DOI: 10.1007/s00267-002-0036-4

Integrating Bioassessment and Ecological Risk Assessment: An Approach to Developing Numerical Water-Quality Criteria

RYAN S. KING*
CURTIS J. RICHARDSON
Nicholas School of the Environment
and Earth Sciences
Duke University
Box 90328
Durham, North Carolina 27708 USA

ABSTRACT / ioassessment is used worldwide to monitor aquatic health but is infrequently used with risk-assessment objectives, such as supporting the development of defensible, numerical water-quality criteria. To this end, we present a generalized approach for detecting potential ecological thresholds using assemblage-level attributes and a multimetric index (Index of Biological Integrity—IBI) as endpoints in response to numerical changes in water quality. To illustrate the approach, we used existing macroinvertebrate and surface-water total phosphorus (TP) datasets from an observed P gradient and a P-dosing experiment in wetlands of the south Florida coastal plain nutrient ecoregion. Ten assemblage attributes were identified as potential metrics using the observational data, and five were validated in the experiment. These five core metrics were subjected individually and as an aggregated Nutrient–IBI to nonparametric changepoint analysis (nCPA) to estimate cumulative probabilities of a threshold response to TP. Threshold responses were evident for all metrics and the IBI, and were repeatable through time. Results from the observed gradient indicated that a threshold was ≥50% probable between 12.6 and 19.4 μg/L TP for individual metrics and 14.8 μg/L TP for the IBI. Results from the P-dosing experiment revealed ≥50% probability of a response between 11.2 and 13.0 μg/L TP for the metrics and 12.3 μg/L TP for the IBI. Uncertainty analysis indicated a low (typically ≥5%) probability that an IBI threshold occurred at ≤ 10 μg/L TP, while there was ≥95% certainty that the threshold was ≤ 17 μg/L TP. The weight-of-evidence produced from these analyses implies that a TP concentration > 12–15 μg/L is likely to cause degradation of macroinvertebrate assemblage structure and function, a reflection of biological integrity, in the study area. This finding may assist in the development of a numerical water-quality criterion for TP in this ecoregion, and illustrates the utility of bioassessment to environmental decision-making.

Bioassessment has become a widely accepted technique for monitoring aquatic health in streams, lakes, and wetlands throughout the world (Rosenberg and Resh 1993). Bioassessment has a long history in Europe (reviewed by Cairns and Pratt 1993) and has more recently become popular in North America, largely in response to the mandate of §101(a) of the Clean Water Act (CWA) to restore and maintain the biological integrity of the USA's waters (Karr 1981). One bioassessment approach that has received considerable attention in the USA is the multimetric approach (sensu Karr 1981). Multimetric indices, such as the Index of Biological Integrity (e.g., Karr and Chu 1997), are an aggregation of a suite of biological attributes that represent key elements of structure or function of an aquatic assemblage and show a consistent, predictable response to human influence. The strength of multimetric assessments lies in their ability to integrate multiple facets of biological condition (Barbour and others 1995), and thus provide an overall indication of biological integrity (Karr and Dudley 1981; Angermeier and Karr 1994).

One potentially important but underutilized application of multimetric bioassessment is supporting the development of numerical water-quality criteria (Miltner and Rankin 1998; Dodds and Welch 2000). The premise of bioassessment is that resident biota in a water body are natural integrators of environmental conditions and thus can reveal the effects of episodic changes in water quality as well as cumulative pollution (Rosenberg and Resh 1993). Nevertheless, development of water-quality criteria has historically been based on laboratory tests on individual species or solely on chemical endpoints without accounting for the assemblage-level consequences (Barbour and others 2000). The United States Environmental Protection

KEY WORDS: Biological assessment; Ecological thresholds; Everglades; Index of Biological integrity; Macroinvertebrates; Nutrients; Phosphorus; Risk assessment; Wetlands

Published online May 13, 2003.

*Author to whom correspondence should be addressed at: Smithsonian Environmental Research Center, 647 Contees Wharf Road, Box 28, Edgewater, Maryland 21037 USA; *email:* KingRy@si.edu

Agency (USEPA) has recognized the shortcomings of this former approach and its inconsistency with goals of the CWA (USEPA 1998a). In response, the USEPA has issued a comprehensive plan for the development of scientifically defensible, numerical water-quality criteria. The plan emphasizes the need for the inclusion of assemblage-level endpoints in criteria development, and that the criteria need to be stratified into different regions and types of water bodies (USEPA 1998a). Metrics used in bioassessment may be well suited for this purpose.

Here we extend the multimetric bioassessment concept to directly supporting the development of numerical water quality criteria (Barbour and others 1995). Unlike traditional multimetric approaches, which are based primarily on observational data, our approach relies on a coupling of observational and experimental datasets to elucidate potential cause-effect linkages (e.g., Daehler and Strong 1996; Lemly and Richardson 1997; Beyers 1998; Adams and Greeley 2000). This approach allows the development of metrics that are diagnostic and stressor-specific, a limitation of most bioassessment techniques in use today. For example, multimetric indexes have historically been developed along gradients of general types of human influence (e.g., urban land-use) over a broad geographic area (Karr and Chu 1997). While the description of biotic responses to general disturbance is useful for assessing status and trends of aquatic health, these assessments were not developed to characterize the effects of specific stressors on biological endpoints (Norton and others 2000; USEPA 2000b; Griffith and others 2001). Thus, traditional multimetric indexes have a limited capacity to diagnose causes of impairment or estimate the risk associated with a stressor (Suter 2001). Therefore, our goal was to identify biological attributes that responded to a specific stressor in a specific region and water body type. These attributes would serve as measurement endpoints to estimate levels of a stressor that may result in a high risk of degradation to biological integrity (USEPA 1998b).

To illustrate the approach, we estimated levels of surface-water total phosphorus (TP) that affected macroinvertebrate assemblages in wetlands of a nutrient-sensitive ecoregion using existing, published datasets (King and Richardson 2002; Qian and others, in press; King and Richardson, in press). We defined macroinvertebrate structure and function as our assessment endpoint, assemblage attributes as measurement endpoints, and TP as the stressor—however, any biological endpoint or stressor of concern could be substituted. Ultimately, the broad objective of this paper is to show how assemblage-level data can be used in a risk-based framework to quantify potential ecological thresholds, which, in turn, can be used to support environmental decision-making.

Methods

Study Area

Data used for this study were collected in Water Conservation Area 2A (WCA-2A) in the northern Everglades of Florida, USA (26° 15' N, 80° 23' W). WCA-2A is located in the south Florida coastal plain nutrient ecoregion, an area considered P-sensitive by USEPA (2000a). WCA-2A is a 43,280 ha diked wetland landscape, with water-control structures governing the inflow and outflow of surface water. Inflow primarily occurs along the northern levee through three water-control structures (S10-A, C, and D) on the Hillsboro Canal, a conduit for outflow from Lake Okeechobee and P-enriched runoff from the Everglades Agricultural Area (EAA). Inflow from the Hillsboro Canal has induced a steep longitudinal eutrophication gradient in WCA-2A due primarily to excessive inputs of P (SFWMD 1992). Surface-water and soil P has been shown to be elevated above natural, background concentrations up to 7 km into the interior of WCA-2A (e.g., DeBusk and others 1994; McCormick and others 1996; SFWMD 2000). TP in these interior, reference areas of WCA-2A typically ranges between 5–10 μg/L, while often exceeding 100 μg/L in areas near inflow structures on the Hillsboro Canal (Vaithiyanathan and Richardson 1998; SFWMD 2000). Maps and greater detail about physical and chemical characteristics of the study area are provided in Davis and Ogden (1994), Richardson and others (1999), SFWMD (2000), and King and Richardson (2002).

Observational Data

The first dataset was observational and collected along a 10-km TP gradient in WCA-2A by King and Richardson (2002). In this study, 126 stations were sampled for surface-water TP (μg/L) and macroinvertebrate assemblage composition (density of taxa). Sampling stations extended from a highly impacted region near the canal inflow structures into the interior of WCA-2A, which was defined as a reference area (e.g., SFWMD 2000; King and Richardson 2002). For this analysis, we only used stations located in open-water sloughs (n = 37) to reduce variability associated with different habitats and because the experimental data also were limited to sloughs.

Surface-water TP sample collection, sample storage, and analysis (external standards, blanks, spikes) were in accordance with QA/QC protocols mandated by the Florida Department of Environmental Protection and standard methods (APHA 1992). Due to the large spatial extent of this study, TP was collected only once during October 1998. TP concentrations at slough stations ranged from 4.5 to 50.4 µg/L, consistent with long-term observations for sloughs along this P gradient (SFWMD 2000).

Macroinvertebrate sampling and sample processing were based on a slight modification of protocols used by FDEP (1996; SOP #BA-7, 8) and USEPA (1997b; Barbour and others 1999). A D-framed dip net was used to collect a 1.5-m ^{2} composite sample from each station. Sampling was conducted in October of 1998, simultaneous with TP collections. Macroinvertebrates were identified to the lowest practical taxonomic level (usually species), and data were expressed as number of individuals/m ^{2}. A total of 202 taxa from 37 samples were included in the slough-station dataset. Greater detail on methods is presented in King and Richardson (2002).

Experimental Data

The experimental dataset was obtained from a P-dosing study in the interior, reference area of WCA-2A where TP concentrations average < 10 µg/L (Vaithiyanathan and Richardson 1998; Richardson and others 2000; King and Richardson, in press; Qian and others, in press). Two P-dosing sites, each with six mesocosms (12 mesocosms in total), were constructed in adjacent open-water sloughs in 1992. Mesocosms were 2-m wide and 8-m long flumes and were constructed around natural, undisturbed slough habitat. Mesocosms were oriented parallel to surface-water flow and closed at the upstream end. P was dosed from the closed end of each mesocosm downstream toward the open end. P was dosed in the form of soluble reactive phosphate (SRP) continuously from 1992–1998. Each flume was assigned one of six P treatments, ranging from walled and unwalled control treatments (no P added above background concentrations; 0.25 g/m ^{2}/y TP) up to 8.2 g/m ^{2}/y P. This design created experimental P gradients both among and within mesocosms (i.e., gradients in concentrations down the length of each flume due to uptake and dilution).

Sampling stations were established at positions 2, 4, and 6 m down the length of each mesocosm (36 sampling stations in total). Thus, measured P concentrations at stations were a product of physical, biogeochemical, and biological factors that resulted from different, controlled input concentrations, just as along the P gradient (Richardson and others 2000). For this analysis, this was desirable because our research question specifically dealt with estimating an ecological threshold based on a measured concentration of TP in surface waters, as mandated by the Everglades Forever Act (1994) and USEPA (2000a). Because each station had unique TP and macroinvertebrate data associated with it and spatial autocorrelation among stations was minimal (King and Richardson, in press), stations could validly be considered independent observations (Hurlbert 1984).

Surface-water TP was collected biweekly at each sampling station throughout the majority of the six-year experiment following QA/QC protocols used in the observational gradient study. Because there were many observations from each station, TP data were expressed as geometric means in accordance with the Everglades Forever Act (EFA 1994) and USEPA guidelines (2000a). Geometric means were calculated for a six-month period prior to each macroinvertebrate collection, and provided an integrated estimate of long-term TP exposure at each station (note, however, that USEPA [1998b] recommends arithmetic means for describing chronic exposure, thus geometric means were conservative estimates of TP exposure). Geometric means of TP ranged from 5.8 to 60.9 µg/L among stations, very similar to the range of values observed along the P gradient.

Macroinvertebrates were collected four times during 1996–1998 (September 1996, January 1997, February 1998, September 1998) at each station using FDEP protocols (1996; SOP #BA-13). Macroinvertebrates were collected using Hester-Dendy (HD) artificial substrates because active sampling methods (e.g., dip nets) would have significantly disturbed the habitat in the experimental mesocosms. King (2001) demonstrated that HD samples were effective for characterizing the macroinvertebrate assemblage of Everglades sloughs. In addition, assemblage attributes evaluated in the experiment were required to respond in the same manner (increase or decrease and temporal repeatability) to TP as those measured along the observed gradient to be considered metrics. Moreover, we were not interested in comparing the absolute values of attributes between the two studies; rather, we were interested in the levels of TP that elicited changes in attribute values, which is completely independent of any potential differences in the magnitudes of attribute values. Thus, any biases associated with differences in sampling methods between the two studies were eliminated because attributes selected as metrics were demonstrated to show the same response using both methods.

A composite of three HD samples were collected from each station (n = 36) on each date (n = 4). Macroinvertebrates were identified to the lowest prac-

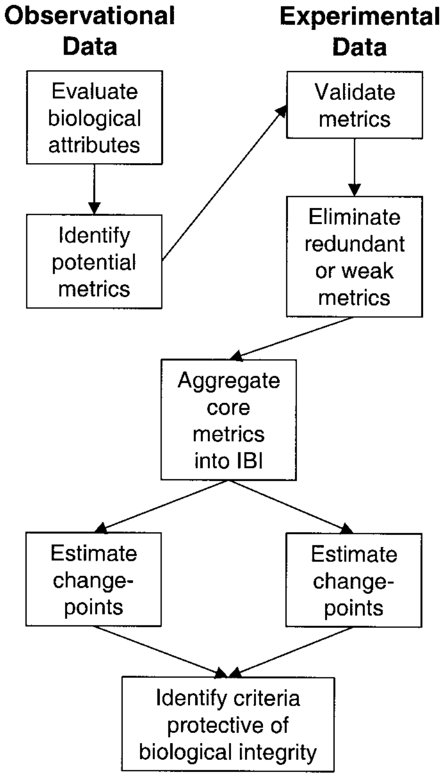
Observational Data
Experimental Data
Evaluate biological attributes
Validate metrics
Identify potential metrics
Eliminate redundant or weak metrics
Aggregate core metrics into IBI
Estimate change-points
Estimate change-points
Identify criteria protective of biological integrity

Figure 1. Conceptual framework for developing numerical water-quality criteria using bioassessment.

tical taxonomic level (usually species), and expressed as number of individuals/m ^{2}. A total of 123 taxa from 144 samples were included in the experimental dataset. Comparisons of macroinvertebrate assemblage composition between unwalled and walled control mesocosms showed that composition in the walled controls was not different than the unwalled controls (King and Richardson, in press). Thus, the experiment was representative of the reference condition. Greater detail on the P-dosing experiment is provided in Richardson and others (2000), King (2001), and King and Richardson (in press).

Metric Development and Analytical Approach

Our approach was patterned after the conceptual framework of multimetric development outlined by Barbour and others (1995). Our initial step (Step 1) was to select a suite of assemblage-level attributes and use the observational data to evaluate the response of these attributes to TP (Figure 1). We supposed that if attributes did not exhibit a response in the "real" world, then these should not be tested experimentally (Daehler and Strong 1996; Lemly and Richardson 1997; Adams and Greeley 2000). Thus, Step 2 was the identification of a suite of *candidate* metrics, which would then be scrutinized more fully using the experimental data. Attributes that met several selection criteria using the experimental data were subsequently validated as TP metrics (Steps 3 and 4). Selected metrics were aggregated into an IBI-type multimetric index, which we termed a *Nutrient-IBI,* in addition to being assigned as individual biological endpoints for analysis (Step 5). Data from both the observational and experimental studies were then analyzed using changepoint analysis to estimate levels of TP that could be expected to change biological condition (Step 6). We defined a detectable change in the mean and/or variance of an attribute of macroinvertebrate structure and function, coupled with uncertainty estimates, as an indication of an ecological threshold response to TP. Because our data spanned observed and experimental gradients from reference conditions (TP < 10 μg/L) to highly P-enriched conditions, we argued that such changes represented a significant deflection from the reference condition, and consequently, degradation of biological integrity. This argument was also consistent with the Everglades Forever Act (1994), which mandated that a TP criterion for this region should not result in an imbalance of flora and fauna representative of the natural Everglades. Thus, in the final step (Step 7) we synthesized results from the changepoint analysis to identify levels of TP that were likely to be protective of biological integrity, as reflected by the metrics of macroinvertebrate structure and function.

Step 1. Select assemblage attributes. The first step toward metric evaluation was to select a variety of attributes that represented key elements of the structure and function of macroinvertebrate assemblages found in the reference area of the observed P gradient. Attributes were selected from four general classes: (1) taxonomic composition, (2) species richness and diversity, (3) tolerance/intolerance, and (4) trophic structure (Barbour and others 1999). In all, over 50 attributes were selected, with the majority falling under the taxonomic composition category. As recommended by Barbour and others (1995), we used relative (percent) rather than absolute abundance in calculating attributes except those of richness/diversity because percentage metrics have been shown to be more robust and reliable and were more likely to reflect structural changes resulting from nutrients. Barbour and others (1999) provided a summary of potential benthic metrics, which helped direct our selection process.

Composition attributes, expressed as percent of total numerical abundance, were selected according to the dominant major taxonomic groups present in the study, which corresponded to families (e.g., percent

Chironomidae), orders or classes (e.g., percent Odonata), or a combination of higher groups with relatively similar habits or food preferences (e.g., percent Microcrustacea).

An additional composition attribute was Bray-Curtis dissimilarity (BCD, percent dissimilarity), a coefficient shown to be a robust and ecologically interpretable index of changes in taxonomic composition (Faith and others 1987; Legendre and Legendre 1998). BCD was calculated using the taxa (n = 202) abundance data (standardized using $\log_{10}$ (x + 1) transformation; Legendre and Legendre 1998). Because it is multivariate, BCD was ordinated using nonmetric multidimensional scaling (nMDS), rotated using varimax rotation, and extracted as univariate scores along nMDS Axis 1 (McCune and others 1997; Legendre and Legendre 1998). The objective in the use of nMDS was to recover a multivariate assemblage pattern that corresponded to a gradient in TP concentration, and to produce individual sample scores that could be used for analysis.

Richness and diversity attributes included total number of taxa (richness per unit area, or areal richness; Larsen and Herlihy 1998), numerical richness (total number of taxa per 300 individuals, or NR300; Larsen and Herlihy 1998), Shannon-Wiener diversity (H'), and number of taxa belonging to several major taxonomic groups (e.g., number of taxa of Chironomidae).

Tolerance/intolerance attributes were derived using a list of taxa (species) shown to either disappear at low levels of P enrichment or proliferate with high levels of P enrichment in the Everglades (King 2001). A relatively small proportion (< 20%) of taxa were considered either highly tolerant or highly sensitive. These attributes were expressed as the percent of total numerical abundance comprised of taxa shown to be tolerant or sensitive to P enrichment.

Trophic-structure attributes were selected according to the predominant functional feeding groups in the study area, which were predators, filterers, scrapers, and gatherers (Merritt and Cummins 1996; Barbour and others 1999). Trophic attributes were expressed as percent of total numerical abundance.

Step 2: Identify potential metrics. As recommended by several authors who have developed multimetric indexes (Barbour and others 1996; Fore and others 1996; Karr and Chu 1997), we graphically evaluated the response of assemblage attributes to TP concentrations along the observed gradient. Attributes with values that either increased or decreased monotonically with TP were identified as potential metrics. Attributes that either did not respond or showed very weak responses were eliminated from consideration. Attributes that responded unimodally were also discarded because values were similar at low and high concentrations of TP.

Step 3: Validate metrics. The suite of potential metrics identified from the observed gradient were further evaluated using the experimental data. We graphically examined each attribute separately for each of the four sampling dates. Attributes were discarded if they did not respond, showed very weak responses, or showed unimodal responses to TP on more than one sampling date. Attributes that responded in a different direction than the observed gradient (e.g., an increase with TP in the experiment while a decrease with TP along the observed gradient) were deemed too variable and also discarded. Thus, attributes that met all criteria as metrics responded to TP (1) in the real world, (2) in an experimental setting, (3) in the same direction in both studies, and (4) repeatedly over time.

Step 4: Eliminate redundant metrics. Metrics used in a multimetric index are intended to individually capture some variation not explained by other metrics. Collinear metrics do not add new information to an index, and may weight it too heavily toward one facet of biological condition. Thus, it was important to cull metrics that were excessively redundant before proceeding. A Pearson product-moment correlation matrix was used to evaluate collinearity among metrics ($r > 0.90$; Kleinbaum and others 1988; Barbour and others 1996). When pairs or sets of metrics were deemed collinear, the metric that showed the strongest, most consistent response to TP was retained.

Step 5: Aggregate core metrics into IBI. Metrics that met all selection criteria formed a core set to construct a multimetric index, which we termed a Nutrient Index of Biological Integrity (Nutrient-IBI). Typically, metric values are assigned a tiered score of 1, 3, or 5, ranging from poor to good, based on an arbitrary cutoff for each of the three tiers (Barbour and others 1995; Karr and Chu 1997). While this approach has been shown to be effective, we chose to scale the continuous metric values from 0 to 1 (low to high condition) to avoid making value judgments about tiers of condition (Suter 2001). This scaling procedure gave each metric continuous values and equal weight in the IBI. Metrics with low values at low TP were first inverted so that the raw minimum value was scaled to highest condition. The IBI score was the sum of all metric values for each observation, scaled from 0 to 5 (5 = highest condition).

Step 6: Estimate changepoints. We estimated potential threshold responses in the measurement endpoints to numerical levels of TP using nonparametric changepoint analysis (nCPA), a technique explicitly designed for detecting threshold responses using ecological data (Qian and others, in press). Nonparametric changepoint analysis is a derivative of a family of techniques historically used in classification and divisive partitioning of ecological data (e.g., Pielou 1984). This analysis is based on the idea that a structural change in an ecosystem may result in a change in both the mean and the variance of an ecological response variable used to indicate a threshold. When observations are ordered along a stressor gradient, a changepoint is a value that separates the data into the two groups that have the greatest difference in means and/or variances. This can also be thought of as the degree of within-group variance relative to the between group variance, or *deviance* (*D*) (see Venables and Ripley 1994 and Qian and others, in press, for details). Analytically, the nCPA examines every point along the stressor gradient and seeks the point that maximizes the reduction in deviance. Thus, each stressor value is a potential changepoint and is associated with a deviance reduction:

$$\Delta_i = D - (D_{\leq i} + D_{> i}), \qquad (1)$$

where D is the deviance of the entire data set $y_1, \cdots, y_n$, $D_{\leq i}$ is the deviance of the sequence $y_1, \cdots, y_i$, and $D_{>i}$ is the deviance of the sequence $y_{i+1}, \cdots, y_n$, where $i = 1, \cdots, n$. The changepoint r is the i value that maximizes $\Delta_i : r = \max_i \Delta_i$.

There is one particular value of the predictor y (in this case, TP) that maximizes the reduction in deviance in the response data (in this case, the selected metrics); however, there is uncertainty associated with that value. It is unlikely that any one value of TP is the only value that could represent a changepoint. In reality, depending on the acuteness of the biological change in response to TP, several observations of TP could represent the changepoint, each with varying probabilities. Thus, to assess the risk associated with particular levels of TP, nCPA incorporates estimates of uncertainty in the changepoint (Qian and others, in press). These estimates are calculated using a bootstrap simulation (Efron and Tibshirani 1993). This simulation resamples (with replacement) the original dataset and recalculates the changepoint with each simulation. Bootstrap simulations are repeated 1,000 times. The result is a distribution of changepoints that summarizes the uncertainty among multiple possible changepoints. This uncertainty is expressed as a cumulative probability of a changepoint based on the relative frequency of each changepoint value in the distribution. To illustrate, a cumulative probability curve is shown in Figure 2 for the percent sensitive taxa metric in response to TP from the observed P gradient. Here, there is at least a 5% cumulative probability, or risk, that a detectable change in the percentage of sensitive taxa occurs at or below 13.3 µg/L TP. In other words, 5% of the bootstrap simulations resulted in a changepoint that was ≤ 13.3 µg/L TP. To fully visualize the range of uncertainty, the cumulative probability curve is extended to the highest level of TP that resulted in a changepoint in at least one of the simulations (Figure 2). Thus, the cumulative probability curve depicts the range of TP values that could potentially represent a changepoint and illustrates a cumulative level of risk associated with each TP value.

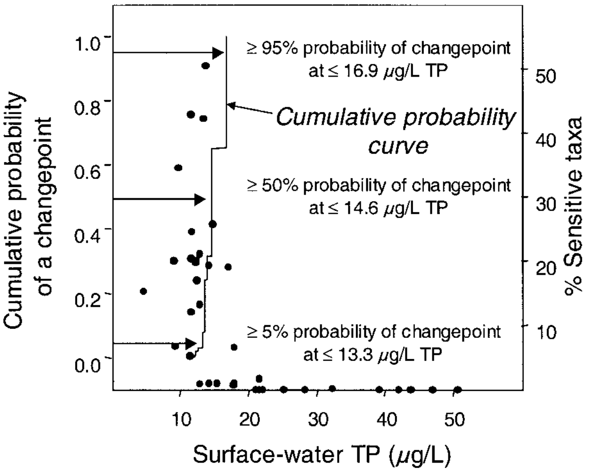

Figure 2. Illustration of the cumulative probability of a changepoint estimated for an individual metric in response to surface-water TP. The cumulative probability curve describes the cumulative risk of a change in a response variable (% sensitive taxa, y-axis [right side]; depicted by filled circles) associated with a range of stressor values. Cumulative probabilities are calculated using 1,000 bootstrap simulations. Any given location along the curve corresponds to a specific cumulative probability of a changepoint (y-axis [left side]) at a specific level of TP (x-axis). In this example, there was at least a 5% cumulative probability, or risk, that a detectable change in the mean and/or variance of the % sensitive taxa metric occurred at or below 13.3 µg/L TP. In other words, ≥5% of the bootstrap simulations resulted in a changepoint that was ≤ 13.3 µg/L TP. Similarly, there was ≥50% risk of a changepoint ≤ 14.6 µg/L TP, while there was ≥95% probability that a changepoint occurred ≤ 16.9 µg/L TP. Data are from the observed P-gradient study.

An additional factor to consider when using nCPA is an estimate of the probability of Type I error. A χ^2 test statistic (1 df) can be used to evaluate the likelihood that an observed changepoint is real (Qian and others, in press). However, we only used this statistic to help assess the likelihood that changepoints with relatively wide cumulative probability distributions represented real biological changes, as uncertainty around the changepoint was a much more relevant issue (Suter 1996; Germano 1999; Johnson 1999).

Changepoint analysis works best when stressor-response relationships are nonlinear or heteroscedastic, properties very common to ecological data. For strong linear relationships, the analysis will find a significant changepoint but uncertainty will be high. Preliminary examination of the observational and experimental data revealed that all relationships were nonlinear and/or heteroscedastic, thus were well suited for nCPA. We estimated changepoints for individual metrics and the IBI using the observational and experimental datasets. Analyses were conducted for each date separately using the experimental data to better evaluate temporal variability in threshold responses. Analyses were conducted using the custom function "chngp.nonpar" (Qian and others, in press) in S-Plus 2000 (Mathsoft, Inc., Seattle, WA, USA).

Step 7: Identify criteria protective of biological integrity. We graphically concatenated the results from the observational and experimental studies to help identify levels of TP that were protective of biological integrity, as reflected by the metrics of macroinvertebrate structure and function. We interpreted a cumulative probability of a changepoint ≥50% to imply that a threshold response for a certain endpoint was more likely than not to occur at the respective predictor level of TP. We evaluated the range of TP levels that resulted in a ≥50% likelihood of a threshold response for individual metrics and the IBI, and contrasted this range of values between the observational and experimental data. Similarly, we contrasted the range of TP levels that had low (5%) and high (95%) probabilities of resulting in a threshold response to better characterize the risk to macroinvertebrate structure and function. However, it is important to note that the level of risk that scientists, managers, and decision-makers may be willing to accept will most certainly depend on a variety of ecological, economic, and social factors. Thus, our evaluation of cumulative probabilities of a changepoint at 5%, 50%, and 95% should not be implied to be an endorsement of these levels as the only levels of risk that should be evaluated in the criteria development process. Our approach provides a continuum of risk for each level of a stressor, and our focus on these three levels was largely necessitated by the limitation in presenting levels of risk for every possible changepoint.

Results

Ten of the metrics evaluated using the observational P-gradient data showed clear responses to TP and were identified as potential metrics. Of these 10 candidate metrics, five exhibited consistent responses to TP in the P-dosing experiment: BCD, percent tolerant taxa, percent sensitive taxa, percent Oligochaeta (aquatic worms), and percent predators. Results from correlation analysis among these five metrics indicated that no pair was collinear ($r < 0.90$), thus each metric was sufficiently unique to retain as core metrics. These five metrics were subsequently analyzed individually and as an aggregated Nutrient-IBI using nCPA.

Changepoints were detected for all selected metrics and the IBI using the observational P-gradient data (Table 1). Probabilities of Type I error (*P* in Table 1) were all quite low, indicating that it was highly likely that changepoints were real and represented a threshold response. The cumulative probability distributions generated from nCPA indicated that a changepoint was ≥50% probable between 12.6 and 19.4 μg/L TP for individual metrics and 14.8 μg/L TP for the IBI (Table 1, Figures 2–5). These changepoints represented biologically significant shifts in assemblage structure and function. Sensitive taxa dropped from a mean of over 21% to only 1.3% above 14.6 μg/L TP (Table 1, Figure 2). Conversely, tolerant taxa increased from only 2.2% to nearly 20% above 17.7 μg/L TP (Table 1, Figure 3). Percent Oligochaeta, a group of aquatic worms, nearly doubled when TP exceeded 13 μg/L. Mean BCD values (nMDS Axis 1 scores) were highly negative to the left of the 50% probability of a changepoint, while highly positive to the right, indicating a markedly different species assemblage once a cumulative probability of 50% had been exceeded (Table 1). Elevated TP also resulted in functional changes, reducing the proportion of predators in the assemblages from a mean of 9.2% to only 3.4% at TP levels above 12.6 μg/L. Finally, mean IBI scores above 14.8 μg/L were reduced by one-half when compared to IBI scores below that concentration (Table 1, Figure 4). In addition to these changes in means, all of these metrics exhibited distinct changes in variances that corresponded to TP changepoints (e.g., Figures 2–4).

Results from the P-dosing experiment mirrored those of the observed P gradient. Changepoints were evident for all metrics and the IBI, and were repeatable through time. Overall, median threshold responses from the four dates were ≥50% probable between 11.2 and 13.0 μg/L TP for individual metrics and 12.3 for the IBI (Table 1, Figures 3–5). Means and variances of metric values above and below the 50% level of risk were very similar to the biologically significant changes observed along the P gradient, and highly suggested that the changepoints represented threshold responses to TP (Table 1).

The cumulative probability distributions of changepoints indicated that there was a relatively tight range of TP levels likely to result in degradation in biological

Table 1. Results from nonparametric changepoint analysis showing cumulative probabilities of a threshold response for individual metrics and the aggregated IBI at specific levels of TP from the experimental and observational studies

		Cumulative Probability of a Changepoint (TP, μg/L)				Mean Metric Value (± 1 SE)[a]	
Metric	Study (Date)	5%	50%	95%	*P**	Left	Right
Bray-Curtis Dissimilarity (BCD)[b]	Experimental (Sep 1996)	10.1	12.3	18.4	0.0012	−0.75 (0.16)	0.44(0.15)
	Experimental (Jan 1997)	11.1	11.6	12.8	0.0001	−0.95 (0.14)	0.48(0.09)
	Experimental (Feb 1998)	10.1	10.5	10.7	0.0007	−0.80 (0.20)	0.54(0.10)
	Experimental (Sep 1998)	8.3	10.8	13.9	0.0006	−0.79 (0.23)	0.46(0.12)
	Observational (Oct 1998)	15.2	19.4	21.4	0.0002	−0.98 (0.13)	0.61(0.25)
% Sensitive Taxa	Experimental (Sep 1996)	7.4	14.5	25.7	0.1207	9.2 (3.8)	3.4 (2.3)
	Experimental (Jan 1997)	8.7	11.3	11.8	0.0032	21.3 (4.8)	8.0 (1.5)
	Experimental (Feb 1998)	7.1	11.4	18.7	0.0122	7.9 (1.6)	3.6 (1.1)
	Experimental (Sep 1998)	6.8	9.8	11.6	0.0414	4.7 (1.7)	0.9 (0.4)
	Observational (Oct 1998)	13.3	14.6	16.9	0.0013	21.2 (3.1)	1.3 (1.1)
% Tolerant Taxa	Experimental (Sep 1996)	10.2	12.1	19.0	0.0016	5.6 (2.4)	24.2 (3.3)
	Experimental (Jan 1997)	11.3	14.0	16.4	0.0008	7.5 (1.8)	23.3 (3.4)
	Experimental (Feb 1998)	9.1	10.7	12.1	0.0162	3.3 (1.5)	18.5 (4.1)
	Experimental (Sep 1998)	7.1	10.7	14.4	0.0098	7.2 (3.0)	20.9 (3.8)
	Observational (Oct 1998)	14.6	17.7	25.5	0.0002	2.2 (1.2)	20.0 (5.0)
% Oligochaeta	Experimental (Sep 1996)	9.6	13.3	25.7	0.0026	7.3 (6.2)	41.6 (8.4)
	Experimental (Jan 1997)	8.8	12.7	18.1	0.0154	17.6 (5.0)	38.5 (7.0)
	Experimental (Feb 1998)	9.0	18.4	21.6	0.0262	22.2 (4.2)	40.0 (7.2)
	Experimental (Sep 1998)	8.3	12.6	13.9	0.0035	7.7 (4.4)	41.5 (6.0)
	Observational (Oct 1998)	11.4	13.0	16.9	0.0212	33.9 (5.3)	57.9 (4.2)
% Predators	Experimental (Sep 1996)	7.6	11.1	18.1	0.0162	21.0 (12.2)	4.7 (1.4)
	Experimental (Jan 1997)	7.9	11.7	12.8	0.0006	18.3 (4.5)	5.0 (1.1)
	Experimental (Feb 1998)	7.1	10.2	10.6	0.0221	8.7 (1.2)	4.4 (0.6)
	Experimental (Sep 1998)	8.3	14.5	21.2	0.1694	9.4 (3.3)	5.6 (2.6)
	Observational (Oct 1998)	8.9	12.6	16.4	0.0419	9.2 (3.7)	3.4 (1.2)
Nutrient Index of Biological Integrity (Nutrient–IBI)	Experimental (Sep 1996)	11.9	13.6	13.8	0.0003	3.4 (0.2)	1.9 (0.1)
	Experimental (Jan 1997)	9.7	11.7	12.8	0.0002	3.9 (0.2)	2.2 (0.1)
	Experimental (Feb 1998)	9.1	10.6	12.1	0.0122	2.9 (0.2)	2.1 (0.1)
	Experimental (Sep 1998)	10.0	13.0	14.2	0.0020	2.9 (0.1)	1.8 (0.1)
	Observational (Oct 1998)	12.3	14.8	16.9	0.0004	3.0 (0.2)	1.5 (0.2)

**P* = Probability of Type I error, indicating the likelihood that there was no changepoint in the response data.

[a] Mean (± 1 SE) metric values to the left and right of the level of TP corresponding to ≥50% cumulative probability of a changepoint.

[b] BCD values were expressed as standardized nMDS Axis 1 scores (see Methods for greater detail).

condition (Table 1, Figures 3–5). Both the observational and experimental data revealed that there was a low (5%) probability that a threshold response occurred ≤ 10 μg/L TP for some metrics. There was high (≥95%) certainty that the threshold was ≤ 20 μg/L TP for the majority of individual metrics. Aggregating the individual metrics into the IBI reduced this range of variability, however. Results indicated a 5% probability that a threshold response for the IBI occurred at or below 9 μg/L TP (experimental) and 12.3 μg/L TP (observational), whereas there was ≥95% certainty that a threshold response occurred ≤ 15 μg/L TP (experimental) and ≤ 17 μg/L TP (observational) (Table 1, Figures 4 and 5). Although these differences were relatively small, the lower changepoints from the P-dosing experiment than the observed P gradient implied that changepoints from the P-dosing experiment might have been conservative estimates of TP levels that may pose a risk to macroinvertebrate structure and function.

Discussion

Can Bioassessment Be Used To Develop Numerical Water-Quality Criteria?

Bioassessments generally are performed with the intent of detecting impairment in an aquatic ecosystem, which usually implies degraded water quality. Despite the fundamental linkage between bioassessment and water quality, there are surprisingly few examples of

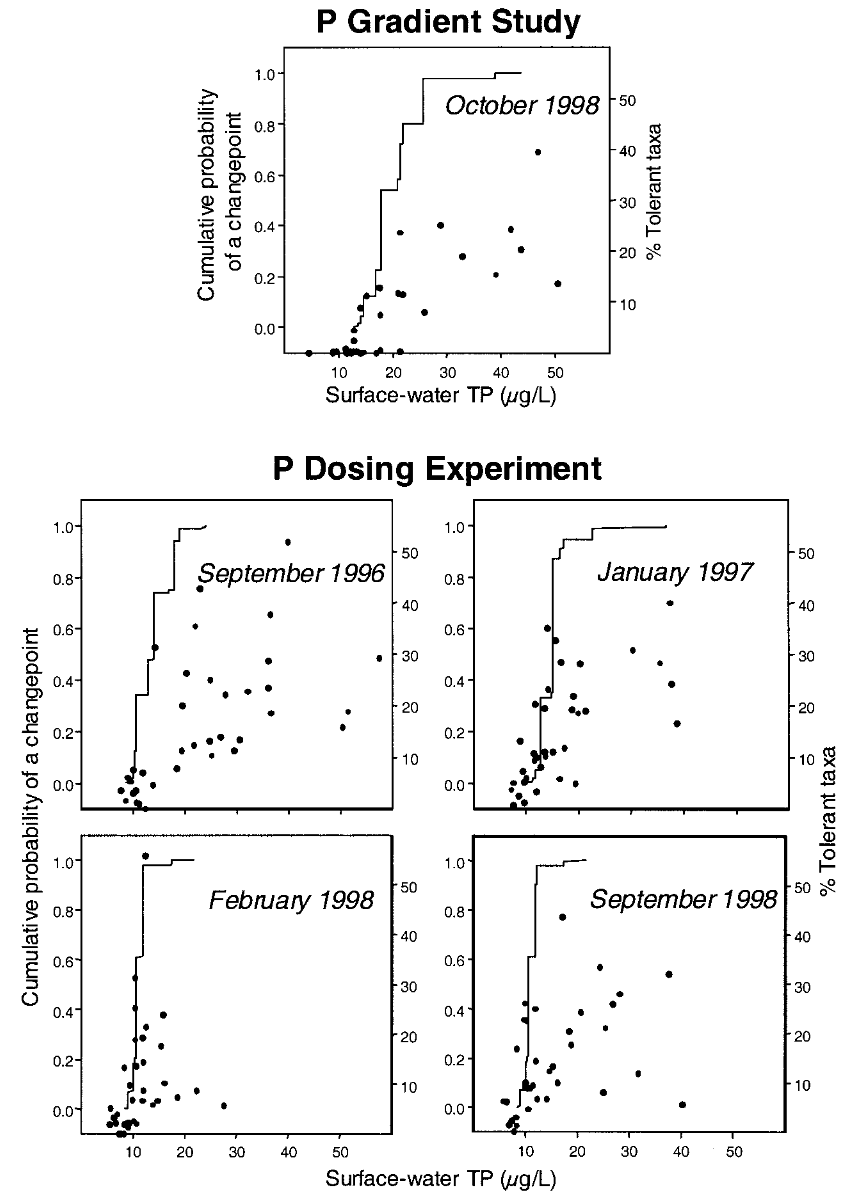

Figure 3. Cumulative probabilities of changepoints for the percent of tolerant taxa metric in response to surface-water TP. Results are shown for the observational P-gradient study and the four dates from the P-dosing experiment.

bioassessment used explicitly to support the development of numerical water-quality criteria (see Dodds and Welch 2000). One of the primary reasons for this is that traditional bioassessments, such as multimetric approaches, are intentionally developed to capture the effect of a wide range of stressors to biological integrity. This lack of specificity results in ambiguity about the potential cause(s) of impairment and, consequently, the levels of a stressor that may result in a threshold response. However, the results of our study provide evidence the multimetric approach to bioassessment is robust and appears to be easily adaptable to a particular stressor, such as nutrients. We identified several attributes of macroinvertebrate assemblages that responded to surface-water TP using observational, real-world data. Experimental data provided evidence that changes observed in the observational study were indeed due to P enrichment. Temporal replication from the experiment also indicated that, despite seasonal variation, attributes responded in a consistent direction (increase or decrease) to TP. These core metrics also responded repeatedly over time to TP. Finally, threshold responses were detected at similar levels of TP among different metrics and across several dates. Thus, this approach was consistent with the water-quality criteria development strategy proposed by the USEPA

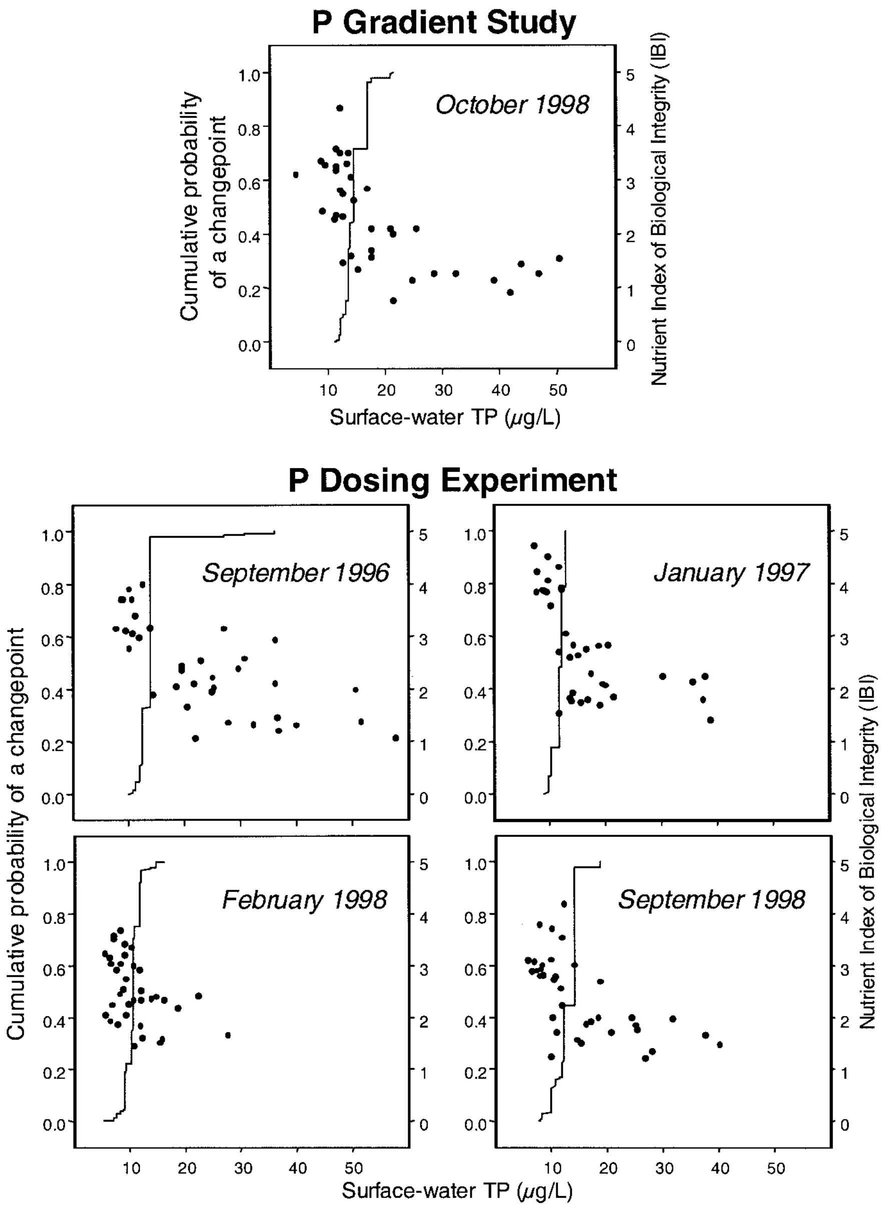

Figure 4. Cumulative probabilities of changepoints for the Nutrient-IBI in response to surface-water TP. Results are shown for the observational P-gradient study, and the four dates from the P-dosing experiment.

(1998a), as our findings (1) established a cause-effect linkage between TP and biological attributes within a given nutrient ecoregion, (2) estimated levels of TP that may cause biological changes, and (3) estimated uncertainty in TP levels that may lead to degradation of biological integrity.

The use of ecological experiments may be the most critical step in the validation of numerical criteria using bioassessment. Descriptive, correlative studies are often very useful for generating hypotheses but often are insufficient for establishing cause-effect linkages (e.g., Beyers 1998). A number of recent studies have shown creative ways to use descriptive biomonitoring data to ascribe causation using a stressor-identification framework (e.g., Norton and others 2000; Griffith and others 2001; Cormier and others 2002). However, without experimental evidence, it is still very difficult to eliminate other potential causes of an observed biological response to a candidate stressor (USEPA 2000b). Moreover, it is nearly impossible to quantify the uncertainty associated with additive or synergistic effects of multiple stressors in an aquatic ecosystem without first isolating a single stressor using an experiment. This latter point is particularly critical in the context of numerical criteria development because the level of a stressor that apparently results in an observed threshold response

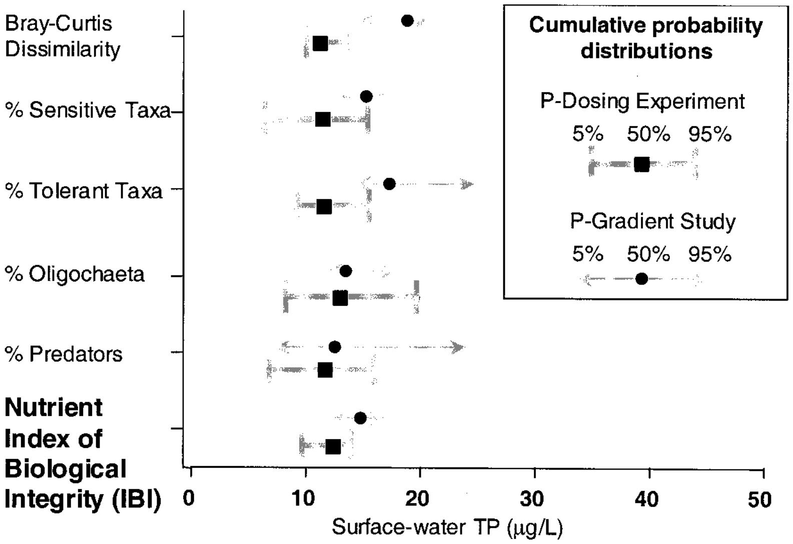

Figure 5. Synthesis of results from the P-gradient study and P-dosing study for the identification of a TP criterion protective of biological integrity. Median values from the four dates in the P-dosing experiment were used for the ≥5%, 50%, and 95% cumulative probabilities.

may be confounded by another, perhaps unmeasured, factor (Suter 2001). For these reasons, we highly recommend the collection of experimental data to support observed assessments used for numerical criteria development.

Conversely, experimental studies suffer from some important limitations. Most are not conducted at the appropriate scale (e.g., watershed) and need to be coupled with observational research to help validate the applicability of experimental findings to the real world (e.g., Daehler and Strong 1996; Lemly and Richardson 1997; Adams and Greeley 2000). In our approach, we relied on a descriptive study to identify biological attributes that may have been affected by TP. Once identified, these biological attributes were further examined in a long-term P-dosing experiment to corroborate their P sensitivity and estimate TP changepoints. Without the observational study, however, it would have been difficult to extrapolate the experimental findings to the much larger scale of the study area. By coupling the two studies, each provided evidence that the other study could not, which made for a much stronger case about the levels of TP that were likely to degrade macroinvertebrate structure and function.

One potential criticism of our approach is that it may be impractical for state, tribal, other regulatory agencies that have limited funding to conduct long-term exposure studies such as the P-dosing experiment we illustrated here. While large-scale (spatial, temporal or both) experiments are probably too costly in most situations, small in situ microcosm or mesocosm studies may provide sufficient evidence to support an observational finding. For example, Clements and others (2002) provided an excellent illustration of the coupling of small experiments with descriptive data. Here, lab experiments, small in situ exposure experiments, and large-scale observational studies afforded strong inference about the levels of heavy metals that affected biota in Rocky Mountain streams. Similar examples also exist for nutrients (e.g., Hart and Robinson 1990; Perrin and Richardson 1997; Lemly and King 2000). Thus, it seems that experiments can be a practical addition to the criteria development process if efficiently and purposefully designed.

Detecting Threshold Responses with Changepoint Analysis

Estimation of risk should be a critical step in developing numerical water-quality criteria. Risk analysis requires a tangible, numerical estimate of the levels of a stressor that are likely to result in an effect on an assessment endpoint. However, the most commonly employed types of data analyses—hypothesis-testing statistics—are insufficient and possibly misleading when used for this purpose (e.g. Germano 1999; Johnson 1999). Suter (1996) provided a thorough review of the problems with hypothesis testing in ecological risk assessment, most notably the inability of the approach to provide a clear estimate of expected or observed effects and associated uncertainties related to a predictor variable. In contrast, our results suggest that nCPA has potential to be a useful analytical tool in the development of criteria because of the easily interpretable, numerical estimates it affords. Rather than asking the question "is there a statistically significant relationship between predictor *x* and response *y*?" as implied with most hypothesis-testing statistics, this risk-based analysis more explicitly asks "what level of predictor *x* results in a threshold response of *y*, and how uncertain is this threshold?" Using this analysis, we were able to identify levels of TP that were likely to result in a threshold response in the macroinvertebrate assemblage as well as provide an estimate of the cumulative probability that a particular level of TP would elicit a threshold response. Although we included a χ^2 significance test (1 df) to assess the likelihood that changepoint was real, this test was of limited value because such tests provide little information about the risk of a threshold response at various levels of TP. Thus, we contend that results from hypothesis testing fail to provide enough information to decision-makers, and generally be avoided for supporting the development of numerical criteria.

Another advantage of nCPA is that it is particularly appropriate for ecological data analysis because it makes few assumptions about the distributional properties of data (Qian and others, in press). A deviance reduction algorithm, nCPA considers both the mean and the variance of response variables, contrary to most parametric techniques that focus only on the mean (Breiman and others 1984; Sokal and Rohlf 1995). Most parametric techniques require that data meet the assumptions of homogeneity of variances (e.g., ANOVA) or homoscedasticity (regression) despite the fact that changes in the variance may be equally informative as changes in the mean (e.g., Palmer and others 1997). For example, in ecological risk assessment, a fitted function that describes the dose-response relationship between a measurement endpoint and level of exposure to a contaminant is often used to estimate the magnitude of effect on the endpoint at a particular contaminant concentration (effective concentration, or EC) (Suter 1993). However, distributional properties of most metrics used in bioassessment are not conducive for these types of fitted models and, we argue, are not appropriate. In our study, many threshold responses were detected due to dramatic changes in variance in metric values with increasing levels of TP (e.g., Figure 2). This change in variance would have violated the assumptions of commonly employed parametric statistics but was paramount to the detection of levels of TP that resulted in a changepoint in our study.

While nCPA was effective in this study and has advantages over other many other methods for this application, one potential criticisms of nCPA is that it may not detect a low-level changepoint if a second, competing changepoint occurs at a higher concentration. First, we recommend that all data be examined graphically before any analysis is conducted so that the shape of the response can be evaluated (e.g., Karr and Chu 1997). If multiple changepoints are evident, a tree-based, recursive approach (i.e., tree regression; Breiman and others 1984) can be used to help isolate the lower changepoint. Here, the model splits the data into multiple subsets rather than just two. The subset of data above the upper changepoint can be discarded, and nCPA conducted on the lower subset of data. In this study, all primary changepoints occurred at low concentrations, although bootstrapping revealed that, in a few instances, a second, slightly weaker change also occurred at a higher concentration and subsequently skewed the upper range of the cumulative probabilities (e.g., Figure 3). Because nCPA is an extension of recursive-partitioning techniques such as tree regression, they are compatible and may provide a tactical, conservative means of detecting secondary changes at low concentrations if a primary response occurs at a greater concentration.

In our study, we defined macroinvertebrate structure and function as our assessment endpoint, and used a stressor-identification process to select five individual biological metrics and a multimetric index, the Nutrient-IBI, as measurement endpoints. We analyzed the individual metrics separately because we were concerned about the effect of blending metrics into one score on our threshold estimates. Of particular concern was that some metrics might have responded at different levels of TP, thus the IBI would have found the middle of this response range and possibly underestimated the risk posed by lower levels of TP. Conversely, we recognized that aggregating the individual metrics into the IBI might have increased the signal-to-noise ratio and allowed us to detect assemblage-levels changes that may have been clouded by variability at the individual-metric level. In reality, most of the individual attributes responded at a relatively similar levels of TP as the IBI, but the IBI overall had a tighter range of cumulative probabilities of a threshold response to TP than the individual metrics. However, there was modest deviation in the TP changepoints between the IBI and some metrics, suggesting that aggregating the responses into one index may have masked the variation in responses among individual attributes of the macroinvertebrate assemblage. Considering that biological responses to other stressors in other regions could lead to a wider range of changepoints than observed in this study, it is important to recognize this potential artifact of multimetric indexes. Moreover, the reduction in variance of individual metric values that invariably results from aggregating them into a multimetric index may actually eliminate biologically relevant changes in variance that could be detected using nCPA. Thus, we highly recommend the analysis of individual metrics in addition to an aggregated multimetric index to better characterize the range of levels of a candidate stressor that pose a risk to different facets of biological condition.

A final consideration when using nCPA is that it is a just a statistical tool, and any tool can be used inappropriately. We used nCPA for quantifying the cumulative probability that a particular level of TP resulted in a biologically significant change in macroinvertebrate structure and function, as expressed by the selected metrics (measurement endpoints). However, as with any statistical technique, nCPA may detect a statistical change in the data that may not represent a biologically significant change—clearly, the definition of biological significance is a subjective one and will vary among scientists and decision-makers. However, our results indicated that means and variances of assemblage attributes to the left and right of the $\geq$50% cumulative probabilities of changepoints differed markedly, sometimes by a factor $>$ 10. We contend that these changepoints represented TP levels that resulted in a qualitatively different biological community, as expressed by various attributes of assemblage structure and function, and were indicative of biologically significant changes.

Conclusions and Recommendations

Bioassessment and ecological risk assessment are inherently complementary in nature (Pittinger and others 2000). We presented a generalized approach for integrating these two assessment systems for the purpose of supporting numerical water-quality criteria. The strengths of the approach are the establishment of cause-effect linkages and the estimation of numerical thresholds. Moreover, the results are easy to interpret and communicate to environmental decision-makers and the public (Schiller and others 2001).

In this study, the weight-of-evidence produced from these analyses implied that a TP criterion $>$ 12–15 μg/L is likely to cause degradation of macroinvertebrate structure and function, a reflection of biological integrity, in at least this area of the south Florida coastal plain nutrient ecoregion. Our results also indicated that there is a very low (typically $\leq$5%) probability that an IBI threshold response would occur at $\leq$ 10 μg/L TP, while there is $\geq$95% certainty that a threshold would occur at $\leq$ 17 μg/L TP. However, this study only considers the macroinvertebrate component of biological integrity. The purpose of this study is not to imply that macroinvertebrate attributes should be the only endpoints used to assign a water-quality criterion to a region and water body. On the contrary, we highly recommend the evaluation of the responses of multiple biological endpoints from a variety of indicator groups across multiple trophic levels to better identify criteria protective of biological integrity. It is also important to recognize that the establishment of numerical criteria is ultimately a societal decision that will be based on a host of factors. However, these results do provide some compelling evidence that bioassessment can be used in a risk-assessment framework to identify critical levels of pollution, and ultimately guide environmental decision-making. Although the approach seems promising, it remains to be seen how well it will perform in different geographic regions and water bodies of the USA and other parts of the world.

Acknowledgments

We thank S. Qian for writing the S-Plus function for the nCPA method, J. Johnson, K. Nicholas, and L. Karppi for collecting water samples, and W. Willis, J. Rice, and P. Heine for conducting TP analyses. The critical reviews of N. Detenbeck, J. Karr, D. Lemly, S. Mozley, D. Urban, and three anonymous reviewers improved the manuscript. Primary funding was provided by a grant from the EAA Environmental Protection District to the Duke University Wetland Center. RSK was partially supported by a grant from the United States Environmental Protection Agency's Science to Achieve Results (STAR) Estuarine and Great Lakes (EaGLe) program through funding to the Atlantic Slope Consortium, USEPA Agreement #R-82868401. Although the research described in this article has been funded in part by the United States Environmental Protection Agency, it has not been subjected to the agency's required peer and policy review and therefore does not necessarily reflect the views of the agency and no official endorsement should be inferred.

Literature Cited

Adams, S. M., and M. S. Greeley. 2000. Ecotoxicological indicators of water quality: using multi-response indicators to assess the health of aquatic ecosystems. *Water Air and Soil Pollution* 123:103–115.

Angermeier, P. L., and J. R. Karr. 1994. Biological integrity versus biological diversity as policy directives—protecting biotic resources. *Bioscience* 44:690–697.

APHA (American Public Health Association). 1992. Standard methods for the evaluation of water and wastewater, 18 thedition. American Public Health Association, Washington, DC.

Barbour, M. T., J. B. Stribling, and J. R. Karr. 1995. Multimetric approach for establishing biocriteria and measuring biological condition. Pages 63–77 *in* W. S. Davis, and T. P. Simon (eds.)., Biological assessment and criteria: Tools for water resource planning and decision-making. CRC Press, Boca Raton, Florida.

Barbour, M. T., J. Gerritsen, G. E. Griffith, R. Frydenborg, E. McCarron, J. S. White, and M. L. Bastian. 1996. A framework for biological criteria for Florida streams using benthic macroinvertebrates. *Journal of the North American Benthological Society* 15:185–211.

Barbour M. T., J. Gerritsen., B. D. Snyder, and J. B. Stribling. 1999. Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates, and fish. EPA 841-0B-99-002. U. S. Environmental Protection Agency, Office of Water. Washington, DC.

Barbour, M. T., W. F. Swietlik, S. K. Jackson, D. L. Courtemanch, S. P. Davies, and C. O. Yoder. 2000. Measuring the attainment of biological integrity in the USA: A critical element of ecological integrity. *Hydrobiologia* 422:653–464.

Beyers, D. W. 1998. Causal inference in environmental impact studies. *Journal of the North American Benthological Society* 17:367–373.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. Classification and regression trees. Wadsworth and Brooks/Cole, Monterey, CA.

Cairns, J., and J. R. Pratt. 1993. A history of biological monitoring using benthic macroinvertebrates. Pages 10–27 *in* D. M. Rosenberg, and V. H. Resh (eds.)., Freshwater biomonitoring and benthic macroinvertebrates. Chapman and Hall, New York.

Clements, W. H., D. M. Carlisle, L. A. Courtney, and E. A. Harrahy. 2002. Integrating observational and experimental approaches to demonstrate causation in stream biomonitoring studies. *Environmental Toxicology and Chemistry* 21:1138–1146.

Cormier, S. M., G. W. Norton, G. W. Suter, D. Altfater, and B. Counts. 2002. Determining the causes of impairments in the Little Scioto River, Ohio, USA: Part 2. Characterization of stress. *Environmental Toxicology and Chemistry* 21:1125–1137.

Daehler, C. C., and D. R. Strong. 1996. Can you bottle nature? The roles of microcosms in ecological research. *Ecology* 77:663–664.

Davis, S. M., and J. C. Ogden. 1994. Everglades: The ecosystem and its restoration. St. Lucie Press, Boca Raton, FL.

DeBusk, W. F., K. R. Reddy, M. S. Koch, and Y. Wang. 1994. Spatial distribution of soil nutrients in a northern Everglades marsh: Water Conservation Area 2A. *Soil Science Society of America Journal* 58:543–552.

Dodds, W. K., and E. B. Welch. 2000. Establishing nutrient criteria in streams. *Journal of the North American Benthological Society* 19:186–196.

Efron, B., and R. J. Tibshirani. 1993. An introduction to the bootstrap. Chapman and Hall, London.

Faith, D. P., P. R. Minchin, and L. Belbin. 1987. Compositional dissimilarity as a robust measure of ecological distance. *Vegetation* 69:57–68.

FDEP (Florida Department of Environmental Protection)—1996. Standard operating procedures manual—Benthic macroinvertebrate sampling and habitat assessment methods: 1. Freshwater streams and rivers. Florida Department of Environmental Protection, Tallahassee, FL.

Fore, L. S., J. R. Karr, and R. W. Wisseman. 1996. Assessing invertebrate responses to human activities: evaluating alternative approaches. *Journal of the North American Benthological Society* 15:212–231.

Germano, J. D. 1999. Ecology, statistics, and the art of misdiagnosis: The need for a paradigm shift. *Environmental Reviews* 7:167–190.

Griffith, M. B., P. R. Kaufmann, A. T. Herlihy, and B. R. Hill. 2001. Analysis of macroinvertebrate assemblages in relation to environmental gradients in Rocky Mountain streams. *Ecological Applications* 11:489–505.

Hart, D. D., and C. T. Robinson. 1990. Resource limitation in a stream community: phosphorus enrichment effects on periphyton and grazers. *Ecology* 71:1494–1502.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54:187–211.

Johnson, D. H. 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management* 63:763–772.

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

Karr, J. R., and D. R. Dudley. 1981. Ecological perspectives on water-quality goals. *Environmental Management* 5:55–68.

Karr, J. R., and E. W. Chu. 1997. Biological monitoring and assessment: Using multimetric indexes effectively. EPA 235-R97-001. University of Washington, Seattle.E. W.

King, R. S. 2001. Dimensions of invertebrate assemblage organization across a phosphorus-limited everglades landscape. Ph.D. Dissertation, Duke University, Durham, NC..

King, R. S., and C. J. Richardson. 2002. Evaluating subsampling approaches and macroinvertebrate taxonomic resolution for wetland bioassessment. *Journal of the North American Benthological Society* 21:150–171.

King, R. S., and Richardson C. J. In press. Macroinvertebrate and fish responses to experimental P additions in Everglades sloughs. *In*C. J. Richardson (ed). The Everglades experiments: Lessons for ecosystem restoration. Springer-Verlag, New York .

Kleinbaum, D. G., L. L. Kupper, and E. E. Muller. 1988. Applied regression analysis and other multivariable methods. Duxbury Press, Belmont, CA.

Larsen, D. P., and A. T. Herlihy. 1998. The dilemma of sampling streams for macroinvertebrate richness. *Journal of the North American Benthological Society* 17:359–366.

Legendre, P., and L. Legendre. 1998. Numerical ecology, second edition. Elsevier, Amsterdam.

Lemly, A. D., and R. S. King. 2000. An insect-bacteria bioindicator for detecting detrimental nutrient enrichment in wetlands. *Wetlands* 20:91–100.

Lemly, A. D., and C. J. Richardson. 1997. Guidelines for risk assessment in wetlands. *Environmental Monitoring and Assessment* 47:117–134.

McCormick, P. V., P. S. Rawlik, K. Lurding, E. P. Smith, and F. H. Sklar. 1996. Periphyton-water quality relationships along a nutrient gradient in the northern Florida Everglades. *Journal of the North American Benthological Society* 15:433–449.

McCune, B., J. P. Dey, J. E. Peck, D. Cassell, K. Heiman, S. Will-Wolf, and P. N. Neitlich. 1997. Repeatability of community data: species richness versus gradient scores in large-scale lichen studies. *Bryologist* 100:40–46.

in R. W. Merritt, and K. W. Cummins (eds.). 1996. An introduction to the aquatic insects of North America, third edition. Kendall/Hall, Dubuque, IA.

Miltner, R. J., and E. T. Rankin. 1998. Primary nutrients and the biotic integrity of rivers and streams. *Freshwater Biology* 40:145–158.

Norton, S. B., S. M. Cormier, M. Smith, and R. C. Jones. 2000. Can biological assessment discriminate among types of stress? A case study for the eastern cornbelt plains ecoregion. *Environmental Toxicology and Chemistry* 19:1113–1119.

Palmer, M. A., C. C. Hakenkamp, and K. Nelsonbaker. 1997. Ecological heterogeneity in streams: why variance matters. *Journal of the North American Benthological Society* 16:189–202.

Perrin, C. J., and J. S. Richardson. 1997. N and P limitation of benthos abundance in the Nechako River, British Columbia. *Canadian Journal of Fisheries and Aquatic Sciences* 54:2574–2583.

Pielou, E. C. 1984. The interpretation of ecological data: A primer on classification and ordination. John Wiley and Sons, New York.

Pittinger, C., K. Thornton, S. B. Norton, and M. Barbour. 2000. The converging paths of ecological assessment and ecological risk assessment. *SETAC Globe, May 2000*: .

Qian, S. S., R. S. King, and C. J. Richardson. Two statistical methods for the detection of environmental thresholds. *Ecological Modelling* In press.

Richardson, C. J., G. M. Ferrell, and P. Vaithiyanathan. 1999. Nutrient effects on stand structure, resorption efficiency, and secondary compounds in Everglades sawgrass. *Ecology* 80:2182–2192.

Richardson, C. J., P. Vaithiyanathan, R. J. Stevenson, R. S. King, C. A. Stow, R. G. Qualls, and S. S. Qian. 2000. The ecological basis for a phosphorus (P) threshold in the Everglades: Directions for sustaining ecosystem structure and function. Duke Wetland Center Publication 00-02. Duke University, Durham, NC.

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

Schiller, A., C. T. Hunsaker, M. A. Kane, A. K. Wolfe, V. H. Dale, and G. W. Suter., et al2001. Communicating ecological indicators to decision makers and the public. *Conservation Ecology* 5:1–26.

SFWMD (South Florida Water Management District). 1992. Surface water improvement plan for the Everglades. Supporting information document. South Florida Water Management District, West Palm Beach, FL.

SFWMD (South Florida Water Management District). 2000. 2000 Everglades consolidated report. Supporting information document. South Florida Water Management District, West Palm Beach, FL.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry, third edition. W. H. Freeman and Co., New York.

Suter, G. W. 1993. Ecological risk assessment. Lewis Publishers, Chelsea, MI.

Suter, G. W. 1996. Abuse of hypothesis testing statistics in ecological risk assessment. *Human and Ecological Risk Assessment* 2:331–347.

Suter, G. W. 2001. Applicability of indicator monitoring to ecological risk assessment. *Ecological Indicators* 1:101–112.

USEPA (United States Environmental Protection Agency). 1997. Field and laboratory methods for macroinvertebrate and habitat assessment of low gradient, nontidal streams. Mid-Atlantic Coastal Streams Workgroup, Environmental Services Division, Region 3, Wheeling, WV.

USEPA (United States Environmental Protection Agency). 1998a. National strategy for the development of regional nutrient criteria. EPA 822-R-98-002. Office of Water, Washington, DC.

USEPA (United States Environmental Protection Agency). 1998b. Guidelines for ecological risk assessment. EPA/630/R-95/002F. Office of Research and Development. Risk Assessment Forum, Washington, DC.

USEPA (United States Environmental Protection Agency). 2000a. Ambient water quality recommendations: information supporting the development of state and tribal nutrient criteria for wetlands in nutrient ecoregion XIII. EPA 822-B-00-023. Office of Water, Washington, DC.

USEPA (United States Environmental Protection Agency). 2000b. Stressor identification guidance manual. EPA 822-B-00-025. Office of Research and Development, Washington, DC.

Vaithiyanathan, P., and C. J. Richardson. 1998. Biogeochemical characteristics of the Everglades sloughs. *Journal of Environmental Quality* 27:1439–1450.

Venables, W. N., and B. D. Ripley. 1994. Modern applied statistics with S-Plus. Springer, New York.