

Principles for the Development of Contemporary Bioassessment Indices for Freshwater Ecosystems

Andrew L. Garey and Leonard A. Smock

Contents

1	Introduction	234
2	The Reference Condition Approach	237
2.1	Alternatives to the Reference Condition Approach	238
2.2	Reference Site Screening	239
2.3	Reference Site Classification	240
3	Predictive Modeling of Aquatic Assemblages	244
3.1	The Observed-to-Expected (<i>O/E</i>) Index	246
3.2	Advances in Predictive Modeling	247
4	Index Development and Performance Evaluation	249
4.1	Numerical Range	249
4.2	Accuracy and Precision	249
4.3	Metric Redundancy	252
4.4	Metric Aggregation and Scoring	253
4.5	Index Validation	253
5	Expert Interviews: Challenges and Important Considerations in Bioassessment	254
6	Conclusions	257
	References	259

Abstract Bioassessment can be broadly defined as the use of biota to assess the nature and magnitude of anthropogenic impacts to natural systems. We focus on an important and specific type of bioassessment: the use of ecological assemblages, primarily fish, macroinvertebrates, and algae, as indicators of anthropogenic impairment in aquatic systems. Investigators have long known that biota provide spatially and temporally integrative indicators of impairment. This chapter provides an introduction to the process of developing assemblage-level indices that provide

A.L. Garey (✉) • L.A. Smock
VCU Rice Rivers Center, Virginia Commonwealth University, Richmond, VA, USA
e-mail: gareyal@vcu.edu

quantitative estimates of the ecological integrity of freshwater ecosystems. We discuss important developments made in the latter half of the twentieth century which are still relevant and useful for bioassessment, as well as more recent developments that have improved the effectiveness of bioassessment strategies. Throughout the chapter, we focus on analytical approaches for improving the effectiveness of bioassessment indices for detecting anthropogenic impairment. In the concluding section of the chapter, we widen our perspective and include excerpts from discussions with three expert practitioners on topics that are more broadly applicable to the assessment of the ecological integrity of aquatic systems. The major challenge for all bioassessment programs is to separate the effects of anthropogenic impairment on biota from the effects of natural environmental variability unrelated to impairment. Analytical developments, such as advanced predictive modeling techniques, coupled with emerging technologies and the development of large-scale bioassessment programs will continue to increase our ability to meet this challenge and to improve our understanding of how aquatic assemblages are affected by anthropogenic impairment.

Keywords Aquatic ecosystems • Bioassessment • Biomonitoring • Biotic assemblages • Predictive modeling

1 Introduction

The US Environmental Protection Agency (USEPA) defines biological assessment as the “. . . evaluation of the condition of a waterbody using biological surveys and other direct measurements of the resident biota in surface waters” [1]. Investigations that fall under this broad definition may be focused on any level of biological organization, from studies of subcellular effects of toxic compounds [2] to ecosystem-scale assessments using multiple taxonomic assemblages [3]. The terms biological assessment, bioassessment, biological monitoring, and biomonitoring are often used interchangeably. For clarity, we restrict our discussion to the term bioassessment.

The value of aquatic organisms as pollution indicators has been recognized by scientists for over 100 years. The Saprobien system of Kolkwitz and Marsson [4], most probably the first bioassessment index, was a system for quantitatively rating the tolerance of aquatic organisms to sewage pollution, much akin to modern pollution tolerance values. This concept has been adapted and modified many times, and both the concept and use of the word “saprobity” persist in contemporary literature [5, 6]. The practice of bioassessment invokes the concept of biological integrity, defined as “the capability of supporting and maintaining a balanced, integrated, adaptive community of organisms having a species composition, diversity, and functional organization comparable to that of natural habitat of the region” [7, 8]. Practitioners conducting bioassessments assume that biotic integrity reflects

overall ecological integrity, which describes the state of an ecosystem with respect to biology as well as physical and chemical factors [8]. Therefore, the purpose of using biota to assess environmental conditions is that they integrate the effects of all environmental factors to which they are exposed over their entire life-spans and habitat ranges [8, 9]. However, because biota are responsive to such a multitude of environmental factors acting over multiple temporal and spatial scales, determining clear and unambiguous relationships between biota and anthropogenic impairment remains a challenging and active area of research.

This chapter addresses the development of numerical indices, based on biological assemblage-level data, to make inferences regarding anthropogenic stress to freshwater ecosystems. We follow the framework of Fauth et al. [10] in defining the terms community and assemblage. Communities refer to all organisms within the spatial boundaries of the system of interest. For bioassessment, the spatial boundaries of communities are generally artificial constructs, rather than distinct, natural boundaries, and are chosen based on some combination of scientific, logistical, and political criteria. The term assemblage refers to a taxonomically defined subset of a given community, for example, the benthic macroinvertebrate assemblage of a stream system.

The general objective of all bioassessments is to separate the signal of anthropogenic impairment effects from the noise of effects related to natural variations in space and time that are not related to anthropogenic impairment. Evaluation of the relative importance of these two effects requires measurement or estimation of variables related to anthropogenic impairment, which we refer to hereafter as stressors, as well as those related to natural variation, which we refer to as natural environmental variables.

The assemblages chosen for bioassessments depend on the expertise and resources available to investigators, public interest, and on those that are most expected to respond to anthropogenic stress. Algae, fish, and macroinvertebrates are the most commonly used assemblages, and numerous examples of useful bioassessment indices exist for each. Investigations comparing these assemblages commonly show that they respond differently to anthropogenic stress, and each represents a unique aspect of ecological integrity [11–13]. Therefore, we focus on describing the analytical methods used for the development of contemporary indices, not on comparing the usefulness of different assemblages. We do not address descriptions of field and laboratory methods, but do note that sampling methodology [14, 15], sampling effort [16, 17], and taxonomic resolution [18, 19] have important and well-documented effects on bioassessments. Our focus is on perennial streams and rivers, as these systems dominate the literature and are the focus of most bioassessment programs. We also provide examples from lakes, impoundments, and wetlands when they enhance our discussion. The analytical methods presented here are also applicable to other aquatic systems and assemblage types.

Biological integrity is defined by one or a series of bioassessment metrics, which are quantitatively defined aspects of assemblages that are expected to vary in response to impairment. Some investigators favor the use of multiple metrics,

Table 1 Selected fish, macroinvertebrate, and algae metrics used as indicators of anthropogenic impairment

Metric category	Assemblage	Metric	Impairment response	Description
Composition and diversity	Fish	Percent of nonnative species [20]	Increase	Percent of species that are nonnative
	Macroinvertebrates	EPT richness [21]	Decrease	Number of taxa (genera or species) in the orders Ephemeroptera, Plecoptera, and Trichoptera
	Algae	Similarity to reference [22]	Decrease	Bray–Curtis similarity index of test site compared to reference sites
Trait based	Fish	Percent lithophilic spawners [17]	Decrease	Percent of individuals that spawn on coarse gravel substrate
	Macroinvertebrates	Percent collector–gatherer taxa [23]	Increase	Percent of taxa in the collector–gatherer functional feeding group
	Algae	N-heterotrophic taxa [24]	Increase	Number of taxa that derive nitrogen from the uptake of amino acids
Pollution tolerance	Fish	Tolerance value [25]	Increase	Weighted average tolerance value based on an impairment gradient derived by principal components analysis (PCA)
	Macroinvertebrates	Percent tolerant taxa [23]	Increase	Percent of individuals with tolerance value >6 (10-point scale, increasing with tolerance). Tolerance values derived based on best professional judgment and literature review
	Algae	Alkaliphilous taxa richness [26]	Increase	Number of taxa (genera or species) primarily occurring at pH > 7
Individual condition	Fish	Percent of individuals with anomalies [27]	Increase	Percent of individuals with deformities, erosion, lesions, or tumors
	Macroinvertebrates	Toxic score index [28]	Increase	Index of effluent toxicity based on mentum deformities of <i>Chironomus</i> spp.
	Algae	Percent deformed cells [29]	Increase	Percent of diatom frustules with deformities

which are aggregated within a multimetric index (MMI). MMIs provide a checks-and-balances system to account for variable responses of metrics to multiple stressors. Others prefer single-metric indices, most notably the observed-to-expected (*O/E*) index, which we introduce in Sect. 3.1. Developers of MMIs commonly group metrics based on the general type of ecological information they express. Metrics from different ecological categories are included in MMIs in order to reduce the redundancy of information and increase the explanatory power of indices. In practice, a vast array of assemblage-level metrics has been used for aquatic bioassessments.

To provide relevant examples, we have assembled a short list of fish, macroinvertebrate, and algae metrics, which we group into four broad categories (Table 1). Diversity and composition metrics are taxonomy-based metrics associated with assemblage characteristics such as richness, evenness, diversity, and dominance. Trait-based metrics incorporate information on ecological habits, habitats, morphology, life history, and life cycle characteristics of populations in the assemblage of interest. Pollution tolerance metrics are numerical ratings of the degree to which individuals in the assemblage are tolerant to stressors. Individual condition metrics are associated with visually apparent morphological anomalies of individual specimens. In addition to these general metric types, the absolute abundance of fish and biomass of algae are also sometimes used, although the absolute abundance of macroinvertebrates is rarely used.

In Sect. 2, we introduce the most widely used method for bioassessment, the Reference Condition Approach (RCA) [30–32]. In Sect. 3, we discuss predictive modeling of aquatic assemblages, which is conducted to control for the effects of natural environmental variation in order to obtain an unambiguous determination of anthropogenic effects. Once selected and properly calibrated for natural environmental variation, metrics are used individually, or are aggregated within an MMI, to provide a scoring system that reflects the assemblage-inferred level of anthropogenic impairment at a given study site. This process, as well as methods for evaluating the performance of metrics and indices, is reviewed in Sect. 4. In Sect. 5, we take a broader perspective and present interviews with three experts who provide valuable insights into some of the most important emerging issues and challenges in the field of bioassessment.

2 The Reference Condition Approach

Reference conditions serve as surrogates for negative controls, representing the assemblage characteristics at test sites that would occur in the absence of impairment. Following the RCA, reference conditions are derived from assemblage data at least-disturbed reference sites (*sensu* Stoddard et al. [32]). In comparison to experimental studies, where variation among replicates is carefully controlled and expected to be minimal, variation among reference site assemblages is high and difficult to predict. Replicate samples from the same site are generally insufficient

to account for this variation. Therefore, the RCA calls for the use of multiple reference sites in order to adequately account for the effects of natural environmental variation on assemblages [30–32].

We use the term reference to broadly encompass streams considered to be in least-impaired conditions, though as discussed in Sect. 2.2, the actual level of impairment at reference sites is highly variable among studies [32]. We use the term impaired to refer, in general, to sites that are subject to the deleterious effects of stressors and the term test sites to refer to those of unknown impairment status (i.e., those for which bioassessments are needed). This chapter focuses on the RCA, although alternative approaches for estimating reference conditions may be used when system conditions and data availability warrant (Sect. 2.1). Application of the RCA proceeds by first screening multiple potential reference sites to determine if they reflect appropriate least-impaired conditions (Sect. 2.2), then classifying the screened reference sites as to expected or quantified patterns of variability among their biotic assemblages (Sect. 2.3).

2.1 Alternatives to the Reference Condition Approach

Though not often available, data describing past assemblages may provide valuable information for inferring reference conditions. Investigations of sediment records, historical accounts of landscape conditions, and museum records have been used to infer past environmental conditions and assemblage composition in aquatic systems [33–36]. Historical approaches, while important, lack broad applicability for bioassessment. In lotic ecosystems, sediment deposition is generally insufficient to provide a historical record. Moreover, data that describe assemblage composition before anthropogenic development occurred may reflect conditions that are no longer attainable given the effects of factors acting at large spatial and temporal scales such as atmospheric deposition of pollutants and global climate change.

When anthropogenic impacts are spatially discernable, a paired-site approach may be useful. For example, lotic sites impacted by point source pollution such as mine effluent [37, 38] or municipal wastewater [39] may be paired with upstream sites above the source of stressors. Plafkin et al. [40] referred to paired upstream sites as controls, though this implies that confounding natural environmental factors on the upstream–downstream comparison are being actively controlled, which is generally not feasible. For bioassessments conducted over large spatial scales (e.g., ecoregions), the paired-site approach is problematic because much assemblage variation is driven by local-scale differences in natural environmental variables [41, 42]. Assemblage variation among reference sites, that is, variation that is not likely caused by impairment, is typically much larger than variation among replicate samples collected at a single site. Therefore, comparisons of replicate samples from a single reference and test site pair often do not provide a realistic representation of the effects of impairment on assemblage characteristics [43].

2.2 Reference Site Screening

Reference site screening is the process of estimating the degree of anthropogenic impairment at study sites, such that those with acceptably low stress levels may be designated as reference sites. Screening criteria vary among studies based on both data availability and on the opinions of investigators as to which criteria are most appropriate. Some have advocated for the use of professional judgment regarding whether observed assemblages represent reference conditions [44, 45]. Professional judgment may also be applied to the environmental conditions at the sites, providing a more independent, and potentially less biased, means of estimating impairment status [46]. Advocates of professional judgment often favor consensus opinions among groups of experts [45, 47]. Potential criticisms of such approaches include a lack of empirical support for decisions regarding reference designations and circular logic when sites are designated as reference based on the assemblage that is also used for bioassessment. However, Davies and Jackson [45] recently showed that the opinions of trained experts were highly consistent when rating ecological integrity based on assemblage data, leading the authors to assert that well-informed professional opinions provide reliable, ecologically relevant benchmarks for bioassessment.

Independent stressor variables (i.e., those not related to the assemblage used for bioassessment) are often used to estimate impairment status. Screening approaches that employ both professional judgment and independent stressor variables produce reference sites that are of higher ecological integrity than those selected using either approach alone [48]. Variables used to estimate anthropogenic stress include physicochemical water quality parameters, land-cover features derived using a geographic information system (GIS) that indicate development, and physical habitat quality assessments based on site observations. Studies often employ a filtering approach, whereby threshold levels for each measured stressor variable are set to designate sites as reference or impaired [21, 49]. The filtering criteria used to select reference sites are generally evaluated with an all approach for reference sites and any approach for non-reference sites. For example, Blocksom et al. [49] required that sites meet all of the reference criteria listed in Table 2 to be designated as reference, but considered sites impaired if any one of the impaired criteria was met.

The level of anthropogenic impairment considered acceptable varies greatly among studies, and many authors fail to provide clear descriptions of how reference conditions are defined. Recognizing this problem, Stoddard et al. [32] advocated for the use of the terms minimally disturbed condition (MDC) to describe expected conditions in the absence of substantive anthropogenic disturbance, least-disturbed condition (LDC) to describe the best available conditions present under current disturbance regimes, and best attainable conditions (BAC) to describe the expected conditions when all avoidable sources of anthropogenic influence are removed (BAC is generally intermediate between LDC and MDC). MDC sites, within virtually unimpacted, near pristine watersheds, are rarely available, and most

Table 2 Filtering criteria used by Blocksom et al. [49] to differentiate between reference and impaired streams in the Mid-Atlantic Highlands (USA)

Criterion	Reference	Impaired
pH	NA	<5
Sulfate	<400 µeq/l	>1,000 µeq/l
Acid neutralizing capacity	>50 µeq/l	NA
Chloride	<1,000 µeq/l	>1,000 µeq/l
Total phosphorus	<20 µeq/l	>100 µeq/l
Total nitrogen	<750 µeq/l	>5,000 µeq/l
Mean RBP score	>15	<10

RBP refers to the rapid bioassessment habitat protocol of Barbour et al. [1], where scores range from 0 to 20

It is possible for sites to have intermediate characteristics and therefore fall between the two classes

NA no filter specified

often reference sites are chosen based on the best available conditions (LDC sites; e.g., [50, 51]).

The reliability of bioassessments depends largely on the existence of a sufficient number of reference sites to encompass the range of natural variability in the study region. Given the pervasiveness of human impacts on aquatic ecosystems, achieving a sufficient number of reference sites is often difficult, and may be impossible if standards regarding the acceptable level of impairment are unrealistically high [50, 51]. Impairment is highly variable among geographic regions because development pressure is nonuniform. Recognizing this, Yates and Bailey [50] developed a novel strategy for selecting reference sites that allowed for flexibility in the standards used for screening to select LDC sites within study regions exposed to different impairment regimes. This is a pragmatic strategy, as flexibility in the level of impairment allowed is unavoidable in areas where impairment is pervasive; however, clear comparisons of bioassessment results among studies are hindered when the standards used to select reference sites vary.

2.3 Reference Site Classification

Reference sites that successfully pass the screening process are used to predict assemblage conditions under minimal impairment. Classification of reference sites is intended to increase the precision and accuracy of these predictions by grouping sites inhabited by similar ecological assemblages. Broadly, there are two major types of classification systems: (1) typologies, wherein sites are grouped based on spatial proximity and/or similarity in their natural environmental variation, and (2) biotic classifications, which employ statistical analyses on assemblage data to group sites. We begin by discussing the two as distinct approaches, although bioassessment programs often use elements of both to develop the best site classifications. Throughout this section, we discuss analytical techniques that are

specialized for the analysis of assemblage data. McCune and Grace [52] and Legendre and Legendre [53] present additional details on most of the analytical methods presented here. Software for conducting most of the techniques is available free of cost in the R statistical programming language [54].

The objective of classification is to maximize assemblage similarity within classes, while maintaining sufficient replication to allow for statistical comparisons of test sites with the reference classes. Many measures of assemblage similarity exist [52, 53]. The Bray–Curtis coefficient [55] for abundance data and the Sorenson coefficient [56, 57], the equivalent of Bray–Curtis for presence-or-absence data, are most commonly used. Both are well suited to the numerical structures of assemblage datasets. Similarity may also be calculated using bioassessment metric values in place of taxonomic data (a sound but under-used technique, [58]).

Similarity is summarized as the mean within-class similarity (W , the mean of all pairwise similarities of sites within classes) and the mean between-class similarity (B , the mean of all pairwise similarities of sites not in the same class). The precision of a classification is described by the relationship of W to B , referred to as the classification strength [58]. High classification strength is indicated by a large positive difference or large ratio of within- to among-class similarity (i.e., high $W-B$ or W/B). Predictions regarding assemblage conditions are most reliable when classification strength is high. Multivariate techniques such as MEANSIM [58], analysis of similarity [59], multiresponse permutation procedure [60], and nonparametric, multivariate analysis of variance [61] are used to test the hypothesis that classification strength is higher than expected by chance, providing an indication of whether the classification improves the reliability of predictions regarding reference conditions.

2.3.1 Typological Site Classification

Typological site classifications are based on a priori judgments regarding the site conditions that best group reference sites with similar assemblages. Early typological approaches focused on coarse-scale, map-based classifications (e.g., ecoregions) [62, 63]; however, typological classifications that do not account for the effects of local-scale variables typically exhibit much lower classification strength than biotic classifications [43, 64, 65]. Typological classifications are a convenient and useful tool that should be at least considered as an initial step toward site classification [65, 66]. Like all classifications, the effectiveness of a priori-defined typologies should be assessed by a posteriori, quantitative evaluations of the assemblages of interest [66]. For example, investigators in Virginia (USA) observed a striking difference in stream macroinvertebrate assemblage structure between low-gradient coastal plain sites and upland piedmont and mountain sites, requiring the use of separate bioassessment indices for coastal and non-coastal sites (Fig. 1) [67, 68]. Because assemblages are affected by both regional- and local-scale environmental factors, typological classifications that consider smaller spatial-scale variables as well as large-scale zones may provide comparable, or greater, classification strength than biotic classifications [64, 69].

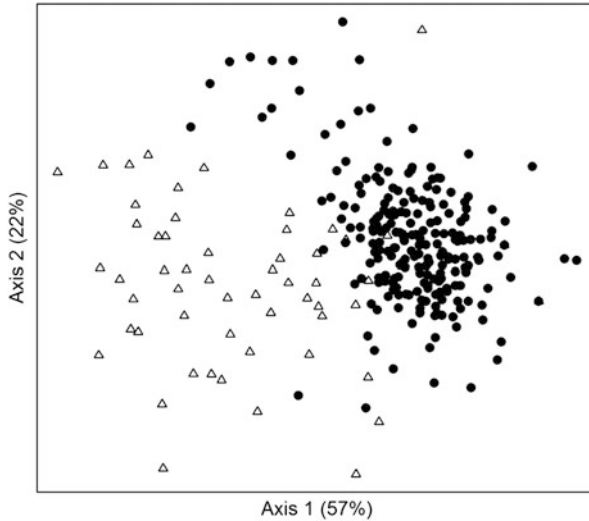


Fig. 1 Nonmetric multidimensional scaling ordination of macroinvertebrate assemblages, identified at the family level, from 269 least-impaired, reference Virginia stream sites. *Symbols: (open triangles) coastal plain sites; (filled circle) non-coastal sites.* Distances between sites correspond to their proximity in Bray–Curtis distance space. Percentages indicate the percent of variance in the Bray–Curtis coefficients explained by the axis coordinates. Adapted from Dail et al. [67] with permission

2.3.2 Biotic Site Classification

Biotic classification of reference sites came to prominence with the introduction of the River Invertebrate Prediction and Classification System (RIVPACS, [70, 71]). Agglomerative cluster analysis is used in many variations of the RIVPACS approach, including those developed for the USA [72] and the Australian River Assessment Scheme (AUSRIVAS, [73]), both of which employ presence/absence data for clustering, and the Benthic Assessment of Sediment (BEAST) method of Canada [74] which clusters based on abundance data. Agglomerative clustering proceeds from the bottom up, progressively grouping sites of increasingly dissimilar taxonomic composition. Most investigators cluster sites based on Bray–Curtis or Sorenson dissimilarity [72–76], although other measures such as Euclidean distance can be employed [77]. Figure 2 shows a cluster analysis of 46 Kentucky (USA) stream reference sites where genus-level macroinvertebrate data were collected by the USEPA.

The standard RIVPACS approach classifies sites using two-way indicator species analysis (TWINSPAN, [79]). In contrast to agglomerative cluster analysis, TWINSPAN is a divisive technique, whereby sites are progressively divided based on taxa (indicator species) that best differentiate them. Also unlike agglomerative clustering, the user cannot choose a dissimilarity measure [80]. TWINSPAN has

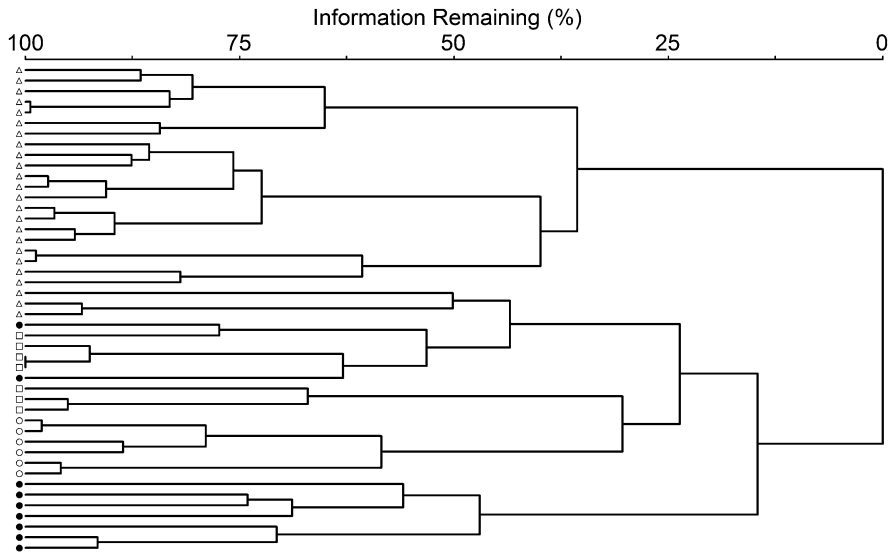


Fig. 2 Agglomerative cluster dendrogram generated using the flexible-beta method ($\beta = -0.30$) on a Bray–Curtis dissimilarity matrix of genus-level macroinvertebrate data at 46 least-impaired Kentucky stream sites. Branch lengths correspond to dissimilarities between sites and clusters. Bioregions are regional classifications as described by Pond et al. [78]: (*open triangle*)—Mountain; (*filled circle*)—Miss. Valley-Interior River; (*open square*)—Pennyroyal; (*open circle*)—Bluegrass. The Mountain and Bluegrass regions separate perfectly. Some overlap occurs for other bioregions because Julian day and latitude (not included in the analysis) were also important variables related to assemblage structure in these bioregions. Data courtesy of Gregory Pond, USEPA

received criticism for poor performance and seemingly arbitrary methodology [52, 81]. However, based on comparisons with other techniques, the developers of RIVPACS concluded that TWINSpan performed well, and the method is still used within the current RIVPACS framework [80].

Relationships among sites can be visualized using a variety of ordination techniques, which reduce the n -dimensional hyperspaces created by ecological distance matrices to fewer (usually 2 or 3) dimensions that best explain the overall pattern of variability (Fig. 1) [52, 53]. When assemblage–environment responses are assumed to be linear, principal components analysis (PCA) is commonly used, whereas when responses are assumed to be unimodal, reciprocal averaging-based techniques such as correspondence analysis (CA) and detrended correspondence analysis (DCA) are often used. We agree with others [52, 82] in preferring nonmetric multidimensional scaling (NMS) to these techniques because NMS includes no assumptions regarding the underlying data distribution and is highly effective at explaining assemblage structure while reducing dimensionality. Ordinations are often used for exploratory purposes, for example, to confirm classifications made using other analyses [76], but also may be used directly for site classification [67, 83].

Multivariate analyses on assemblage data aid in showing the user where distinctions between classes may occur. However, subjectivity in drawing distinctions among classes is unavoidable, as investigators must decide on the appropriate level of similarity at which to consider sites within the same class [84]. The final decision is made as a compromise between including as many reference classes as possible, while still including enough replicate sites within classes to adequately represent within-class assemblage variability among sites. Bowman and Somers [85] recommend “a minimum of 20, but preferably 30–50 reference sites per group,” though this may be overly optimistic given the data constraints experienced in many studies.

2.3.3 Conclusions Regarding Site Classification

Hawkins et al. [43] and Melles et al. [86] draw distinctions between geography-based methods in which reference classes can be clearly delineated within discrete spatial units and geography-independent methods driven by patterns in assemblage variation regardless of physical location. Typological classifications which include map-based delineations of classes are geography dependent, whereas biotic classifications, focused on patterns of assemblage variation, are geography-independent. However, the most effective classifications consider both geography-dependent and geography-independent factors, for example, limitations on the spatial scale over which biotic classifications are developed can increase their classification strength [87]. While biotic classifications provide precise descriptions of the patterns of variability with respect to the assemblage of interest, the resulting classifications may not be applicable to other assemblages. Inclusion of geography-dependent variables that implicitly encompass a wide range of environmental factors provides a more comprehensive classification of sites [88]. A priori typological classification based on large-scale variables (e.g., ecoregions) provides useful, convenient, and easily communicated initial classifications of sites, though classifications are often improved when supplemented by smaller-scale variables that are not spatially discrete (e.g., flow regime; [89]) or not associated with geography (e.g., sampling date; [75]). Good scientific practice requires that the effectiveness of a priori approaches be evaluated with a posteriori evaluations of relationships between classes and biota [64–67, 88].

3 Predictive Modeling of Aquatic Assemblages

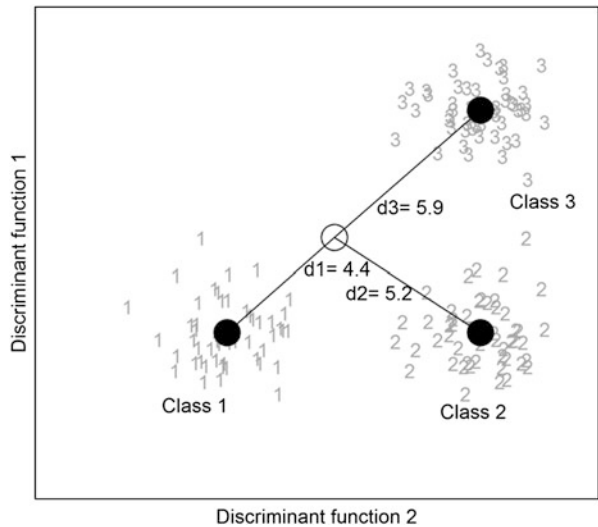
The objective of predictive modeling for bioassessment is to control for the confounding effects of natural environmental variables so that the effects of stressors on metrics can be clearly evaluated. The methods used to meet this objective are as diverse and varied as the assemblages themselves. As an introduction to the core concepts in predictive modeling, we outline the basic steps of the

Table 3 Example calculation of *O/E* based on biotic classes and distances in environmental space shown in Fig. 2

Reference class	Distance from test site in discriminant space	Class probability of test site	Probabilities of capture at reference sites		
			Taxon A	Taxon B	Taxon C
1	4.4	0.47	0.70	0.20	0.10
2	5.2	0.31	0.70	0.25	0.05
3	5.9	0.20	0.20	0.50	0.10
Probability of capture at test site:			0.59	0.29	0.09
Expected richness at test site (<i>E</i>):			0.97		

Data were simulated for example purposes only
 Distances were converted to class probabilities following equations 2 and 3 in Clarke et al. [90] (distances are D^2 values as described in [90])
 All classes contain 20 sites
 The probability of capturing a taxon at the test site is given by multiplying the class probabilities of the test site for each reference class by the corresponding probabilities of capture of the taxon at sites within each reference class
 For example, the probability of capturing Taxon A at the test site is $0.47 \times 0.70 + 0.31 \times 0.70 + 0.20 \times 0.20 = 0.59$
 The probabilities of capture for each taxon are summed to give the expected richness (*E*) at the test site

Fig. 3 Reference and test sites displayed in environmental distance space as defined by two discriminant function axes. Discriminant function scores were simulated from normal distributions with variance = 1. Symbols: (filled circle) reference class centroids, (open circle) test site, (1, 2, 3) reference sites, (d1, d2, d3) distances in discriminant function space between test sites and reference class centroids



RIVPACS framework and present an example of its use in Table 3 and Fig. 3. Although now 30 years old, the framework remains relevant and is still used with little modification of the original methodology [70, 75, 80]. The conceptual basis and technical details of RIVPACS have been described thoroughly by others [71, 72, 90]. Following our introduction to RIVPACS, we present several promising recent advancements in predictive modeling.

3.1 *The Observed-to-Expected (O/E) Index*

The standard metric derived from RIVPACS-type predictive modeling approaches is the observed-to-expected ratio (O/E), a single-metric index that compares the observed taxonomic richness at a study site to the expected richness under minimally impaired conditions. The O/E ratio indicates the degree of “taxonomic completeness” (sensu Hawkins [91]) of the test site. O/E values less than one indicate that taxa expected to be present if the sites were unimpaired are absent.

Following biotic classification of reference sites, RIVPACS employs Multiple Discriminant Analysis (MDA) to develop linear functions that best describe the relationships of natural environmental variables to the biotic classes. The discriminant functions are used to determine the distance, in environmental variable space, of the test site to the biotic reference classes, which in turn are used to estimate the probabilities that the test site belongs in each reference class (referred to here as class probabilities). For all native taxa in the study region, the proportion of reference sites within a given biotic class where a taxon is present represents the probability of observing that taxon at a site in that class (referred to as the probabilities of capture). The probabilities of capture of a given taxon within each reference class, and the class probabilities of the test site for each reference site, are used to estimate the probability of capturing the taxon at the test site assuming unimpaired conditions. The expected richness at the test site (the E in O/E) is given by summing probabilities of capture at the test site for all taxa (see Fig. 3 and Table 3 for additional details).

O/E values greater or less than one indicate departures from what is predicted under unimpaired conditions. Simpson and Norris [92] recommended that O/E values below the tenth percentile of the reference site distribution indicate impairment, with the extent of impairment increasing as the ratio decreases. They also postulated that O/E values greater than one may indicate areas of exceptionally high natural biodiversity or those subject to mild impairment that artificially increases richness.

A common modification to the basic framework is to exclude rare taxa from the analysis, as their inclusion can result in a site receiving an O/E score near one when the assemblage observed deviates considerably from statistical expectations. Several authors have indicated that excluding taxa with probabilities of capture less than 0.5 (producing the $O/E_{0.5}$ index) improves accuracy and precision [72, 93, 94]. As an alternative, Van Sickle [93] adapted the Bray–Curtis dissimilarity measure to compare observed and expected assemblages (referred to as BC) and showed that BC was generally more accurate than O/E for identifying impairment across a wide range of assemblages and study systems. $O/E_{0.5}$ and BC indices developed for Appalachian stream macroinvertebrate assemblages exhibited similar accuracy and precision [75].

3.2 *Advances in Predictive Modeling*

The widespread application of RIVPACS-type models has inspired many alternative approaches. Recognizing that assemblages occur along continuous environmental gradients, investigators have developed nearest neighbor methods that compare the environmental similarity of test sites to each individual reference site, rather than to the average assemblage of each class as is done using RIVPACS [83, 95]. Modeling approaches often skip the biotic classification step and predict assemblage characteristics at reference sites directly using natural environmental variables [96–98]. Direct prediction approaches may allow for different sets of environmental variables to be used as predictors for each taxon. Though appealing in this respect, the development of separate models for each taxon may be overly complex for taxon-rich systems.

In contrast to the long history of predictive modeling for *O/E* indices [70], until recently, developers of MMIs rarely employed predictive modeling to account for natural environmental variability. McCormick et al. [99] used linear regression to control for the effects of watershed size on a fish MMI. Equations derived from the regression of metric values on watershed size at reference sites were applied to test sites, and the residuals from the regression were used to indicate deviations from the expected metric values in the absence of impairment. Oberdorff et al. [100] expanded this approach, modeling metrics based on a suite of natural environmental variables using logistic regression (for presence/absence metrics) and multiple linear regression (for abundance-based metrics). Variations on this residualization technique have been developed for more advanced modeling strategies such as prediction tree approaches (discussed below), improving both the accuracy and precision of MMIs by removing the confounding effects of natural environmental variables [21, 101, 102].

Although conventional techniques such as MDA and linear and logistic regression have provided utility for predictive modeling, several newer methods better account for the variable, often nonlinear and interactive effects of environmental predictors on biota. The generalized additive modeling approach of Yuan [103] shows the flexibility of this nonparametric regression technique for predicting variable responses among different taxa to a suite of environmental factors. Bayesian frameworks provide a comprehensive evaluation of uncertainty in predictive models [104–106] and have been used for this purpose in MMI development. Machine learning techniques, including artificial neural networks (ANNs) and ensemble prediction trees, where models are iteratively trained at prediction to minimize error, have received much recent attention for predictive modeling in ecology [107–110]. Though the method is not yet widely used, support vector machines have performed favorably compared with other machine learning techniques for predicting the occurrence of macroinvertebrate taxa [111, 112].

ANNs structure predictor–response relationships in a manner similar to vertebrate neurological systems. Variables are represented as neurons connected by a multitude of axons representing the possible interrelationships among variables

[113]. ANNs have shown substantial improvement over traditional RIVPACS-type models for predicting the richness of macroinvertebrate and fish assemblages [97, 114].

Prediction tree approaches such as classification and regression trees account for the complex effects of both continuous and categorical predictors by recursively bisecting the dataset into groups that are increasingly similar with respect to the response variable after each division [52, 53, 115]. Ensemble prediction tree approaches such as random forests and boosted regression trees combine the results of hundreds to thousands of trees to reduce prediction error. Random forests have been used to model assemblage metrics directly [21, 101, 116] and to define relationships between environmental variables and predefined biotic classes, effectively replacing MDA as used in RIVPACS [12, 101]. Comparisons of random forests to boosted regression trees, a related ensemble tree method, indicate that the latter may provide superior performance [108, 117].

In the absence of suitable reference sites, investigators have used whole-set approaches that employ all sites in the dataset, rather than only reference sites, to control for the effects of natural environmental variables. Most whole-set approaches involve the use of regression techniques to model the responses of metrics to stressors and then to estimate metric values at the point where the model estimates that no impairment occurs [118, 119]. Because few to no minimally impaired sites are included in these analyses, they are effectively estimates by extrapolation of a stressor–response gradient and therefore may be subject to greater prediction errors than models for which reference sites are available. Such errors, however, may be unavoidable when test sites cannot be matched with comparable reference sites. As an alternative whole-set approach, Chessman and Royal [120] estimated the tolerance limits and preferences of macroinvertebrates to substrate, temperature, and flow conditions across an extensive dataset of Australian rivers. These limits were then used to predict the presence of taxa and derive *O/E* values at test sites, which exhibited stronger correlations with stressor gradients than *O/E* values derived using the AUSRIVAS method.

A case for using the whole-set approach as a replacement for the RCA was recently presented [121]. Data simulations were conducted to model scenarios in which natural environmental variables and stressors affected biotic metrics independently and also interactively. Metrics that were model-adjusted using the whole-set approach exhibited more accurate and precise relationships with the simulated stressor gradient than metrics adjusted using the RCA. The difference in performance was greatest when stressors and natural environmental variables interacted, as the RCA cannot account for such interactions. While the authors present a compelling case, additional field-based empirical comparisons of the whole-set approach to the RCA are needed.

4 Index Development and Performance Evaluation

We begin this section by discussing methods for evaluating the performance of metrics and indices (Sects. 4.1–4.4). For MMIs, these characteristics should be evaluated in order to include the best-performing metrics in the final index. Final metric selection, scoring, and aggregation are discussed in Sect. 4.5. After scoring and aggregation of metrics within an MMI or alternatively the development of an *O/E* index, performance should be re-evaluated using the finished index scores, ideally using independent data not used for index construction (Sect. 4.5). Further information on MMI development has been presented by others [1, 32, 122]. For clarity, these works present index development in a stepwise manner; however, it is important to note that index development is an iterative, rather than a linear process. Metrics that are acceptable based on one criterion (e.g., numerical range, Sect. 4.1) may subsequently be considered unacceptable based on another criterion (e.g., accuracy, Sect. 4.2), requiring the evaluation of new metrics.

4.1 Numerical Range

Assemblage data are often plagued with abundant zeros due to the patchy distribution of biota among habitats, and metrics related to rare taxa typically have narrow numerical ranges. Metrics with limited ranges, and those for which many sites in the dataset exhibit the same value, are unlikely to exhibit clear numerical responses to stressors [123]. Others have presented guidelines for acceptable numerical ranges for metrics, though these vary among studies [123–125]. Simple distribution plots of metric values often provide clear indications of highly limited metrics (e.g., see Fig. 2 in [122]).

4.2 Accuracy and Precision

We broadly define accuracy as the degree to which a given metric or index is quantitatively related to variations in anthropogenic stress. Accurate metrics and indices exhibit low Type II error rates by correctly identifying impairment and low Type I error rates by correctly identifying reference conditions. As others have indicated [126], the precise impairment state of a system, and therefore the absolute accuracy of metrics, can never be truly known. We therefore use the term accuracy to refer to estimated accuracy for identifying impairment, as indicated by relationships of metrics with a priori-selected stressor variables.

Relationships between metrics with continuously varying stressor variables may be expressed using correlation analysis. The objective is often to assess the responsiveness of metrics to overall stressor gradients. PCA is commonly used

to aggregate individual stressor variables into a comprehensive index of impairment. Metrics and indices can then be evaluated as to the strength of their correlation with this aggregate stressor index [124, 127, 128]. Tests of whether metrics differ significantly from reference conditions provide binary results regarding whether a metric classifies sites correctly. Standard approaches such as ANOVA, however, may indicate statistically significant differences that are not biologically meaningful [123, 129]; therefore, specialized techniques have been developed that are more practical for determining whether metrics differ from reference conditions [129–131]. The magnitude of the departure from reference conditions is most important, regardless of statistical significance. To this end, test statistics such as ANOVA F -statistics or t -scores, rather than p -values, are used to determine the degree to which metrics differentiate between reference and stressed sites [116, 123, 125]. Estimates of Type II error rates are given by choosing a threshold value in the reference site distribution that indicates impairment (e.g., 5th or 25th percentile for metrics that decrease with stress and the 95th or 75th percentile for those that increase with stress) and determining the proportion of impaired sites where metric scores exceed this threshold (for metrics that increase with stress), indicating that impairment has not been correctly identified [132]. Barbour et al. [133] developed a similar, graphical approach for evaluating the degree to which metrics and indices discriminate between reference and impaired conditions (Fig. 4). Distribution-based methods such as these are especially susceptible to the confounding effects of outliers, which should be carefully scrutinized to determine whether they are caused by imprecise metrics or site misclassification.

Measures of precision describe the reliability of metrics and indices for consistently indicating site conditions. Those that exhibit high variability that is not attributable to environmental predictors are not useful for bioassessment. Precision is expressed by measures of variability in metric or index values among samples, most commonly as the standard deviation (SD) or coefficient of variation (CV). Variance partitioning is conducted to determine the relative importance of the three primary sources of variation: among-site spatial variation, within-site spatial variation, and temporal variation [29, 134].

Temporal precision is often evaluated using the signal-to-noise ratio (S/N) [135], which is the ratio of metric variance among sites to variance among multiple visits at the same site. When evaluated using both stressed and reference sites, S/N reflects both accuracy and temporal precision. Stevenson et al. [125] set $S/N > 2$ as the acceptable ratio for diatom metrics. Stoddard et al. [123] indicated that acceptable S/N values should vary, based on organisms' generation times, from >1 for algae to >4 for fish (though these preliminary guidelines require further evaluation). Within-site spatial precision is reflected by metric variability in samples collected at the same site and time, which may be affected by sampling error among spatially or temporally replicated samples [134, 136] or by variation among bioassessments employing different protocols [134, 137].

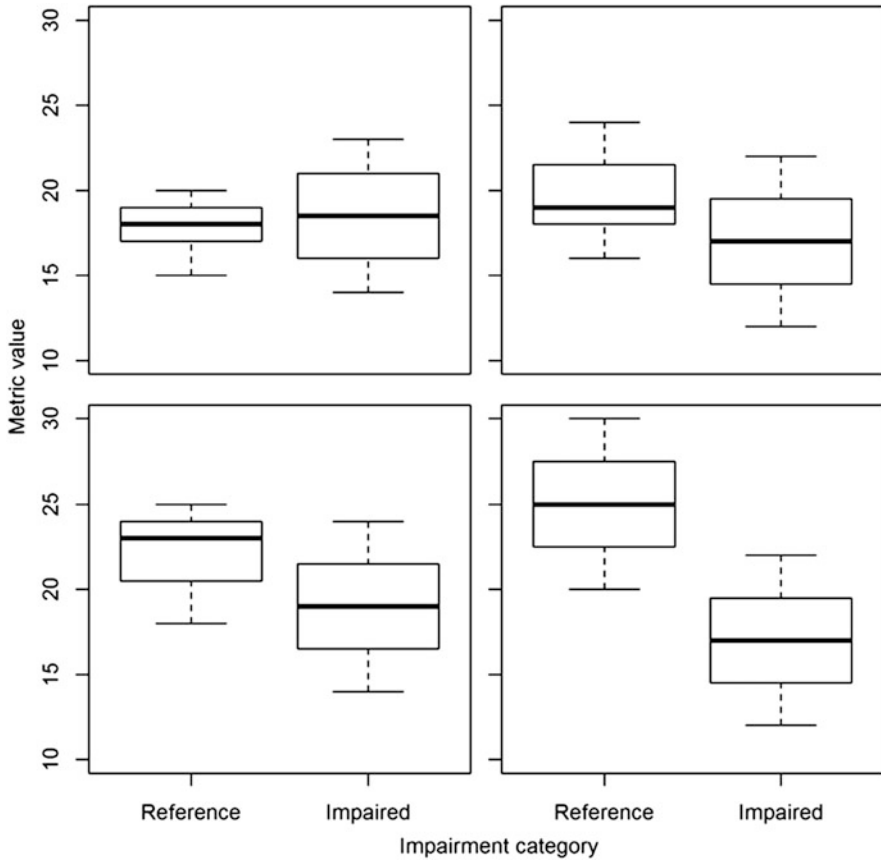


Fig. 4 Box plots of simulated data illustrating the method of Barbour et al. [133] for evaluating the discriminatory power of metrics and indices. Boxes represent 25th and 75th percentiles; whiskers represent non-outlier maximum and minimum values. The metric is expected to decrease with impairment. *Top left:* discriminatory power = 0 (lowest), as the reference and impaired site interquartile ranges (IQRs) overlap and include both medians. *Top right:* discriminatory power = 1 because the IQR overlap includes only one median value. *Bottom left:* discriminatory power = 2 because the overlap does not involve either median. *Bottom right:* discriminatory power = 3 (highest) because the IQRs do not overlap

The among-site precision of metrics and indices may be assessed among sites within the same reference class, thus limiting the confounding effects of environmental variability on the evaluation. For both *O/E* indices and MMIs, the ratio of observed-to-expected metric values at reference sites should differ negligibly from one, and thus the standard deviation (SD) for this ratio should be nearly equal to its coefficient of variation (CV). The distribution of observed-to-expected metric or index values provides a graphical illustration of both accuracy and precision (Fig. 5) [72].

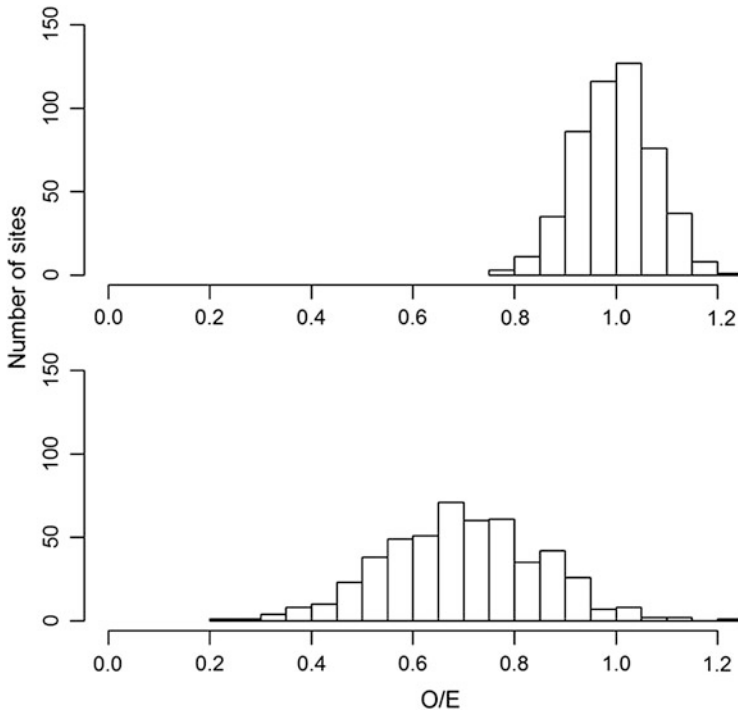


Fig. 5 A graphical comparison of the accuracy and precision of two O/E indices (after methods in [72]). Figure panels depict frequency distributions of O/E values for two indices at reference sites. The data were simulated for example purposes only. The O/E index in the *top panel* is relatively accurate (mean = 1.01) and precise (SD = 0.08). In contrast, the index depicted in the *bottom panel* is less accurate (mean = 0.70) and less precise (SD = 0.16, also note the greater spread of the distribution). This approach can be modified to assess the value distribution of any metric or index at reference sites (e.g., [125])

4.3 Metric Redundancy

Redundant metrics may respond in the same manner to stress and if used together result in overly complex indices that include metrics that do little to increase (and may decrease) accuracy. The problem of redundancy has long been recognized, although the best approach for minimizing it remains unclear. Some investigators prefer to focus on ecological redundancy by including metrics from different ecological categories [122, 128], while others focus on reducing statistical redundancy by evaluating pairwise correlations among metrics and choosing only one metric within each pair that is correlated [124]. Combinations of these approaches may be employed, which consider both ecological and statistical redundancy [138].

Correlations among metrics generally reduce MMI precision and accuracy, but these characteristics appear to be most related to the mean pairwise correlation rather than the maximum correlation among metrics in an index [139]. To best

reduce the mean correlation among a group of potential metrics, multivariate analyses such as PCA and cluster analysis can be used to aggregate correlated metrics [21, 101, 102]. The correlation of metric errors (e.g., residuals of stressor–metric regressions) rather than the correlation of metric values may be more appropriate for judging redundancy, a concept whose applicability should be further evaluated [140].

4.4 *Metric Aggregation and Scoring*

MMI development is completed by aggregating the best-performing metrics to derive an index score. The number of metrics used in the index varies among studies, and the choice is rarely supported by clear empirical justification [139]. Professional judgment is often used to select metrics based on best overall performance, although ordered stepwise processes have been recommended and present more comprehensive and objective options [139, 140]. To express metrics on an equivalent numerical scale, raw values are commonly rescaled to reflect percent or proportional comparability to values in the reference site distribution or to the distribution of all sites producing metric scores on continuous 0–100 or 0–1 point scales that increase with impairment. Blocksom [49] reviewed the details of these and other common scoring methods. After scoring, metrics are nearly always aggregated into an index by simple averaging, although other methods, such as differential weighting based on relative importance [131, 141] or to account for variations in metric precision [131], have been used. Alternative aggregation strategies for MMIs represent yet another area where additional research is needed.

4.5 *Index Validation*

Validation of the index with independent data provides the most comprehensive evaluation of performance. Validation typically proceeds by randomly selecting subsets of impaired and reference sites, which are excluded from the dataset used for index development and used for *a posteriori* evaluation of the performance characteristics described above. The feasibility of index validation depends on the amount of data available, as statistical power is compromised by dividing datasets for this purpose. Categorical approaches for validating index accuracy are data expensive, as the validation set must be divided according to impairment status. When only a few sites are available, index accuracy may be validated by analyzing for correlations of index scores with stressor gradients, which requires fewer validation sites. This approach is especially useful in highly developed landscapes where there are few reference sites [127, 142]. Index accuracy is often prioritized over other performance characteristics, although more thorough validation strategies also evaluate precision [21, 106, 143]. Evaluation of index bias, as indicated by

relationships between indices and natural environmental variables, provides further information on index performance [21, 101, 106, 116].

The effectiveness of predictive models used for reference site classification is often evaluated using cross-validation by constructing models using only subsets of the available data. Leave-one-out cross-validation is a data-efficient method in which reference sites are excluded from the dataset one at a time. After each exclusion, the classification and modeling process is repeated. Each left-out site is then classified using the model constructed without that site. The proportion of agreements between original and cross-validated classifications, relative to the total number of reference sites, is a measure of the effectiveness of the model. This technique can be used with any site classification approach [114, 134, 144]. Null models, which are constructed by predicting metric values across all reference sites, with no classification, are useful for evaluating all types of classifications. If classification strength does not exceed that of the null model, then the classification provides no advantage [145].

5 Expert Interviews: Challenges and Important Considerations in Bioassessment

To provide a broader and more comprehensive perspective, we conducted interviews with three expert practitioners and developers of bioassessment programs. Their responses to our questions, provided here in a question-and-answer format, have been summarized with a focus on emerging issues relevant to bioassessment in aquatic systems.

Expert 1 Michael Barbour, Ph.D.—Adjunct Senior Scientist, Mote Marine Laboratory, Sarasota, FL, USA, and retired Director, Center for Ecological Sciences, Tetra Tech, Inc., Owings Mills, MD, USA

Q: What factors limit the potential for increased use of genetic information in bioassessment surveys? Is it likely that molecular genetic analysis will replace traditional taxonomic approaches, or will these processes be used in conjunction with each other?

A: A major challenge in the use of genetic data for bioassessment will be in determining how reference conditions are expressed and how to account for the effects of natural environmental variability on reference populations. It is unlikely that genetic analysis will replace traditional taxonomic approaches in the near future. Evolving DNA methods, however, should help to decrease taxonomic uncertainty and improve our evaluations of aquatic assemblages.

Q: What are the most important factors to consider in developing a bioassessment program?

A: Adherence to the Critical Elements Process in the design and implementation of bioassessment protocols should provide an objective means of evaluating the rigor of regulatory assessment programs and a basis for comparing data quality

among programs [146]. This process is used to evaluate programs with respect to 13 critical elements within three general categories: study design, methods implementation, and data interpretation. The process considers logistic feasibility and cost-effectiveness, calling for the highest methodological and data-quality standards that are reasonably attainable given existing technological and monetary constraints. When high-quality methods exist, new bioassessment programs should employ methods consistent to these to maximize efficiency (i.e., use of preexisting datasets) and historical significance.

Q: What are the most important recent developments that have improved comparability among assessment programs?

A: In addition to the critical elements process, the Biological Condition Gradient approach (BCG, [45]) provides a framework for developing consistent, meaningful, and understandable aquatic life use standards and is applicable to a wide variety of monitoring strategies and assemblages. The BCG establishes a baseline by employing best professional judgment within an organized framework whereby experts assign bioassessment samples to ecological condition tiers. Biologists trained to use the BCG produce highly consistent evaluations of site conditions. The use of the BCG should greatly facilitate the comparability of bioassessments conducted by different agencies and using different protocols.

Large-scale monitoring programs, such as the USEPA National Aquatic Resource Survey, are of great importance. This nationwide assessment program includes standardized sampling protocols and a probabilistic study design for the assessment of US streams, rivers, lakes, wetlands, and coastal waters. The ongoing intercalibration exercise, a key component of the European Union (EU) Water Framework Directive (EC 2000/60/EC; [147]), and resulting multi-country aquatic ecosystem surveys are other important examples. The major advantage of these programs is the development of consistent and rigorous protocols that allow for large-scale biological assessments of aquatic ecosystems.

Expert 2 Simone D. Langhans, Ph.D.—Humboldt Research Fellow, Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany

Q: What factors are most important in limiting comparability among bioassessment schemes?

A: Variability in the definition of reference conditions (i.e., the allowable amount of impairment within the reference dataset) can hinder the comparison of index scores from different assessment programs. Index scores are typically an expression of how similar a site is from the reference state; therefore, the use of similar reference criteria facilitates comparability among assessment indices.

The expression of index scores in a continuous manner, rather than as categorical ratings, is helpful when aggregating scores derived from different assessment schemes. Due to their discrete nature, categorical ratings may differ for scores that are actually quite similar; therefore, the most reliable aggregated indices are based on continuous scoring systems. For management purposes, categorical ratings can be applied after aggregation. Langhans et al. [141] present a method for standardizing metric scores or attribute measures among indices to a continuous

0–1 scale. The method can be applied to both categorical and continuous scales and preserves the best professional judgment of developers regarding the interpretation of score or measurement values.

Q: What aspects of the relationships of anthropogenic activities and ecological integrity are most misunderstood by nonscientists?

A: The watershed (or catchment) concept is of great importance, though often not understood by the layperson. Following from the simple concept that water flows downhill, water bodies integrate the effects of human activities everywhere within their watersheds. Responsible management of aquatic ecosystems must consider not only in-stream and local effects but also effects at much larger spatial scales. The range of relevant spatial scales that should be considered increases with the size of the water body, with the largest river systems integrating the effects of anthropogenic activities over thousands of square kilometers.

Q: In considering streams in different natural settings and at different positions along the general gradient of impairment, which are the most important candidates for preservation or restoration?

A: How best to prioritize conservation and restoration efforts for aquatic ecosystems is currently a popular and important topic in the EU. A strategy that considers the conservation of existing ecological integrity and the restoration of impaired systems simultaneously is best. When biological assemblage objectives are given high priority, the most effective areas for restoration are those in close proximity to high-quality conservation areas because the conservation areas provide sources for recolonization. For example, Tonkin et al. [148] evaluated the likelihood of recolonization by invertebrates at 21 river restoration sites in Germany. They determined that the density of occurrence of a taxon at surrounding sites (proportion of sites with the taxon present) and the distance to the nearest potential source site were important factors for predicting recolonization.

Expert 3 Gregory J. Pond, M.S.—Aquatic Biologist, USA Environmental Protection Agency, Region III, Wheeling, WV, USA

Q: Given the myriad protocols currently employed to conduct assemblage-level assessments, what considerations should be made by investigators and managers to select the most accurate, precise, and cost-effective strategies?

A: Protocols should be flexible, thoroughly documented in standard operating procedures, and based on the varying assessment goals, characteristics of the system being studied and available funding. For example, species-level macroinvertebrate data may in some cases produce the most effective assessment results. In other situations, temporal variation among samples may produce confusion at high levels of taxonomic resolution, for example, if errors in identification increase at times of the year when early instars predominate in the samples. In such cases, coarser taxonomic resolution at the genus or family level may be necessary to avoid inconsistency among samples.

If assessment on a large spatial scale is a priority, investigators may use less time-intensive methods to assess a greater number of sites within the time and monetary constraints of the project. In an attempt to provide a spatially

comprehensive assessment of Pennsylvania (USA) waters, the state's Department of Environmental Protection (DEP) conducted rapid surveys focused on macroinvertebrate assemblage characteristics that could be evaluated in the field by trained biologists without the need of extensive sampling and laboratory processing. This protocol allowed DEP to conduct an initial screening of several thousand stream sites over a 2-year period within budgetary constraints. However, the accuracy and precision of field-based rapid surveys such as this are likely far lower than would be expected from more intensive sampling and processing protocols that produce quantitative, genus- or species-level datasets.

Data consistency and comparability are also of great concern. When USEPA conducts bioassessments, the protocols developed by the state are typically followed. This insures that the data collected are comparable with those produced by state biologists and that the methods have been calibrated for the region of interest. Natural variability of system-specific characteristics should also be considered. For example, Virginia (USA) is currently developing a new protocol for swamp streams, which have not been previously assessed for regulatory purposes.

Q: What are your thoughts regarding the use of continuous environmental variables within a predictive modeling framework versus typological approaches for reference site classification?

A: Large-scale typological approaches have generally been insufficient in accounting for the variation in natural environmental factors that affect biota among aquatic systems. Typological approaches, however, are convenient, easy to use and understand, and can be effective within relatively small and homogenous geographic regions. Natural environmental gradients often persist within typological categories, and care should be taken to ensure that typological approaches do not oversimplify these gradients. Predictive modeling is more analytically intensive and requires more precise data, but generally provides more reliable results in highly heterogeneous regions. In the development of an *O/E* model for Central Appalachian streams, Pond and North [75] determined that subcoregion [149], Julian day and latitude were the most important natural predictors of reference macroinvertebrate assemblages. For that study region, a predictive modeling approach was chosen as the best strategy, given the importance of the continuous variables Julian day and latitude. A typological approach, where reference conditions are developed at the subcoregion level, could also be effective; however, the effects of latitude and seasonality should be carefully observed and potentially controlled.

6 Conclusions

In this chapter, we provided an introduction to the major components of assemblage-level bioassessments of aquatic systems. Macroinvertebrates, fish, and algae are the most commonly used assemblages, although the methods described here are applicable to, and have been successfully used with, other biotic

assemblages [150–152]. One emerging strategy is in the use of prokaryote assemblages, which has historically been limited because many prokaryotes are not readily cultured in the laboratory. However, emerging technologies that allow for quantification of assemblage composition through DNA sequencing have largely removed this limitation, making prokaryote assemblage assessments an emerging new option for bioassessment of aquatic systems [152, 153]. Rapidly evolving DNA sequencing methods have the potential to greatly enhance not only bioassessments using prokaryotes but also those using assemblages that have traditionally been evaluated by identification of specimens based on morphological characteristics [154–157].

Regardless of the assemblage type chosen or the methods used for identifying taxa in the assemblage, the most challenging aspect of bioassessments has been, and remains, the difficulty in separating environmental effects on assemblages that are the result of naturally varying factors such as climate and geology from those caused by anthropogenic factors. The use of the RCA, coupled with advanced predictive modeling methods such as machine learning techniques, has enhanced our ability to predict how assemblages should vary based on natural environmental factors. Such enhanced predictive power should ultimately allow for more accurate determination of assemblage variation patterns that indicate impairment. Despite these advancements, predictive modeling and the use of the RCA are greatly confounded by the lack of suitable reference sites in many regions. To this end, alternative strategies that employ both impaired and reference sites to derive expected reference conditions have been proposed [118–121] and warrant further evaluation to determine their widespread applicability. Because of the scarcity of reference sites in many regions and the high potential for complex interactions between natural environmental factors and stressors, the development of additional data-efficient methods for predicting expected assemblages under unimpaired conditions and for quantifying deviations from these expectations is much needed.

An additional challenge for contemporary bioassessment programs is the shifting baseline syndrome (*sensu* Hawkins et al. [43]) wherein future climate change is likely to alter temperature and precipitation regimes globally, thus changing the assemblage compositions that might reasonably be expected under minimally impaired conditions. To meet this challenge, spatially and temporally extensive monitoring is essential to derive realistic reference conditions. Several large-scale assessment programs have been recently developed, such as the EU Water Framework Directive, the US Geological Survey's National Water Quality Assessment program, the US EPA's National Aquatic Resources Survey, the US National Science Foundation's National Ecological Observatory Network, and the Canadian Biological Monitoring network. These programs include rigorous and thoroughly documented bioassessment protocols focused on monitoring aquatic assemblages over large spatial and long temporal scales. Data produced by these important programs will enhance our ability to overcome the inherent challenges in evaluating ecological integrity when least-impaired reference conditions are rare, highly variable among regions, and changing in response to global climate change.

Acknowledgments We wish to thank Michael Barbour, Simone Langhans, and Gregory Pond for providing the material that resulted in our “interviews with the experts” section. Their insights have provided an invaluable contribution to this work. This is VCU Rice Rivers Center Contribution Number 49.

References

1. Barbour MT, Gerritsen J, Snyder B et al (1999) Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates and fish, 2nd edn. EPA 841-B-99-002. U.S. Environmental Protection Agency, Office of Water, Washington, DC
2. Ahrens L, Bundschuh M (2014) Fate and effects of poly- and perfluoroalkyl substances in the aquatic environment: a review. *Environ Toxicol Chem* 33:1921–1929
3. Mueller M, Pander J, Geist J (2014) A new tool for assessment and monitoring of community and ecosystem change based on multivariate abundance data integration from different taxonomic groups. *Environ Syst Res* 3(1):12. doi:10.1186/2193-2697-3-12
4. Kolkwitz R, Marsson M (1902) Grundsätze für die biologische beurtheilung des wassers, nach seiner flora und fauna. *Mitteilungen der Prüfungsansalt für Wasserversorgung und Abwasserreinigung* 1:1–64
5. Junqueira MV, Friedrich G, de Araujo PR (2010) A saprobic index for biological assessment of river water quality in Brazil (Minas Gerais and Rio de Janeiro states). *Environ Monit Assess* 163:545–554
6. Usseglio-Polatera P, Bournaud M, Richoux P et al (2000) Biological and ecological traits of benthic freshwater macroinvertebrates: relationships and definition of groups with similar traits. *Freshw Biol* 43:175–205
7. Frey DG (1977) Biological integrity of water—an historical approach. In: *The integrity of water: proceedings of a symposium*, US Environmental Protection Agency, Washington, DC, 10–12 Mar 1975
8. Karr JR, Dudley DR (1981) Ecological perspective on water quality goals. *Environ Manage* 5:55–68
9. Barbour MT, Paul MJ (2010) Adding value to water resource management through biological assessment of rivers. *Hydrobiologia* 651:17–24
10. Fauth J, Bernardo J, Camara M et al (1996) Simplifying the jargon of community ecology: a conceptual approach. *Am Nat* 147:282–286
11. Carlisle DM, Hawkins CP, Meador MR et al (2008) Biological assessments of Appalachian streams based on predictive models for fish, macroinvertebrate, and diatom assemblages. *J North Am Benthol Soc* 27:16–37
12. Bae M, Kwon Y, Hwang S et al (2011) Relationships between three major stream assemblages and their environmental factors in multiple spatial scales. *Ann Limnol—Int J Lim* 47: S91–S105
13. Pilière A, Schipper AM, Breure AM et al (2014) Comparing responses of freshwater fish and invertebrate community integrity along multiple environmental gradients. *Ecol Indic* 43:215–226
14. Carter JL, Resh VH (2001) After site selection and before data analysis: sampling, sorting, and laboratory procedures used in stream benthic macroinvertebrate monitoring programs by USA state agencies. *J North Am Benthol Soc* 20:658–682
15. Potapova M, Charles DF (2005) Choice of substrate in algae-based water-quality assessment. *J North Am Benthol Soc* 24:415–427
16. Li J, Herlihy A, Gerth W et al (2001) Variability in stream macroinvertebrates at multiple spatial scales. *Freshw Biol* 46:87–97

17. Pritt JJ, Frimpong EA (2014) The effect of sampling intensity on patterns of rarity and community assessment metrics in stream fish samples. *Ecol Indic* 39:169–178
18. Lenat DR, Resh VH (2001) Taxonomy and stream ecology—the benefits of genus- and species-level identification. *J North Am Benthol Soc* 20:287–298
19. Jiang X, Xiong J, Song Z et al (2013) Is coarse taxonomy sufficient for detecting macroinvertebrate patterns in floodplain lakes? *Ecol Indic* 27:48–55
20. Uzarski DG, Burton TM, Cooper MJ et al (2005) Fish habitat use within and across wetland classes in coastal wetlands of the five Great Lakes: development of a fish-based index of biotic integrity. *J Great Lakes Res* 31:171–187
21. Chen K, Hughes RM, Xu S et al (2014) Evaluating performance of macroinvertebrate-based adjusted and unadjusted multi-metric indices (MMI) using multi-season and multi-year samples. *Ecol Indic* 36:142–151
22. Smucker NJ, Vis ML (2009) Use of diatoms to assess agricultural and coal mining impacts on streams and a multiassemblage case study. *J North Am Benthol Soc* 28:659–675
23. Blocksom KA, Kurtenbach JP, Klemm DJ et al (2002) Development and evaluation of the lake macroinvertebrate integrity index (LMII) for New Jersey lakes and reservoirs. *Environ Monit Assess* 77:311–333
24. Pan Y, Hughes RM, Herlihy AT et al (2012) Non-wadeable river bioassessment: spatial variation of benthic diatom assemblages in Pacific Northwest rivers, USA. *Hydrobiologia* 684:241–260
25. Whittier TR, Hughes RM, Lomnický GA et al (2007) Fish and amphibian tolerance values and an assemblage tolerance index for streams and rivers in the western USA. *Trans Am Fish Soc* 136:254–271
26. Black RW, Moran PW, Frankforter JD (2011) Response of algal metrics to nutrients and physical factors and identification of nutrient thresholds in agricultural streams. *Environ Monit Assess* 175:397–417
27. Zhu D, Chang J (2008) Annual variations of biotic integrity in the upper Yangtze River using an adapted index of biotic integrity (IBI). *Ecol Indic* 8:564–572
28. Lenat DR (1993) Using mentum deformities of Chironomus larvae to evaluate the effects of toxicity and organic loading in streams. *J North Am Benthol Soc* 12(3):265–269
29. Fore LS, Grafe C (2002) Using diatoms to assess the biological condition of large rivers in Idaho (USA). *Freshw Biol* 47:2015–2037
30. Reynoldson T, Norris R, Resh V et al (1997) The reference condition: a comparison of multimetric and multivariate approaches to assess water-quality impairment using benthic macroinvertebrates. *J North Am Benthol Soc* 16:833–852
31. Bailey RC, Kennedy MG, Dervish MZ (1998) Biological assessment of freshwater ecosystems using a reference condition approach: comparing predicted and actual benthic invertebrate communities in Yukon streams. *Freshw Biol* 39:765–774
32. Stoddard JL, Larsen DP, Hawkins CP et al (2006) Setting expectations for the ecological condition of streams: the concept of reference condition. *Ecol Appl* 16:1267–1276
33. Birks H, Line J, Juggins S et al (1990) Diatoms and pH reconstruction. *Philos Trans R Soc Lond B Biol Sci* 327:263–278
34. Hughes RM, Kaufmann PR, Herlihy AT et al (1998) A process for developing and evaluating indices of fish assemblage integrity. *Can J Fish Aquat Sci* 55:1618–1631
35. Thoms M, Ogden R, Reid M (1999) Establishing the condition of lowland floodplain rivers: a palaeo-ecological approach. *Freshw Biol* 41:407–423
36. Shirey PD, Cowley DE, Sallénave R (2008) Diatoms from gut contents of museum specimens of an endangered minnow suggest long-term ecological changes in the Rio Grande (USA). *J Paleolimnol* 40:263–272
37. Nelson SM, Roline RA (1996) Recovery of a stream macroinvertebrate community from mine drainage disturbance. *Hydrobiologia* 339:73–84

38. Clark JL, Clements WH (2006) The use of in situ and stream microcosm experiments to assess population- and community-level responses to metals. *Environ Toxicol Chem* 25:2306–2312
39. Applegate JM, Baumann PC, Emery EB et al (2007) First steps in developing a multimetric macroinvertebrate index for the Ohio River. *River Res Appl* 23:683–697
40. Plafkin JL, Barbour MT, Porter KD, et al (1989) Rapid bioassessment protocols for use in streams and rivers: benthic macroinvertebrates and fish. EPA 444-4-89-001. U.S. Environmental Protection Agency, Office of Water, Washington, DC
41. Mykrä H, Heino J, Muotka T (2007) Scale-related patterns in the spatial and environmental components of stream macroinvertebrate assemblage variation. *Global Ecol Biogeogr* 16:149–159
42. Hoeinghaus DJ, Winemiller KO, Birnbaum JS (2007) Local and regional determinants of stream fish assemblage structure: inferences based on taxonomic vs. functional groups. *J Biogeogr* 34:324–338
43. Hawkins CP, Olson JR, Hill RA (2010) The reference condition: predicting benchmarks for ecological and water-quality assessments. *J North Am Benthol Soc* 29:312–343
44. Ganasan V, Hughes RM (1998) Application of an index of biological integrity (IBI) to fish assemblages of the rivers Khan and Kshipra (Madhya Pradesh), India. *Freshw Biol* 40:367–383
45. Davies SP, Jackson SK (2006) The biological condition gradient: a descriptive model for interpreting change in aquatic ecosystems. *Ecol Appl* 16:1251–1266
46. Mazor RD, Reynoldson TB, Rosenberg DM et al (2006) Effects of biotic assemblage, classification, and assessment method on bioassessment performance. *Can J Fish Aquat Sci* 63:394–411
47. Hawkes H (1998) Origin and development of the biological monitoring working party score system. *Water Res* 32:964–968
48. Whittier TR, Stoddard JL, Larsen DP et al (2007) Selecting reference sites for stream biological assessments: best professional judgment or objective criteria. *J North Am Benthol Soc* 26:349–360
49. Blocksom KA (2003) A performance comparison of metric scoring methods for a multimetric index for Mid-Atlantic Highlands streams. *Environ Manage* 31:670–682
50. Yates AG, Bailey RC (2010) Selecting objectively defined reference sites for stream bioassessment programs. *Environ Monit Assess* 170:129–140
51. Blocksom KA, Flotemersch JE (2005) Comparison of macroinvertebrate sampling methods for nonwadeable streams. *Environ Monit Assess* 102:243–262
52. McCune B, Grace JB (2002) Analysis of ecological communities. MjM Software Design, Glendale Beach
53. Legendre P, Legendre LF (2012) Numerical ecology. Elsevier, Amsterdam
54. R Core Team (2014) R: a language and environment for statistical computing. <http://www.r-project.org/>
55. Bray JR, Curtis JT (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecol Monogr* 27:325–349
56. Dice LR (1945) Measures of the amount of ecologic association between species. *Ecology* 26:297–302
57. Sorenson T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *Kongelige Danske Videnskabernes Selskab* 5:4–7
58. Van Sickle J, Hughes RM (2000) Classification strengths of ecoregions, catchments, and geographic clusters for aquatic vertebrates in Oregon. *J North Am Benthol Soc* 19:370–384
59. Clarke K, Green R (1988) Statistical design and analysis for a “biological effects” study. *Mar Ecol Prog Ser* 46:213–226
60. Mielke PW Jr, Berry KJ, Johnson ES (1976) Multi-response permutation procedures for a priori classifications. *Commun Stat Theory Methods* 5:1409–1424
61. Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecol* 26:32–46

62. Omernik JM (1987) Ecoregions of the conterminous United States. *Ann Assoc Am Geogr* 77:118–125
63. Hughes RM, Larsen DP, Omernik JM (1986) Regional reference sites: a method for assessing stream potentials. *Environ Manage* 10:629–635
64. Aroviita J, Mykra H, Muotka T et al (2009) Influence of geographical extent on typology- and model-based assessments of taxonomic completeness of river macroinvertebrates. *Freshw Biol* 54:1774–1787
65. Hawkins CP, Norris RH, Gerritsen J et al (2000) Evaluation of the use of landscape classifications for the prediction of freshwater biota: synthesis and recommendations. *J North Am Benthol Soc* 19:541–556
66. Dallas HF (2004) Spatial variability in macroinvertebrate assemblages: comparing regional and multivariate approaches for classifying reference sites in South Africa. *Afr J Aquat Sci* 29:161–171
67. Dail MR, Hill JR, Miller RD (2013) The Virginia coastal plain macroinvertebrate index. Technical bulletin WQA/2013-002. Virginia Department of Environmental Quality, Richmond
68. Burton J, Gerritsen J (2003) A stream condition index for Virginia non-coastal streams. Tetra-Tech, Owings Mills
69. Hawkins CP, Vinson MR (2000) Weak correspondence between landscape classifications and stream invertebrate assemblages: implications for bioassessment. *J North Am Benthol Soc* 19:501–517
70. Wright J, Moss D, Armitage P et al (1984) A preliminary classification of running-water sites in Great Britain based on macro-invertebrate species and the prediction of community type using environmental data. *Freshw Biol* 14:221–256
71. Wright JF, Sutcliffe DW, Furse MT (2000) Assessing the biological quality of freshwaters. RIVPACS and other techniques. Freshwater Biological Association, Ambleside
72. Hawkins CP, Norris RH, Hogue JN et al (2000) Development and evaluation of predictive models for measuring the biological integrity of streams. *Ecol Appl* 10:1456–1477
73. Smith M, Kay W, Edward D et al (1999) AusRivAS: using macroinvertebrates to assess ecological condition of rivers in Western Australia. *Freshw Biol* 41:269–282
74. Reynoldson TB, Bailey R, Day K et al (1995) Biological guidelines for freshwater sediment based on Benthic Assessment of Sediment (the BEAST) using a multivariate approach for predicting biological state. *Aust J Ecol* 20:198–219
75. Pond GJ, North SH (2013) Application of a benthic observed/expected-type model for assessing Central Appalachian streams influenced by regional stressors in West Virginia and Kentucky. *Environ Monit Assess* 185:9299–9320
76. Pardo I, Gómez-Rodríguez C, Abraín R et al (2014) An invertebrate predictive model (NORTI) for streams and rivers: sensitivity of the model in detecting stress gradients. *Ecol Indic* 45:51–62
77. Linke S, Bailey RC, Schwindt J (1999) Temporal variability of stream bioassessments using benthic macroinvertebrates. *Freshw Biol* 42:575–584
78. Pond G, Call S, Brumley J et al (2003) The Kentucky macroinvertebrate bioassessment index: derivation of regional narrative criteria for headwater and wadeable streams. Kentucky Department for Environmental Protection, Division of Water, Frankfort. http://water.ky.gov/Documents/QA/MBI/Statewide_MBI.pdf
79. Hill MO (1979) TWINSPAN: a FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. Cornell University, Ithaca
80. Davy-Bowker J, Clarke R, Corbin T (2008) River invertebrate classification tool—final report. Scotland and Northern Ireland Forum for Environmental Research, Edinburgh
81. Oksanen J, Minchin PR (1997) Instability of ordination results under changes in input data order: explanations and remedies. *J Veg Sci* 8:447–454

82. Austin M (2013) Inconsistencies between theory and methodology: a recurrent problem in ordination studies. *J Veg Sci* 24:251–268
83. Linke S, Norris RH, Faith DP et al (2005) ANNA: a new prediction method for bioassessment programs. *Freshw Biol* 50:147–158
84. Olden JD (2003) A species-specific approach to modeling biological communities and its potential for conservation. *Conserv Biol* 17:854–863
85. Bowman MF, Somers KM (2005) Considerations when using the reference condition approach for bioassessment of freshwater ecosystems. *Water Qual Res J Can* 40:347–360
86. Melles S, Jones N, Schmidt B (2014) Evaluation of current approaches to stream classification and a heuristic guide to developing classifications of integrated aquatic networks. *Environ Manage* 53:549–566
87. Ode PR, Hawkins CP, Mazor RD (2008) Comparability of biological assessments derived from predictive models and multimetric indices of increasing geographic scope. *J North Am Benthol Soc* 27:967–985
88. Aroviita J, Koskenniemi E, Kotanen J et al (2008) A priori typology-based prediction of benthic macroinvertebrate fauna for ecological classification of rivers. *Environ Manage* 42:894–906
89. Joy M, Death R (2002) Predictive modelling of freshwater fish as a biomonitoring tool in New Zealand. *Freshw Biol* 47:2261–2275
90. Clarke RT, Wright JF, Furse MT (2003) RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. *Ecol Model* 160:219–233
91. Hawkins CP (2006) Quantifying biological integrity by taxonomic completeness: its utility in regional and global assessments. *Ecol Appl* 16:1277–1294
92. Simpson J, Norris R, Wright J et al (2000) Biological assessment of river quality: development of AUSRIVAS models and outputs. In: Wright J, Sutcliffe D, Furse M (eds) *Assessing the biological quality of fresh waters: RIVPACS and other techniques*. Freshwater Biological Association, Ambleside
93. Van Sickle J (2008) An index of compositional dissimilarity between observed and expected assemblages. *J North Am Benthol Soc* 27:227–235
94. Van Sickle J, Larsen DP, Hawkins CP (2007) Exclusion of rare taxa affects performance of the O/E index in bioassessments. *J North Am Benthol Soc* 26:319–331
95. Bates Prins SC, Smith EP (2007) Using biological metrics to score and evaluate sites: a nearest-neighbour reference condition approach. *Freshw Biol* 52:98–111
96. Moss D, Wright J, Sutcliffe D et al (2000) Evolution of statistical methods in RIVPACS. In: Wright J, Sutcliffe D, Furse M (eds) *Assessing the biological quality of fresh waters: RIVPACS and other techniques*. Freshwater Biological Association, Ambleside
97. Hoang H, Recknagel F, Marshall J et al (2001) Predictive modelling of macroinvertebrate assemblages for stream habitat assessments in Queensland (Australia). *Ecol Model* 146:195–206
98. Waite IR, Brown LR, Kennen JG et al (2010) Comparison of watershed disturbance predictive models for stream benthic macroinvertebrates for three distinct ecoregions in western US. *Ecol Indic* 10:1125–1136
99. McCormick FH, Hughes RM, Kaufmann PR et al (2001) Development of an index of biotic integrity for the Mid-Atlantic Highlands region. *Trans Am Fish Soc* 130:857–877
100. Oberdorff T, Pont D, Hugueny B et al (2002) Development and validation of a fish-based index for the assessment of river health in France. *Freshw Biol* 47:1720–1734
101. Vander Laan JJ, Hawkins CP (2014) Enhancing the performance and interpretation of freshwater biological indices: an application in arid zone streams. *Ecol Indic* 36:470–482
102. Cao Y, Hawkins CP, Olson J et al (2007) Modeling natural environmental gradients improves the accuracy and precision of diatom-based indicators. *J North Am Benthol Soc* 26:566–585
103. Yuan LL (2004) Assigning macroinvertebrate tolerance classifications using generalised additive models. *Freshw Biol* 49:662–677
104. Wintle BA, McCarthy MA, Volinsky CT et al (2003) The use of Bayesian model averaging to better represent uncertainty in ecological models. *Conserv Biol* 17:1579–1590

105. Clark JS (2005) Why environmental scientists are becoming Bayesians. *Ecol Lett* 8:2–14
106. Marzin A, Delaigue O, Logez M et al (2014) Uncertainty associated with river health assessment in a varying environment: the case of a predictive fish-based index in France. *Ecol Indic* 43:195–204
107. Cutler DR, Edwards TC Jr, Beard KH et al (2007) Random forests for classification in ecology. *Ecology* 88:2783–2792
108. De'ath G (2007) Boosted trees for ecological modeling and prediction. *Ecology* 88:243–251
109. Olden JD, Lawler JJ, Poff NL (2008) Machine learning methods without tears: a primer for ecologists. *Q Rev Biol* 83:171–193
110. Goethals PL, Dedecker AP, Gabriels W et al (2007) Applications of artificial neural networks predicting macroinvertebrates in freshwaters. *Aquat Ecol* 41:491–508
111. Hoang TH, Lock K, Mouton A et al (2010) Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam. *Ecol Inform* 5:140–146
112. Feio M, Viana-Ferreira C, Costa C (2014) Combining multiple machine learning algorithms to predict taxa under reference conditions for streams bioassessment. *River Res Appl* 30 (9):1157–1165. doi:[10.1002/rra.2707](https://doi.org/10.1002/rra.2707)
113. Bishop CM (1995) *Neural networks for pattern recognition*. Clarendon, Oxford
114. Olden JD, Joy MK, Death RG (2006) Rediscovering the species in community-wide predictive modeling. *Ecol Appl* 16:1449–1460
115. Breiman L, Friedman J, Olshen R et al (1984) *Classification and regression trees (CART)*. Wadsworth International Group, Belmont
116. Hawkins CP, Cao Y, Roper B (2010) Method of predicting reference condition biota affects the performance and interpretation of ecological indices. *Freshw Biol* 55:1066–1085
117. Waite IR, Kennen JG, May JT et al (2012) Comparison of stream invertebrate response models for bioassessment metrics. *J Am Water Resour Assoc* 48:570–583
118. Baker EA, Wehrly KE, Seelbach PW et al (2005) A multimetric assessment of stream condition in the Northern Lakes and Forests Ecoregion using spatially explicit statistical modeling and regional normalization. *Trans Am Fish Soc* 134:697–710
119. Angradi TR, Pearson MS, Bolgrien DW et al (2009) Multimetric macroinvertebrate indices for mid-continent US great rivers. *J North Am Benthol Soc* 28:785–804
120. Chessman BC, Royal MJ (2004) Bioassessment without reference sites: use of environmental filters to predict natural assemblages of river macroinvertebrates. *J North Am Benthol Soc* 23:599–615
121. Schoolmaster DR Jr, Grace JB, Schweiger EW et al (2013) A causal examination of the effects of confounding factors on multimetric indices. *Ecol Indic* 29:411–419
122. Hering D, Feld CK, Moog O et al (2006) Cook book for the development of a multimetric index for biological condition of aquatic ecosystems: experiences from the European AQEM and STAR projects and related initiatives. *Hydrobiologia* 566:311–324
123. Stoddard JL, Herlihy AT, Peck DV et al (2008) A process for creating multimetric indices for large-scale aquatic surveys. *J North Am Benthol Soc* 27:878–891
124. Blocksom K, Johnson B (2009) Development of a regional macroinvertebrate index for large river bioassessment. *Ecol Indic* 9:313–328
125. Stevenson RJ, Zalack JT, Wolin J (2013) A multimetric index of lake diatom condition based on surface-sediment assemblages. *Freshw Sci* 32:1005–1025
126. Cao Y, Hawkins CP (2011) The comparability of bioassessments: a review of conceptual and methodological issues. *J North Am Benthol Soc* 30:680–701
127. Baptista DF, Henriques-Oliveira AL, Oliveira RBS et al (2013) Development of a benthic multimetric index for the Serra da Bocaina bioregion in Southeast Brazil. *Braz J Biol* 73:573–583
128. Fetscher AE, Stancheva R, Kociolek JP et al (2014) Development and comparison of stream indices of biotic integrity using diatoms vs. non-diatom algae vs. a combination. *J Appl Phycol* 26:433–450

129. McBride GB, Loftis JC, Adkins NC (1993) What do significance tests really tell us about the environment? *Environ Manage* 17:423–432
130. Smith J, Beauchamp J, Stewart A (2005) Alternative approach for establishing acceptable thresholds on macroinvertebrate community metrics. *J North Am Benthol Soc* 24:428–440
131. Bowman MF, Somers KM (2006) Evaluating a novel Test Site Analysis (TSA) bioassessment approach. *J North Am Benthol Soc* 25:712–727
132. Stribling J, Jessup B, Gerritsen J (2000) Development of biological and physical habitat criteria for Wyoming streams and their use in the TMDL process. Tetra-Tech, Owings Mills
133. Barbour M, Gerritsen J, Griffith G et al (1996) A framework for biological criteria for Florida streams using benthic macroinvertebrates. *J North Am Benthol Soc* 15(2):185–211
134. Clarke R, Furse M, Gunn R et al (2002) Sampling variation in macroinvertebrate data and implications for river quality indices. *Freshw Biol* 47:1735–1751
135. Kaufmann PR, Levine P, Peck DV et al (1999) Quantifying physical habitat in wadeable streams. EPA/620/R-99-003, US Environmental Protection Agency, Washington, DC
136. Hose G, Turak E, Waddell N (2004) Reproducibility of AUSRIVAS rapid bioassessments using macroinvertebrates. *J North Am Benthol Soc* 23:126–139
137. Houston L, Barbour M, Lenat D et al (2002) A multi-agency comparison of aquatic macroinvertebrate-based stream bioassessment methodologies. *Ecol Indic* 1:279–292
138. Lyons J (2012) Development and validation of two fish-based indices of biotic integrity for assessing perennial coolwater streams in Wisconsin, USA. *Ecol Indic* 23:402–412
139. Van Sickle J (2010) Correlated metrics yield multimetric indices with inferior performance. *Trans Am Fish Soc* 139:1802–1817
140. Schoolmaster DR, Grace JB, Schweiger EW (2012) A general theory of multimetric indices and their properties. *Methods Ecol Evol* 3:773–781
141. Langhans SD, Lienert J, Schuwirth N et al (2013) How to make river assessments comparable: a demonstration for hydromorphology. *Ecol Indic* 32:264–275
142. Wu W, Xu Z, Yin X et al (2014) Assessment of ecosystem health based on fish assemblages in the Wei River basin, China. *Environ Monit Assess* 186:3701–3716
143. Lunde KB, Resh VH (2012) Development and validation of a macroinvertebrate index of biotic integrity (IBI) for assessing urban impacts to Northern California freshwater wetlands. *Environ Monit Assess* 184:3653–3674
144. Breiman L (2001) Random forests. *Mach Learn* 45:5–32
145. Van Sickle J, Hawkins CP, Larsen DP et al (2005) A null model for the expected macroinvertebrate assemblage in streams. *J North Am Benthol Soc* 24:178–191
146. Yoder CO, Barbour MT (2009) Critical technical elements of state bioassessment programs: a process to evaluate program rigor and comparability. *Environ Monit Assess* 150:31–42
147. Commission E (2000) Water Framework Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. *Off J Eur Community* L327:1–73
148. Tonkin JD, Stoll S, Sundermann A et al (2014) Dispersal distance and the pool of taxa, but not barriers, determine the colonisation of restored river reaches by benthic invertebrates. *Freshw Biol* 59(9):1843–1855
149. Woods A, Omernik J, Brown D et al (1996) Level III and IV ecoregions of Pennsylvania and the Blue Ridge mountains, the Ridge and Valley, and the central Appalachians of Virginia, West Virginia, and Maryland. EPA/600/R-96/077, US Environmental Protection Agency, Washington, DC
150. Bryce S, Hughes R, Kaufmann P (2002) Development of a bird integrity index: using bird assemblages as indicators of riparian condition. *Environ Manage* 30:294–310
151. Kanninen A, Vallinkoski V, Leka J et al (2013) A comparison of two methods for surveying aquatic macrophyte communities in boreal lakes: implications for bioassessment. *Aquat Bot* 104:88–100

152. Washington VJ, Lear G, Neale MW et al (2013) Environmental effects on biofilm bacterial communities: a comparison of natural and anthropogenic factors in New Zealand streams. *Freshw Biol* 58:2277–2286
153. Burgos-Caraballo S, Cantrell SA, Ramírez A (2014) Diversity of benthic biofilms along a land use gradient in tropical headwater streams, Puerto Rico. *Microb Ecol* 68(1):47–59
154. Baird DJ, Hajibabaei M (2012) Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Mol Ecol* 21:2039–2044
155. Pfrender M, Hawkins C, Bagley M et al (2010) Assessing macroinvertebrate biodiversity in freshwater ecosystems: advances and challenges in DNA-based approaches. *Q Rev Biol* 85:319–340
156. Manoylov KM (2014) Taxonomic identification of algae (morphological and molecular): species concepts, methodologies, and their implications for ecological bioassessment. *J Phycol* 50:409–424
157. Stein ED, White BP, Mazor RD et al (2014) Does DNA barcoding improve performance of traditional stream bioassessment metrics? *Freshw Sci* 33:302–311