidate predictors. t (null): t-statistic for the comparison of the raw metric between the reference and stressed samples within the calibration data set. t (pred): t-statistic for the residual metrics. Max R^2: Maximum Pearson correlation coefficient between the metric (or metric residual if variance explained ≥ 10) and the selected metrics. Tolerance, functional feeding group, and habit data were from CAMLnet (2003). Taxa included in the invasiveness metrics were predominantly *Potamopyrgus antipodarum*, *Melanoides tuberculata, Corbicula* sp., and most crayfish species.

Metric	Index	Variance explained	t (null)	t (pred)	Max R^2	Response
Taxonomy						
Taxonomic Richness		28	10.1	15.4	0.67	Decrease
Shannon Diversity	p	19	7.4	10.4	0.45	Decrease
Simpson Diversity		10	4.6	5.0	0.69	Decrease
% Dominant		17	-6.2	-8.8	0.91	Increase
% EPT	n	18	10.2	11.5	0.59	Decrease
% EPT Taxa		34	16.3	15.1	0.62	Decrease
EPT Taxa		43	14.3	18.2	0.82	Decrease
% Coleoptera		12	2.7	8.6	0.27	Decrease
% Coleoptera Taxa		31	6.4	12.6	0.78	Decrease
Coleoptera Taxa		36	7.5	16.8	0.37	Decrease
% Diptera		11	1.2	1.4	0.13	Decrease
% Diptera Taxa		16	-2.5	0.9	0.07	Decrease
Diptera Taxa		1	7.8	11.6	0.29	Decrease
% Chironomidae		11	-0.1	0.0	0.15	Decrease
% Non-insect		8	-10.8	-9.5	0.44	Increase
% Non-insect Taxa	p, n	12	-15.3	-15.9	0.44	Increase
Non-insect Taxa		5	-7.8	-7.1	0.24	Increase
Tolerance						
% Intolerant		24	14.4	17.7	0.61	Decrease
% Intolerant Taxa	p, n	53	18.1	18.1	0.49	Decrease
Intolerant Taxa		51	15.9	18.4	0.74	Decrease
% Tolerant		13	-8.0	-5.2	0.36	Increase
% Tolerant Taxa		25	-14.8	-14.3	0.56	Increase
Tolerant Taxa		9	-9.5	-4.9	0.18	Increase
Tolerance Value	p	26	-12.7	-14.0	0.36	Increase
Functional Feeding Group						
% Collectors		8	-6.4	-9.6	0.23	Increase
% Collector Taxa	p	23	-6.9	-5.1	0.21	Increase
Collector Taxa		13	8.6	13.3	0.44	Decrease

% Predators		10	1.7	1.9	0.15	Decrease
% Predator Taxa		16	2.7	0.0	0.03	Increase
Predator Taxa		12	8.1	8.7	0.39	Decrease
% Scrapers		10	4.0	8.6	0.12	Decrease
% Scraper Taxa		27	2.6	5.5	0.18	Decrease
Scraper Taxa		42	7.0	13.7	0.60	Decrease
% Shredder		17	6.5	9.8	0.20	Decrease
% Shredder Taxa		30	9.5	9.5	0.75	Decrease
Shredder Taxa	p, n	42	10.4	14.8	0.24	Decrease
Habit						
% Burrowers		14	-5.4	-5.6	0.27	Increase
% Burrower Taxa		7	-7.9	-7.4	0.27	Increase
Burrower Taxa		5	4.0	6.9	0.09	Decrease
% Climbers		-6	-0.7	6.0	0.03	Decrease
% Climber Taxa		22	-5.4	-2.7	0.02	Increase
Climber Taxa		20	0.4	7.7	0.21	Decrease
% Clingers	n	6	10.2	10.8	0.27	Decrease
% Clinger Taxa		34	12.3	11.5	0.53	Decrease
Clinger Taxa	p	43	14.3	20.8	0.49	Decrease
% Swimmers		4	-2.3	1.1	0.01	Decrease
% Swimmer Taxa		24	-6.0	-2.9	0.14	Increase
Swimmer Taxa		7	5.3	10.7	0.22	Decrease
Invasiveness						
%Invasive		0	-3.7	-3.7	0.06	Increase
% Invasive Taxa		0	-7.3	-7.3	0.22	Increase
Invasive Taxa		0	-7.7	-7.7	0.19	Increase

Table 5. Summary of selected metrics. Cal: Results for calibration data. Val: Results for validation data. t-statistic: Value from student's t-test comparing reference and stressed sites, using pooled variance. Regression statistics refer to the relationship between predicted and observed values. SE: Standard Error. Slopes marked with an asterisk are significantly different from 1 ($p<0.05$). Intercepts marked with an asterisk are significantly different from 0 ($p < 0.05$). R^2: Pearson correlation between predicted and observed values. Min: Minimum value used for scoring. Max: Maximum value used for scoring.

Metric	Development Set	Variance explained	t-statistic	Slope	SE	Intercept	SE	R^2	Min	Max
Shannon diversity	Cal	20	10.6	1.33	0.02	-0.77*	0.04	0.91	-1.5	0.3
	Val		13.3	0.55*	0.14	1.09*	0.33	0.11		
% Intolerant taxa	Cal	53	17.8	1.17	0.02	-0.06*	0.01	0.93	-0.3	0.1
	Val		15.9	1.09	0.08	-0.03	0.03	0.60		
Tolerance Value	Cal	28	-14.1	1.27	0.02	-1.16*	0.08	0.91	-0.6	2.5
	Val		-10.9	0.91	0.14	0.38	0.61	0.25		
Collector taxa	Cal	12	13.2	1.43	0.02	-4.93*	0.21	0.93	-8.7	2.0
	Val		12.9	0.60*	0.20	4.70*	2.30	0.07		
Shredder taxa	Cal	41	14.7	1.25	0.02	-0.80*	0.06	0.93	-3.5	1.2
	Val		8.4	0.87	0.11	0.18	0.39	0.36		
Clinger taxa	Cal	44	19.6	1.20	0.02	-3.03*	0.27	0.91	-13.8	3.6
	Val		14.8	0.91	0.12	1.47	1.79	0.35		
Coleoptera taxa	Cal	36	17.0	1.24	0.02	-0.73*	0.06	0.92	-4.1	1.2
	Val		16.2	0.82	0.11	0.48	0.37	0.32		
% Noninsect taxa	Cal	12	-16.2	1.40	0.02	-0.06*	0.00	0.91	0.0	0.5
	Val		-19.7	0.62*	0.18	0.06	0.03	0.09		

Table 6. Part A. Means and standard deviations of each site set for each index. Part B. Performance measures to compare indices. For accuracy tests, only reference sites were used. F (cal): F statistic for differences in scores at calibration sites among 5 PSA regions (excluding Central Valley; 467 residual df). F (val): F statistic for differences in scores at validation sites among 5 PSA regions (excluding Central Valley; 112 residual df). VarExp: Variance in index scores explained by natural gradients at reference sites (n=590). Precision: SD: Mean within-site standard deviation for scores at 820 replicated sites. Sensitivity: t.cal: t-statistic for comparing reference and stressed sites in the calibration data set. t.val: t-statistic for comparing reference and stressed sites in the validation data set. VarExp: Variance in index scores explained by stressor gradients at all sites (n=1985).

Part A

		Reference				Stressed				Intermediate	
		Calibration		Validation		Calibration		Validation			
Form	Type	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
CSCI	Predictive	1.01	0.12	0.99	0.16	0.67	0.24	0.61	0.21	0.88	0.20
	Null	1.00	0.20	1.00	0.20	0.62	0.31	0.48	0.25	0.83	0.25
MMI	Predictive	1.00	0.08	0.95	0.16	0.59	0.26	0.52	0.21	0.83	0.21
	Null	1.00	0.25	0.99	0.26	0.53	0.37	0.38	0.29	0.79	0.31
O/E	Predictive	1.02	0.19	1.03	0.19	0.75	0.25	0.71	0.24	0.94	0.23
	Null	1.00	0.21	1.00	0.21	0.70	0.27	0.59	0.25	0.87	0.25

Part B

		Accuracy			Precision	Sensitivity		
Form	Type	F (cal)	F (val)	VarExp	Within-site SD	t.cal	t.val	VarExp
CSCI	Predictive	1.1	1.3	-9	0.08	23	26	53
	Null	49.7	6.8	38	0.08	20	28	65
MMI	Predictive	0.4	1.2	-17	0.08	26	29	63
	Null	44.5	10.0	39	0.11	20	27	63
O/E	Predictive	1.2	1.0	-3	0.11	18	18	32
	Null	23.5	0.9	17	0.10	17	22	48

Figures

Figure 1. Regions and subregions of California. 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.

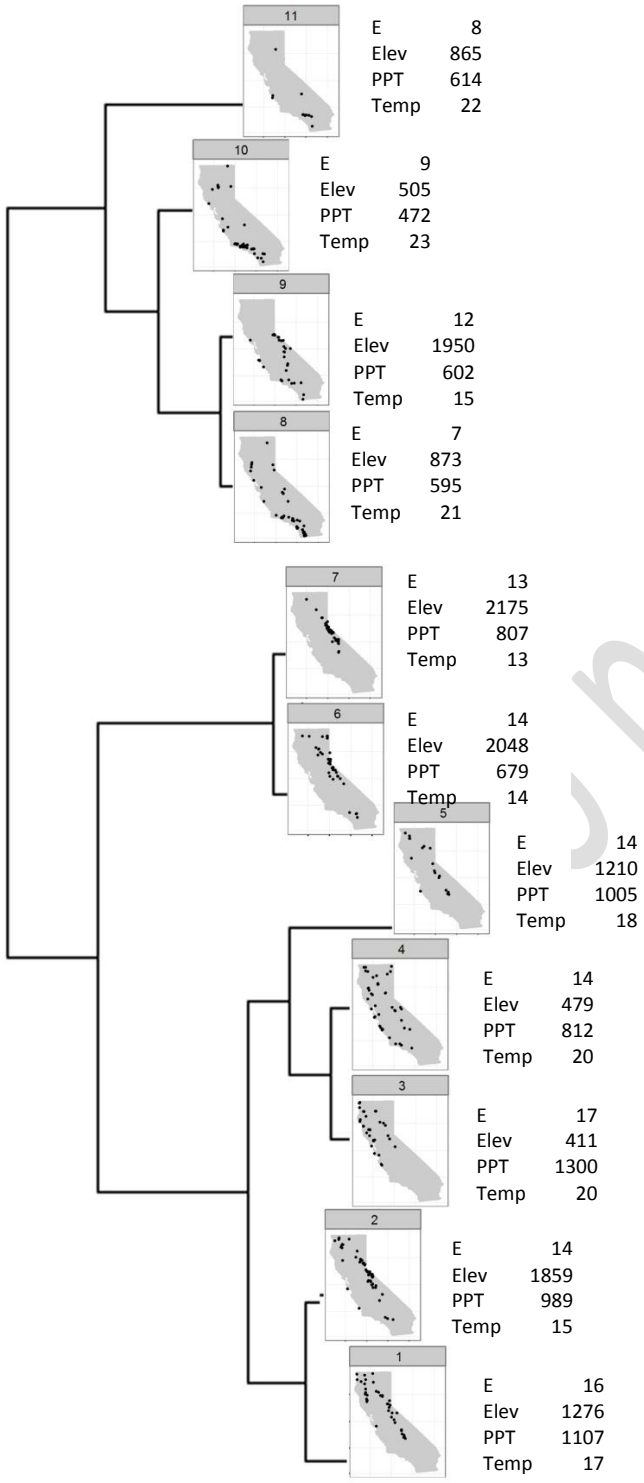

Figure 2. Dendrogram and geographic distribution of each group identified during cluster analysis. Numbers next to leaves are median values for expected number of taxa (E), elevation (Elev), precipitation (PPT), and temperature (Temp).

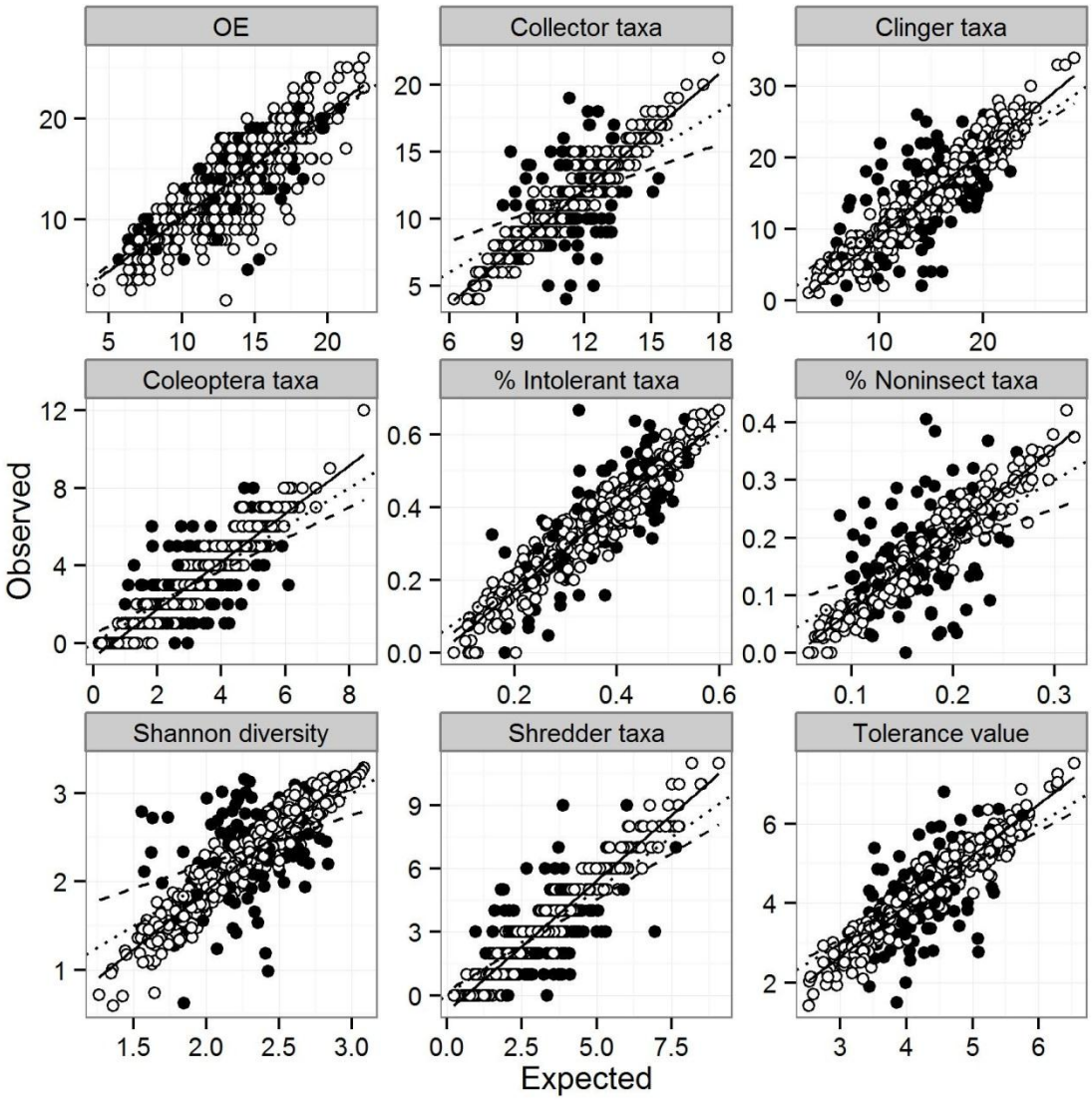

Figure 3. Expected versus observed values at reference sites. White symbols represent calibration sites, and black symbols represent validation sites. The solid line represents the regression for calibration data; the dashed line represents the regression for validation data; and the dotted line represents the line of perfect prediction.

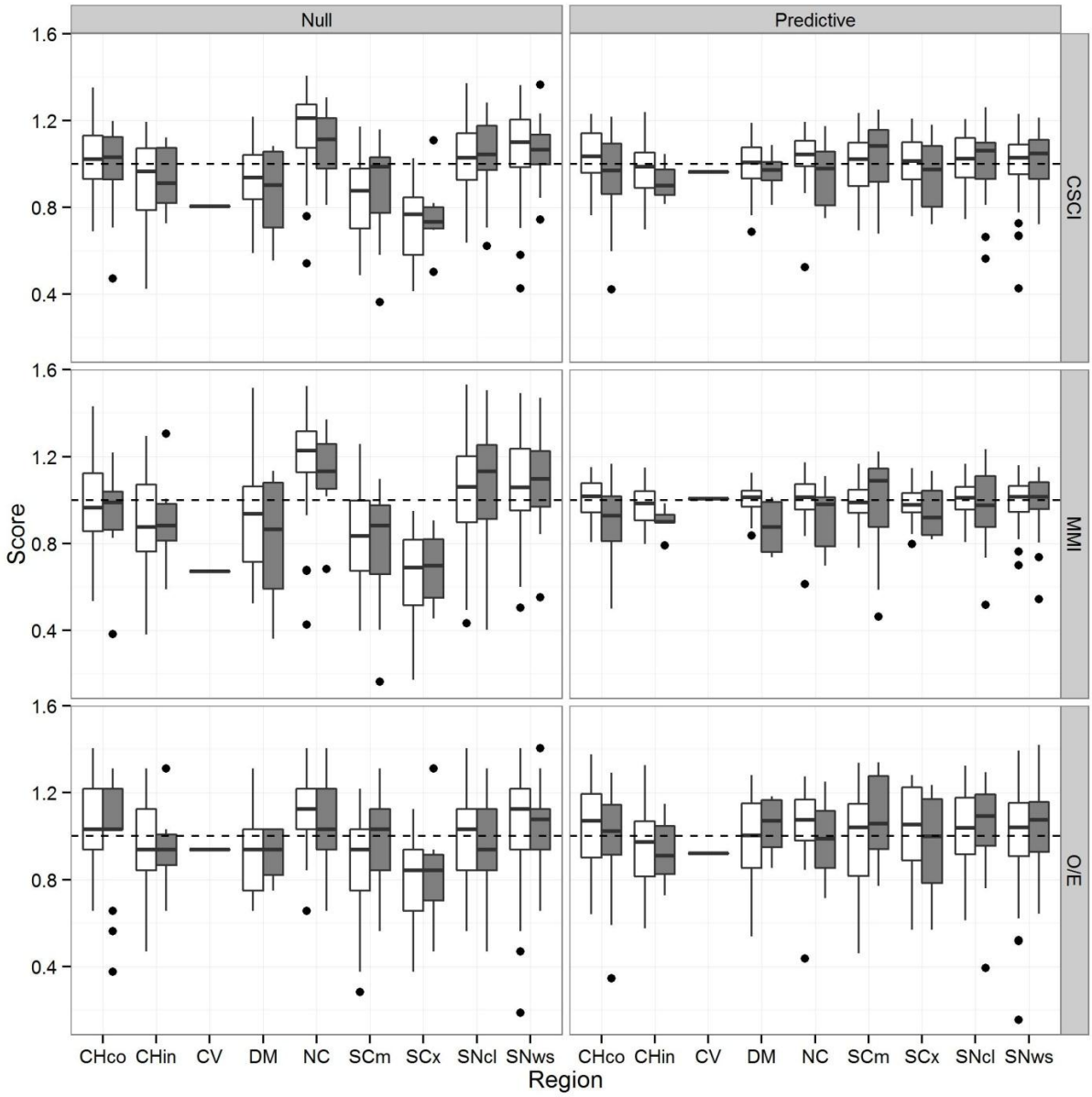

Figure 4. Distribution of scores for null and predictive models for the O/E, MMI, and the aggregated index (CSCI). The horizontal dashed lines indicate the expected value at reference sites (i.e., 1).

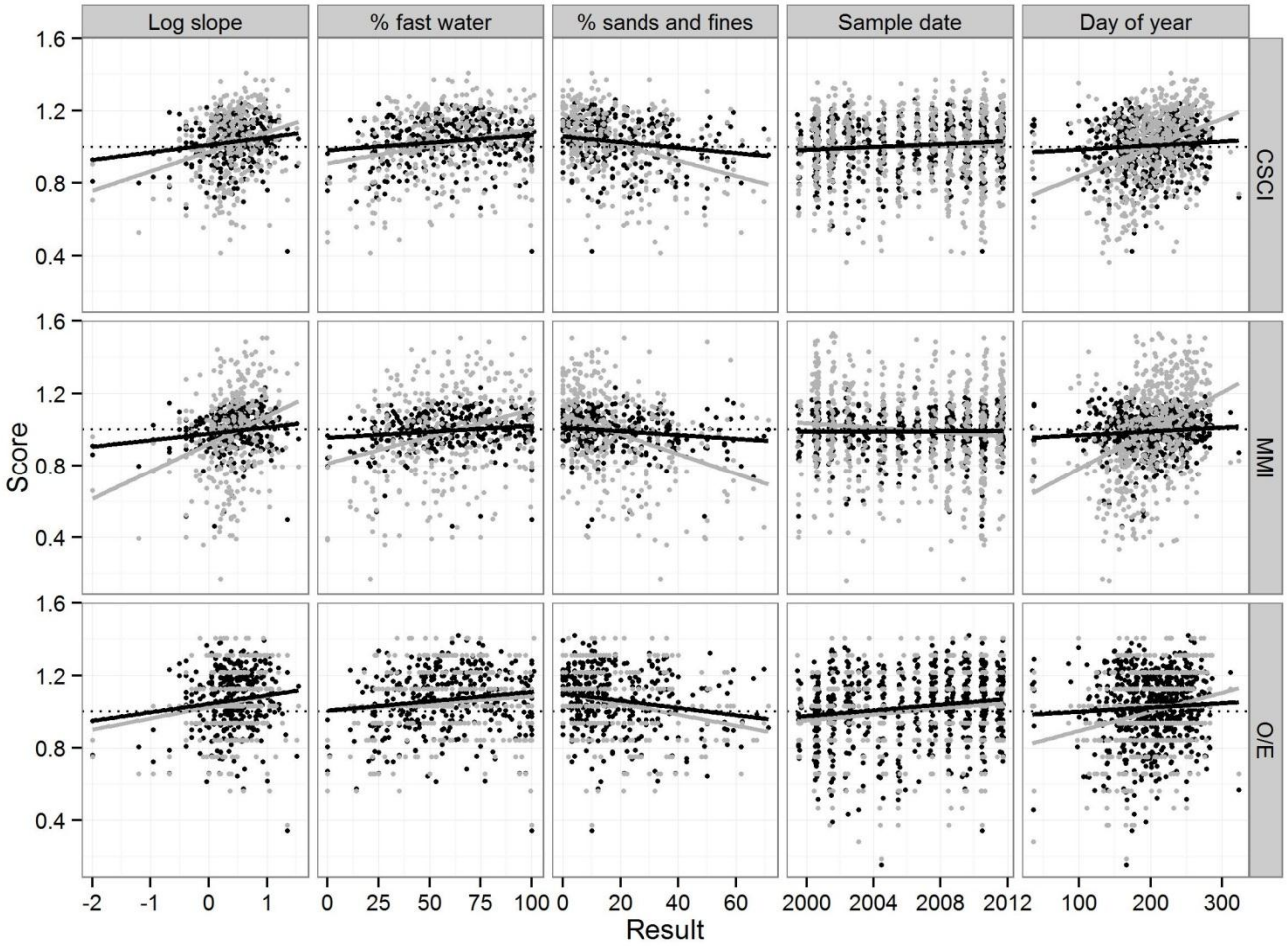

Figure 5. Relationships between scores and three habitat gradients at reference sites for predictive (black symbols and lines) and null (gray symbols and lines) indices. The dotted line indicates a perfect relationship without bias.

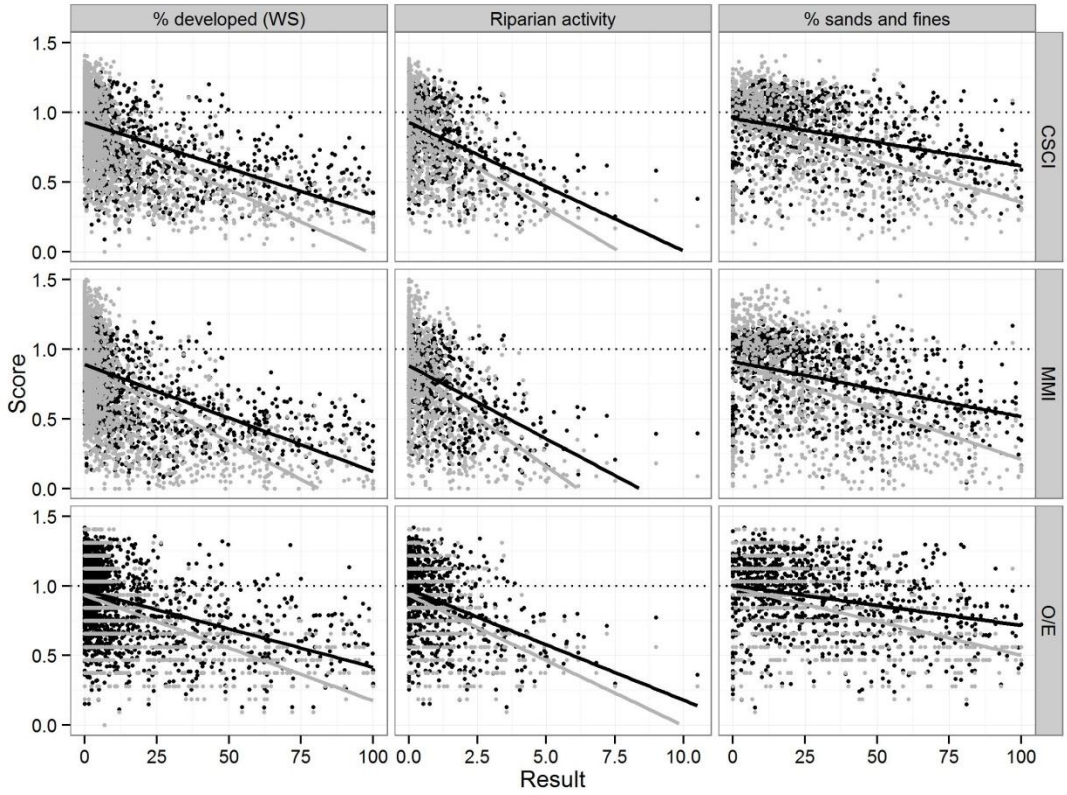

Figure 6. Relationships between scores and selected stressors for predictive (black symbols and lines) and null indices.

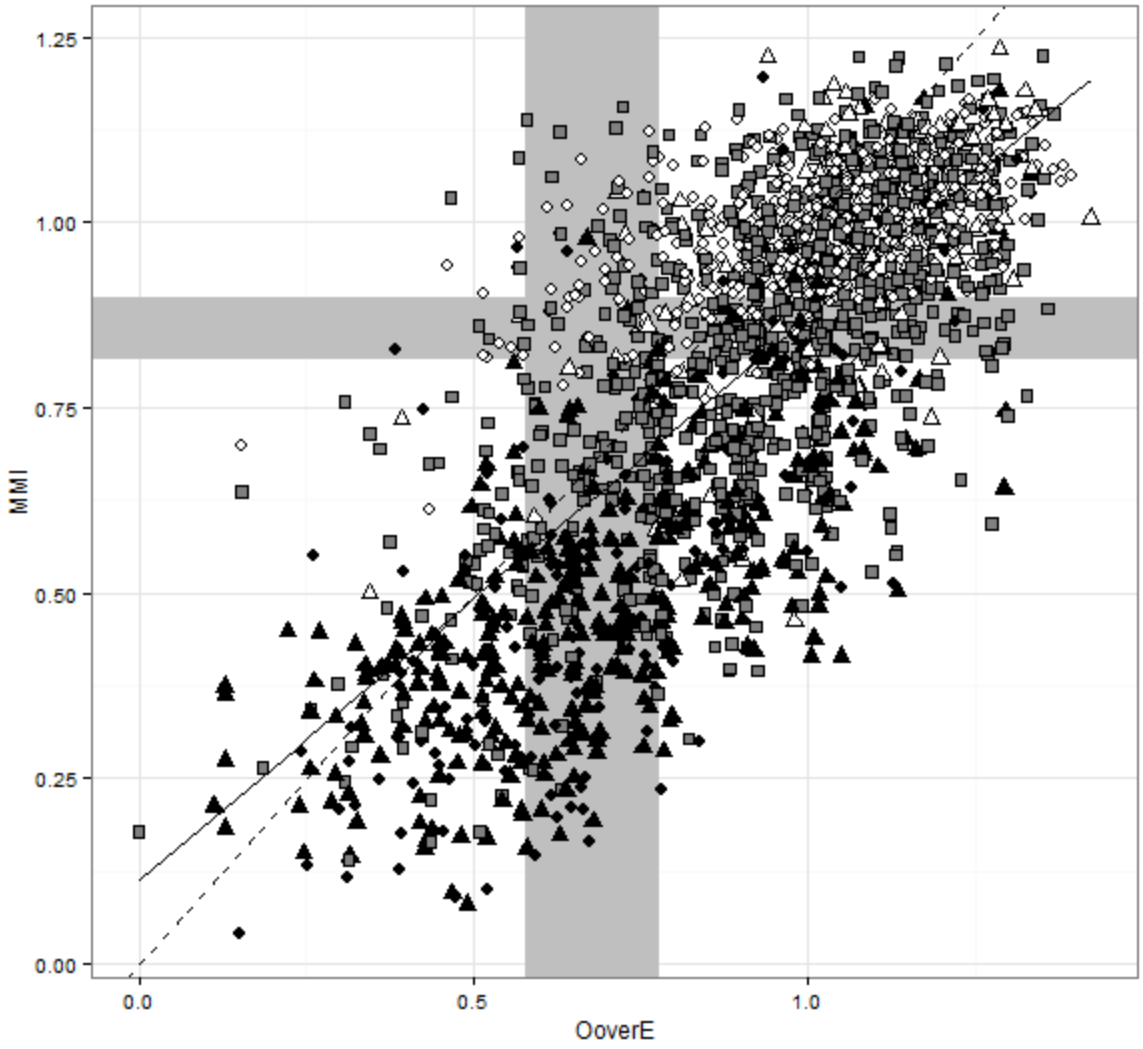

Figure 7. Scores for the O/E versus the predictive MMI. White symbols represent reference sites, black symbols represent stressed sites, and gray symbols represent intermediate sites. Circles represent calibration sites, triangles represent validation sites, and squares represent other sites. The regression line is represented as a solid line, and the line of a perfect relationship is shown as a dashed line. The gray bars indicate regions between the first and tenth percentiles for each index, representing ambiguous areas where agreement was not determined.

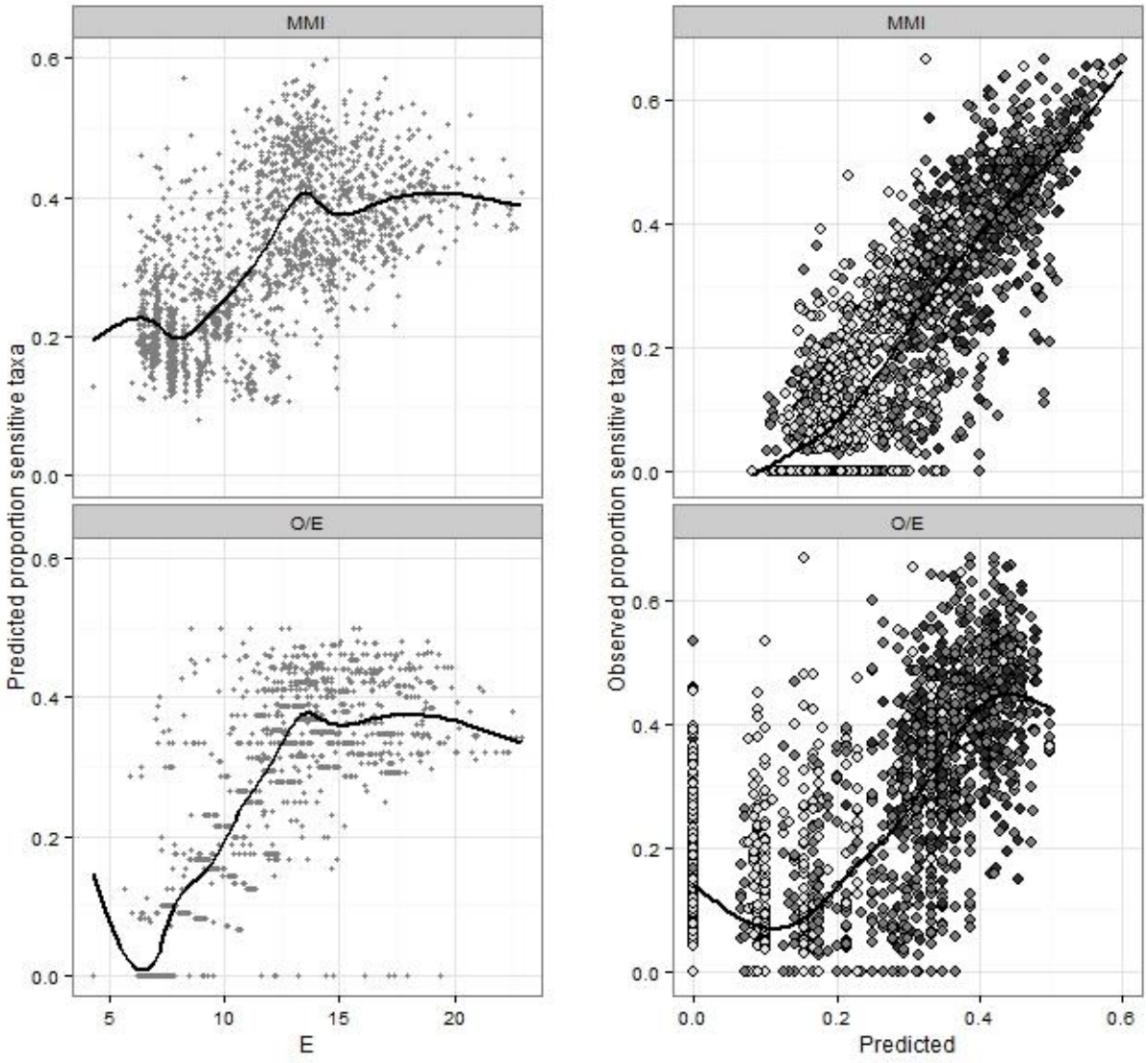

Figure 8. Left panels: Proportion of sensitive taxa expected versus number of expected taxa at all sites. Right panels: Predicted versus observed proportion of sensitive taxa based at reference calibration sites. Dark sybmols represent sites with high (>15) numbers of expected taxa; gray symbols represent sites with moderate (10 to 15) numbers of expected taxa; and white symbols represent sites with low (<10) numbers of expected taxa. The top panels represent predictions based on the MMI, and the bottom panels represent predictions based on the O/E. The solid line represents a smoothed fit from a generalized additive model.

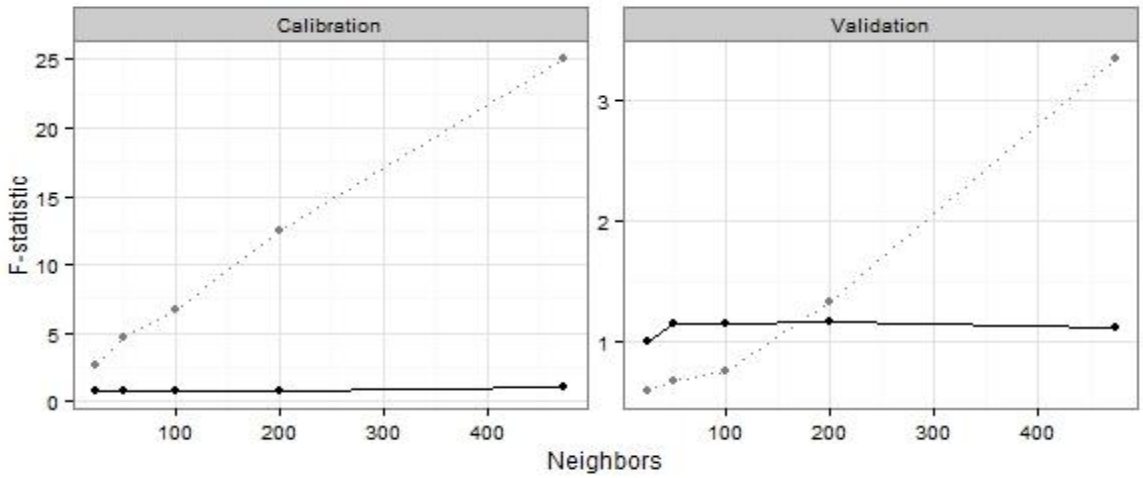

Figure 9. Effects of nearest neighbors thresholds on bias by region. F-statistics are based on ANOVAs of percentile-transformed index scores at reference sites by region. Solid black lines represent results for the predictive index, and gray dotted lines represent results for the null index. Only results for the combined index are presented