Freshwater Biology

Freshwater Biology (2010) **55,** 1066–1085 doi:10.1111/j.1365-2427.2009.02357.x

APPLIED ISSUES

Method of predicting reference condition biota affects the performance and interpretation of ecological indices

CHARLES P. HAWKINS*, YONG CAO* AND BRETT ROPER*,†

*Department of Watershed Sciences, Ecology Center, Western Center for Monitoring and Assessment of Freshwater Ecosystems, Utah State University, Logan, UT, U.S.A.

†Fish and Aquatic Ecology Unit, Forestry Sciences Lab, USDA Forest Service, Logan, UT, U.S.A.

SUMMARY

1. The statistical rigour and interpretability of ecological assessments is strongly affected by how well we predict the biological assemblages expected to occur in the absence of human-caused stress, i.e. the reference condition. In this study, we examined how the specific method used to predict the reference condition affected the performance of two commonly used types of ecological index: RIVPACS-based O/E indices and multimetric indices (MMIs).

2. These two types of index have generally relied on different approaches to predicting the reference condition. For MMIs, some type of regionalisation is typically used to describe the range of metric values among reference sites and hence the expected range at assessed sites. For O/E indices, continuous modelling is used to predict how the biota varies among sites both among and within regions. Because the prediction method differs for these two types of index, it has been impossible to judge if differences in index performance (accuracy, precision, responsiveness and sensitivity) are caused by differences in the way reference condition biota are predicted or by differences in what the indices measure.

3. We used a common data set of 94 reference sites and 255 managed sites and the same potential set of predictor variables to compare the performance of five different MMIs and three O/E indices that were derived from different prediction methods: null models, multiple linear regression (MLR), classification and regression trees, Random Forests (RF) and linear discriminant functions models (LDM). We then calculated values of these indices for samples collected from the managed catchments as well as samples collected from 13 reference sites that were progressively altered in known ways by a simulation programme.

4. Both the type of predictor used and the type of index affected overall index performance. Modelled indices generally had the greatest sensitivity in assessing managed sites as biologically different from reference. Index sensitivity was determined by both an aspect of index precision (10th percentile of reference condition values) and responsiveness. The O/E indices showed the best scope of response to known biological alteration. All three O/E indices decreased linearly in response to simulated alteration in both overall assemblage structure (Bray-Curtis dissimilarity) and taxa loss. The MMIs declined linearly

Correspondence: Charles P. Hawkins, Western Center for Monitoring and Assessment of Freshwater Ecosystems, Department of Watershed Sciences, Ecology Center, Utah State University, Logan, UT 84322-5210, U.S.A.

E-mail: chuck.hawkins@usu.edu

Present address: Yong Cao, Illinois Natural History Survey, University of Illinois at Urbana-Champaign, 1816 S Oak Street, Champaign, IL 61820, U.S.A.

from low to intermediate levels of assemblage alteration but were less responsive between intermediate and high levels of biological alteration.
5. Insights gained from simulations can aid in testing assumptions regarding index response to stress and help ensure that we select indices that are ecologically interpretable and most useful to resource managers.

Keywords: ecological assessments, environmental gradients, modelling, reference condition, simulation

Introduction

Well, blind man, have you seen the elephant? Tell me, what sort of thing is an elephant? (from the Buddhist or Jain parable of the Blind Men and the Elephant).

Many indices have been developed over the last several decades to assess the ecological condition of freshwater ecosystems (e.g. Johnson, Wiederholm & Rosenberg, 1993; Wright, Sutliffe & Furse, 2000; Ziglio, Flaim & Siligardi, 2006). These ecological indices are used to characterise the status of a selected taxonomic group or assemblage with respect to anthropogenic disturbance (e.g. Plafkin *et al*., 1989; Resh, Norris & Barbour, 1995; Karr & Chu, 1998). Indices based on macroinvertebrate assemblages are frequently used because these assemblages are sensitive to anthropogenic alteration of waterways (Cairns & Pratt, 1993) and are widely distributed and easy to sample (Rosenberg & Resh, 1993).

Of the several types of ecological indices that exist, two types are widely used: additive multimetric indices (MMIs) based on the summed standardised values of several different types of assemblage attributes, and O/E indices that describe the departure of taxonomic composition from that expected under reference conditions (e.g. Moss *et al*., 1987; Hawkins, 2006). In addition to differences in the way these indices measure ecological quality, users of these two types of indices typically control for effects of natural environmental condition on index values in different ways: classification for MMIs and continuous modelling for O/E indices. Because of this confounding, it is impossible to determine whether differences in index performance are associated with differences in the ecological attributes they measure or differences in how the reference condition is predicted.

Use of both MMIs and O/E indices depends on a reference condition approach to estimate the biota (or index values) that are expected to occur under minimal human-caused disturbance (Hughes, Larsen & Omernik, 1986; Reynoldson *et al*., 1997; Stoddard *et al*., 2006). For the O/E index, the taxa expected at a site are estimated from models (e.g. River InVertebrate Prediction and Classification System, RIVPACS) that predict how taxon-specific probabilities of detection vary along continuous environmental gradients (Moss *et al*., 1987; Hawkins *et al*., 2000b; Clarke, Wright & Furse, 2003). For MMIs, *a priori* classifications, such as ecoregion (Omernik, 1987, 1995) or stream order, are frequently used to partition natural variability (Barbour *et al*., 1999). However, *a priori* classifications are often ineffective in accounting for much of the natural variance in assemblage structure (e.g. Hawkins *et al*., 2000a; Heino & Mykra, 2006; Herlihy, Hughes & Sifneos, 2006). Modelling techniques akin to what RIVPACS models do, may therefore improve the performance of MMIs. In four recent studies (Baker *et al*., 2005; Pont *et al*., 2006, 2009; Cao *et al*., 2007), the residual values obtained after regressing raw metric values on natural environmental gradients were used as the expected metric values at a site. Although this approach generally appears promising, it has not yet been applied to macroinvertebrate MMIs, especially in the context of improving our ecological interpretation of different indices and our understanding of their comparability (Davies & Jackson, 2006). Moreover, no studies have examined how different modelling approaches affect index performance.

A significant challenge in interpreting and comparing the performance of ecological indices is that we generally do not know how well the indices we develop measure true ecological impairment (Cao & Hawkins, 2005). Simulation is probably the only practical way we can objectively evaluate how accurate an index is relative to known conditions, but it has seldom been used in evaluating the performance of ecological indices (Cao & Hawkins, 2005; Mazor *et al*., 2006).

Our main objective in this study was to evaluate how different prediction methods based on the same set of candidate predictor variables affected the performance of MMIs and O/E indices in detecting biological alteration associated with landscape and waterway modification. We also wanted to determine whether modelled indices assessed biological alteration differently than non-modelled (null) indices. Finally, we wanted to determine how well both non-modelled and modelled indices tracked known changes in assemblage structure.

Methods

Study area and sampling design

We used macroinvertebrate samples that were collected from 94 reference-quality and 300 managed sites that were located on public land within the Interior Columbia River Basin (ICRB), U.S.A. during the summers of 1998–2002 (Fig. 1). The ICRB is 58.4 million ha in size, of which approximately 30.9 million ha are managed by 35 different National Forests and 17 Bureau of Land Management districts (Quigley, Haynes & Graham, 1996). Within the ICRB, landscape alteration associated with land management ranges from minimal in designated wilderness areas to intensive vegetation manipulation and livestock grazing (Henderson *et al.*, 2005). Criteria for reaches to be considered in reference condition included: road densities <0.5 km km^{-2}, <5% of the catchment had been logged, no livestock grazing had occurred in the last 30 years, and no evidence of mining impacts was evident in the catchment. Managed catchments exceeded one or more of these criteria.

We selected sampling sites based on a spatially balanced probabilistic sampling design (Stevens & Olsen, 1999), which ensured that all levels of land management were equally likely to be sampled (Kershner *et al.*, 2004). The sample frame consisted of all 6th level (4000–15 000 hectare) hydrological units (HU). Within each HU, we sampled a low-gradient (<3%) reach that was near the catchment's outlet and on publicly managed land. We sampled

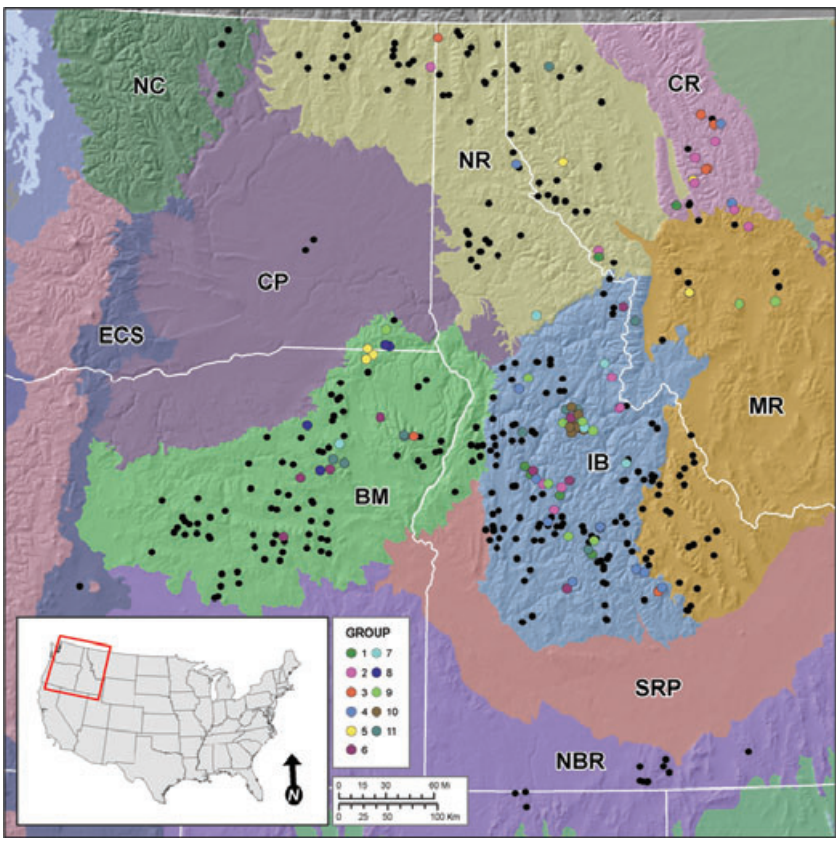

Fig. 1 Map of the study area showing the location of references sites classified by their biological similarity (coloured dots) and the managed sites (smaller, black dots) within the interior Columbia Basin. These reference site classes were used to develop both LDM- and RF-based O/E indices. Colour-shaded areas on the map represent major ecoregions within the basin: BM = Blue Mountains, CP = Columbia Plateau, CR = Canadian Rockies, ECS = Eastern Cascade Slopes and Foothills, IB = Idaho Batholith, MR = Middle Rockies, NBR = Northern Basin and Range, NC = North Cascades, NR = Northern Rockies and SRP = Snake River Plain.

low-gradient reaches because they may be more sensitive to catchment disturbances than other types of reaches (Montgomery & Macdonald, 2002) and thus might serve as sentinel reaches.

Collecting and processing macroinvertebrate samples

During summer (late June to early September), we collected invertebrate samples from four fastest-water habitat units (usually riffles) within each reach. In each habitat unit, we randomly selected two sampling points, at which we used a 0.09-m^2 Surber-type sampler (500 μm mesh size) to collect macroinvertebrates. We pooled these eight samples to give a total sampling area of 0.72 m^2. In the laboratory, approximately 500 individuals were randomly picked and identified to the lowest practical taxonomic level. Most taxa were generally identified with genus. However, chironomids were only identified to subfamily, and oligochaetes, water mites, and ostracods were left at these three course levels of aggregation. We then assigned all taxa to 188 unambiguous operational taxonomic units (OTUs) to provide consistent taxonomic treatment across all samples. The taxonomic level of the OTUs varied from species to phylum: 66% of OTUs were genera or species, 27% were families, and 7% were higher taxonomic levels. Those individuals that could not be placed into an unambiguous OTU were excluded from analyses. After assigning individuals to OTUs and dropping ambiguous taxa, we used a computer program to randomly select up to 300 individuals from each sample to ensure maximum sample comparability in terms of individual counts (samples with <300 individuals were retained).

Predictor variables

To construct models, we used a variety of candidate predictor variables that we extracted from Geographic Information System (GIS) coverages of topography, geology, soils and climate (Table 1). We have found that such map-level predictors usually perform as well as combined map-level and local-level measures of habitat conditions. For assessment purposes, we were mainly interested in exploring those variables that are reasonably invariant to landscape and water way alteration (i.e. they should characterise natural environmental gradients but not be affected by human activity) and which have been previously observed to be associated with variation in stream invertebrate assemblage composition (e.g. Hawkins *et al.*, 2000b; Ostermiller & Hawkins, 2004). Because some predictor variables were missing for some sites, the specific number of managed sites used in index calibration varied between 279 and 300. For model evaluation, we used the 255 managed sites for which all indices were calculated.

Multimetric index development

Of the many metrics that have been proposed for use in MMIs (e.g. Barbour *et al.*, 1999), we considered only those 37 metrics that had been previously used either within or near our study region (Appendix S1). In evaluating these candidate metrics for inclusion in an MMI, we first used the different modelling methods to assess their association with natural environmental gradients and, where appropriate, predict expected metric values under different natural environmental settings. We then evaluated both the original, raw

Table 1 Means and extremes of the 14 map-derived environmental factors used to evaluate effects of natural gradients on indicator values

Site type	Statistic	Lat	Long	C-Area	Alt	Grad	DOY	XAAT	XMxAT	XMiAT	TR	XAP	FFD	WD	XARH
Reference	Mean	45.73	−115.37	1345	1637	1.1	212	3.3	10.3	−3.7	13.9	1007	54	130	63
	Max	48.87	−112.49	5924	2353	3.4	253	7.5	14.1	1.8	18.1	1810	117	182	70
	Min	43.81	−118.87	155	707	0.1	171	0.3	6.3	−7.5	8.3	520	9	91	52
Managed	Mean	44.61	−114.94	832	1990	1.1	199	5.1	11.8	−1.7	13.6	721	76	108	62
	Max	45.50	−113.64	1975	2420	1.8	248	9.9	17.5	3.6	18.1	1724	142	178	72
	Min	43.88	−116.34	182	1237	0.3	152	−0.1	4.9	−7.3	6.7	224	9	58	52

Lat = latitude, Long = longitude, C-Area = catchment area (km^2), Alt = altitude (m.a.s.l.), Grad = reach gradient (%), DOY = day of year sample was taken, XAAT = mean annual air temperature (°C), XMxAT = mean maximum air temperature (°C), XMiAT = mean minimum air temperature (°C), TR = mean annual temperature range (°C), XAP = mean annual precipitation (mm), FFD = number of freeze-free days, WD = number of wet days, XARH = mean annual relative humidity (%).

metrics and the model-adjusted metrics for their ability to discriminate between reference and managed sites and for redundancy in ecological information. After selection of the set of metrics for use in each MMI, we then standardised their values (0–100), calculated their summed values and then rescaled the summed MMI values to a standard scale (0–100).

Partitioning natural variability in MMIs. We developed five different MMIs, all of which were derived from the same invertebrate sample data but which differed in the method used to control for natural biotic variation. We developed a single unadjusted MMI (MMI-A) for the entire ICRB for comparison with the other modelled indices. We did not develop separate MMIs for different landscape classes (i.e. ecoregions), because the vast majority of sites occurred within mountainous regions of similar physiography. For the four other MMIs, we used each of three types of statistical model to account for the effects of natural environmental gradients on index values. We used multiple linear regression (MLR) to derive two MMIs, classification and regression trees (CART) for another and Random Forests (RF) models for the final MMI. One of the MLR-based MMIs was derived by modelling MMI-A values (not individual component metrics). All other models were developed for individual metrics prior to aggregation into an MMI. The MMI models predict the values of metrics (or of MMI directly) expected at a site under reference condition. The size of the difference between observed and expected metric values (the residual) is a measure of the degree of biological alteration. Predictive MMI's are thus conceptually similar to O/E models and differ only in the specific biological index of assemblage condition and how departure from the reference expectation is measured (difference vs. ratio).

Summary of modelling approaches. The modelling approaches we explored differ in the way they select predictor variables, the assumptions they make regarding the type of relationships that exist between biological attributes (e.g. metrics) and environmental features and how predictions are made.

Multiple linear regression (MLR) is a parametric statistical model that assumes that responses are linear (e.g. Draper & Smith, 1981). Stepwise procedures have been commonly used to select the 'best' predictor variables from a set of candidate predictors, but they have been criticised for inconsistent results associated with the specific selection algorithm used, bias in parameter estimation and inappropriate focus on a single model when many others may perform equally well (e.g. Whittingham *et al.*, 2006). We therefore used the leaps() package (Lumley, 2004; version 2.9) for R (version 2.9.2, R Development Core Team 2009) to evaluate a large number of candidate MLR models. There are 2^K-1 possible models where k = number of candidate variables, which makes it difficult to evaluate a large number of candidate predictor variables. The leaps() package uses branch-bound algorithms to efficiently select any specified number of optimal models (Miller, 2002). We used this routine to select the best (based on AIC) 10 models of each model size (i.e. n = 1, 2, 3… predictors). From this list, we chose a single model among those with the lowest AIC values. We did not attempt to model interactions among predictors.

CART models are based on machine learning algorithms and represent a nonparametric, nonlinear approach to prediction or classification (Breiman *et al.*, 1984). CART models identify a series of logical if–then conditions in which sets of observations on a response variable are progressively parsed into bins based on their associations with predictor variables. The CART process selects the best single predictor variable for each split from a potentially large number of predictors to form decision trees that best partition the samplewide variance in the response variable among bins of samples. CART models make no assumptions about the distributional properties of the data and can capture highly nonlinear relationships. In our analyses, we used the tree() routine in the R statistical package with 10-fold cross-validation to identify the optimal number of terminal nodes. Once models were identified, we applied them to both reference and managed data sets and used model residuals as metric values.

RF models incorporate the combined output of many decision tree models such that the RF model predictions are an average of the predictions from those many trees. RF models are built by repeatedly applying a randomly selected subset of the predictor variables to a randomly selected subset of the samples. This bootstrapping-type process is repeated many times to generate a forest of decision trees. The theoretical advantages of RF include those associated with CART as well as that RF models cannot be overfit, and they produce unbiased estimates of modelling error (Breiman, 2001). We used the R implementation

of RF (package randomForest, version 4.5–30, Liaw & Wiener, 2002) and used 3–10 randomly selected variables (depending on metrics) and randomly selected 64% of samples to build models. The final RF model was based on 5000 individual decision trees. As in CART, the final MMI was based on the residual metric values that were derived by calculating the difference between predicted and observed metric values.

For all models, we set an arbitrary threshold that a model had to explain at least 10% of the variation in metric values before we adjusted metric values. We applied this threshold to limit the number of potential predictors to a manageable number and minimise inclusion of predictors that had trivial effects on predicted metric values. We subsequently checked whether exclusion of any variables as a consequence of this procedure resulted in biased predictions (see section on *Evaluating and comparing index performance/Bias* below).

Metric evaluation. We used t-tests to evaluate how well each of the metrics (adjusted metrics as well as those that were unrelated to natural gradients) discriminated between samples from reference and managed sites. Those metrics that failed ($P > 0.1$) this test or responded in a direction contrary to general expectations were not considered further.

Metric redundancy. To identify metrics that were potentially redundant in the information they provided, we calculated Pearson's correlation coefficients for all pairs of metrics. We then selected one of any set of strongly ($r > 0.7$) correlated metrics for use in the MMI. In such cases, we chose the metric that best discriminated between reference and test sites. This process resulted in the selection of 8 - 13 metrics for use in an MMI, with the specific number of metrics selected depending on the modelling approach used.

Re-scaling metrics. We rescaled all selected metrics to have a range of 0–100 (Eq. 1), where higher values indicate less impairment:

$$\text{Standardized Metric Value} = \left(\frac{\text{Site Value} - \text{Min}}{\text{Max} - \text{Min}}\right) \times 100$$

where, *Site Value* = the observed or adjusted value for each metric at a site, *Min* = the minimum site value and *Max* = the maximum site value (see Blocksom, 2003). If the value of a metric increased at managed sites, we subtracted the value produced by equation 1 from 100. This rescaling is commonly used in ecology (Legendre & Legendre, 1998) and is identical to that used by Stoddard *et al*. (2008), except they used the 95th and 5th percentiles for Max and Min, respectively (see also United States Environmental Protection Agency, 2006). The use of percentiles rather than the observed range in values was intended to eliminate the effects of outliers. However, because the definitions of outliers and thresholds are arbitrary, we thought it best to use the full range of data (see Blocksom, 2003; for review of other methods). We then averaged the values of all standardised metrics at a site to give the final MMI value. Finally, we standardised MMI values by dividing original MMI values by the mean value observed at reference sites to allow direct comparison of the precision of MMIs with the O/E indices (Hawkins, 2006), which theoretically have means of one if samples were collected from reference-quality sites.

Development of RIVPACS-type models

We generally followed previously described procedures for developing RIVPACS-type models in the U.S.A. and calculating O/E values (Hawkins *et al*., 2000b; Van Sickle *et al*., 2005). However, we used two modelling approaches to predict the probabilities of class membership for individual sites. Traditional RIVPACS models use linear discriminant functions (LDF) models to predict probabilities of group membership. We also used RF models as an alternative method of predicting group membership. For the LDF modelling, we first used the all-possible-subsets routine of Van Sickle, Huff & Hawkins (2006) to identify the best LDMs from 65 535 possible models (all possible combinations of 17 predictor variables). We defined best as models that were simultaneously parsimonious (fewest predictors), accurate and precise (lowest root mean square error of O/E). For the RF modelling, we used the same general modelling strategy as for metric development (four randomly selected independent variables for each tree, a random selection of 64% of samples for each tree and 5000 trees). We calculated O/E for $P \geq 0.5$ (hereafter O/E_{50}) because O/E based on lower probability thresholds (e.g. >0) generally do not perform as well as O/E_{50} (Hawkins *et al*., 2000b; Ostermiller & Hawkins, 2004; Hawkins, 2006; Van Sickle, Larsen & Hawkins, 2007), but see Clarke & Murphy (2006).

Simulating ecological impairment for index evaluation

We combined multiple samples (N = 3–7) collected at each of 13 reference sites to produce large collections of individuals that we could first alter in known ways and then subsample (300 individuals) to assess index behaviour. These combined samples contained between 1061 and 2308 OTU-level individuals and between 30 and 57 OTU taxa.

To evaluate how well the five MMIs and the three O/E indices quantified known ecological impairment, we followed Cao & Hawkins (2005) in simulating the ecological impairment caused by anthropogenic stress. These simulations were conceptually similar to a series of bioassay experiments in which replicate sets of assemblages were exposed to a series of increasing concentrations of a pollutant. We simulated the taxon-specific effects of stress as:

$$Y_i = X_i[1 - C(1 - TV_i)] \quad (2)$$

where, X_i = the initial number of individuals of taxon i in the initial unaltered assemblage, TV_i = the tolerance value of taxon i, C = a coefficient that controls the level of stress and Y_i = the number of individuals in taxon i after a stress occurs. We created 9 levels of stress in which C ranged from 0 to 3.2 at intervals of 0.4.

We used Hilsenhoff's (1987) tolerance values (HTVs = 0–10) to approximate true differences among taxa in their general sensitivity to stress. These values provide relatively coarse representations of true taxon sensitivities given that they were not developed for western species, but they did provide a way to create generally realistic, assemblage-wide changes in OTU population sizes that occur in response to stress. Under realistic conditions of the type of stress common in these managed catchments (logging, roads, livestock grazing), taxon abundances should decrease or increase depending on a taxon's response to nutrient and habitat (mainly fine sediment and temperature) alterations associated with these activities. To mimic this general behaviour, we chose a TV of 6.5 as the value below which abundances declined and then re-scaled all original TVs as $TV_i = HTV_i/6.5$. If $TV_i > 1$, stress will increase the abundance of taxon i; if $TV_i = 1$, the abundance of the taxon will not be affected; and if $TV_i < 1$, abundance will decrease. When the calculated number of individuals in a taxon dropped below one (fractions were possible), we considered the taxon to be locally extinct.

To incorporate the effects of subsampling error on our simulation results, we randomly drew ten 300-count replicate samples from each site at each stress level. This step created 1170 samples (13 assemblages × 9 stress levels × 10 replicates). We then used the mean values across the 10 replicates at each stress level to describe general trends. Because these 10 replicates were drawn with replacement, they are not truly independent, and we recognise that they will under-estimate true subsampling error. However, we use these mean values only to compare the relative behaviour of the different indices.

We quantified changes in assemblage structure with increasing simulated stress in two ways. First, we measured the percentage of taxa lost from the initial assemblage (large composite prior to 300-count resampling). The maximum % taxa loss in an assemblage depended on the proportion of taxa with $TV_i < 1$, because given the definition of TV_i used here, only those taxa with $TV_i < 1$ can eventually be eliminated. Per cent taxa-loss stabilised at C = 3.2 for all modelled assemblages (i.e. all intolerant taxa were extinct, and the remaining taxa were all increasing), but the range of taxa loss among sites varied between 56 and 80% of OTUs. Second, we measured the overall change in assemblage structure (taxon composition and relative abundance) as the Bray-Curtis (BC) dissimilarity in $\log_{10}$ taxon abundances between the original reference site assemblage and the stressed one. To visually assess how stress affected the 13 different assemblages, we conducted a non-metric multidimensional scaling (NMDS) ordination based on the $\log_{10}$ abundances of each taxon in each collection (i.e. the total number of individuals without fixed-count re-sampling).

Because of its design, O/E might be expected to track true taxa loss better than MMIs. However, changes in BC distances based on log densities should serve as a general measure of overall biological integrity, because BC distance simultaneously incorporates aspects of taxa composition, taxa richness and relative abundance. These three biological attributes describe fundamental aspects of biological integrity as defined by the biological assessment community within the U.S.A. (e.g. see the biological attributes used to describe how assemblages respond to anthropogenic disturbance in Davies & Jackson, 2006).

Evaluating and comparing index performance

The performance of an index can be evaluated based on precision, accuracy/bias, responsiveness and sensitivity. We use these terms as follows.

Precision. After adjusting for the effects of natural gradients, all reference sites should have the same index value other than variation caused by sampling error, i.e. one for both O/E and the MMIs standardised by their reference site means). In this context, the observed variability of index values among reference sites is a measure of how precisely we can estimate this value. We measured precision as the standard deviation (SD) of reference-site index values. To assess whether there is a possibility of precision differing among indices, we also calculated 95 per cent confidence intervals for all precision estimates. We also estimated the 10th percentile values (PER10) for the reference site samples, a threshold value related to precision that we used to infer if assessed samples were in reference condition or not (see Sensitivity below).

Bias. We evaluated bias in two ways. First, we used Random Forest modelling to determine whether remaining variation in reference site index values was systematically associated with naturally occurring conditions. Such biases could occur if specific types of sites were consistently under-predicted relative to other types of sites. Second, we examined how each of the indices tracked known simulated impairment.

Responsiveness. We measured responsiveness in three ways: as the difference in mean index values observed at reference and managed sites, as the Student's *t* value estimated from the comparison (*t*-test with unequal variances) of reference and managed site index values, and as the magnitude of response indices exhibited to simulated stress. We estimated 95 per cent confidence intervals for all managed site index means. The Student's *t* value provides an assessment of effect size (i.e. mean difference between reference and managed sites) adjusted for differences in variance in the index values observed at reference and managed sites.

Sensitivity. We measured sensitivity as the per cent of managed sites that were inferred to be in non-reference condition (NRC). We defined NRCs as index values that fell below the 10th percentile of reference site index values (PER10). The tenth percentile is an arbitrary value that we use for comparison purposes only. We then used McNemar's test (mcnemar.test() in R) to determine which indices were significantly different in sensitivity from one another. We tested all 28 possible comparisons and used the ''false discovery rate'' method (Waite & Cambell, 2006) to adjust P values for multiple comparisons. To determine how sensitivity varied with index precision and responsiveness, we regressed S on PER10 and responsiveness as measured by the Student's *t* statistic.

Results

Environmental setting of reference and managed sites

Both reference and managed sites occurred across a range of geologic and ecoregion settings (Fig. 1, Table 2). However, reference sites were most frequent in granitic catchments, whereas managed sites were more evenly distributed among granitic, metamorphic and volcanic catchments. With respect to ecoregion setting, >50% of reference sites were located in the Idaho Batholith ecoregion, whereas managed sites were more evenly distributed across catchments in the Blue Mountains, Northern Rockies, Idaho Batholith and Middle Rockies ecoregions.

Table 2 Per cent of sites within four different geology classes and nine different level-three ecoregions for each set of sample types

	Geology				Level three ecoregion								
Site type	G	M	S	V	10	11	12	15	16	17	41	77	80
Reference (94)	59	20	2	18	0	19	0	9	54	6	12	0	0
Managed (300)	33	21	6	36	1	27	1	22	33	11	1	1	4

G = granite, M = metamorphic, S = sedimentary, V = volcanic, 10 = Columbia Plateau, 11 = Blue Mountains, 12 = Snake River Plain, 15 = Northern Rockies, 16 = Idaho Batholith, 17 = Middle Rockies, 41 = Canadian Rockies, 77 = North Cascades, 80 = Northern Basin and Range.

Stream taxonomic richness and composition

Sample OTU richness varied approximately three-fold among reference sites (12–37) and approximately seven-fold across managed sites (6–40), showing that the sampled sites represented a wide range of natural and managed ecological conditions within the ICRB. Taxonomic composition also differed substantially among sites (e.g. mean BC dissimilarity between the reference site classes used in O/E modelling = 0.5).

Index development

MMI-A. Mean values of 21 of the 37 candidate metrics differed between reference and managed sites, but three of them (Trichoptera taxa richness, relative abundance (RA) of Trichoptera and number of long-lived taxa) had higher values at managed sites, contrary to expectations. We excluded these three metrics as well as five others that were redundant with more discriminatory metrics, which left 13 metrics for use in MMI-A (Table 3).

MMI-B. MLR accounted for about 15% of the variance in MMI-A values among reference sites and used five environmental variables to adjust for naturally occurring differences in MMI-A values (Table 3).

MMI-C. MLR showed that many individual metrics were associated with natural gradients. Sixteen adjusted and four raw metrics discriminated reference site samples from managed site samples in the expected direction. Four of these 16 metrics were eliminated for redundancy leaving eight adjusted metrics and four unadjusted metric as components of MMI-C (Table 3).

MMI-D. The application of CART to the metrics resulted in selection of eight metrics for use in MMI-D (Table 3). Four of these metrics were significantly associated with natural gradients and required adjustment by CART models.

MMI-E. Use of RF to adjust for natural gradients resulted in selection of 10 metrics for use in MMI-E (Table 3). Three of these metrics were significantly associated with natural environmental gradients and required adjustment by RF models.

O/E indices. Eleven groups were selected from the cluster analysis of 94 reference sites (Fig. 1). The number of sites per group ranged from 4 to 20. The all-possible subsets evaluation led to selection of a LDM with seven predictors: $\log_{10}$ catchment area (km^2), square root of channel slope (%), mean average monthly air temperature (°C), mean between-months temperature range (°C), number of freeze-free days, day of year and volcanic geology (1 or 0). The RF model used all predictor variables, but their relative importance (how frequently each predictor is selected in individual trees) varied considerably. The 10 most frequently used predictors in rank order were as follows: channel slope, catchment area, longitude, number of freeze-free days, latitude, day of year, number of wet days, altitude, mean minimum monthly air temperature and mean maximum monthly air temperature.

Index performance: field data

Precision. Estimates of index precision varied from SD = 0.11 to 0.17, but the 95% confidence intervals for index standard deviations were non-overlapping for only the OE-Null versus all other indices (Table 4). Both modelled O/E indices approached or achieved the precision associated with estimated sampling error (i.e. SD = 0.11). Precision was strongly correlated with the 10th percentile (PER10) of reference site values ($r = -0.88$), the type of statistic that is typically used for drawing inferences regarding the biological status of a site. Indices differed in PER10 values by as much as 0.18 standardised index units.

Accuracy / Bias. Site-specific systematic bias was an issue for only the null indices in which up to 10% of the variance in reference site values was associated with natural environmental features.

Responsiveness. Mean index values estimated at managed sites differed by 0.12 standardised units among indices (Table 4). Inspection of the 95 per cent confidence intervals for mean index values at these sites implied that there were real differences in responsiveness between some but not all indices. The MMI-D, OE-LDM and OE-Null indices appeared to assessed managed sites as most biologically altered and the MMI-B as the least altered. Estimates of the Student's t statistic also implied indices differed in their responsiveness to management (t = 4.5–11.2) and were correlated with mean index values ($r = 0.76$).

Sensitivity. The per cent of managed sites assessed as in NRC based on the PER10 threshold criterion varied markedly (32 to 55) among indices (Table 4). The McNemar tests showed that many indices differed from one another in sensitivity (Table 5). The

Table 3 Metrics used in each multimetric index and the natural environmental factors associated with variation among reference sites in metric values. Metrics were selected based on their ability to discriminate between reference and test sites after adjusting for the effects of natural environmental factors. R^2 values show the amount of variation in the selected metrics that was associated with the predictor variables

MMI	Model	Metrics or index	Predictor variables	R^2
A	None	RA highly sensitive taxa		
		RA collector-filterers		
		RA shredders		
		RA predators		
		RA two most dominant taxa		
		RA Ephemeroptera		
		RA Elmidae		
		Tolerant taxa richness		
		Hilsenhoff's index		
		Predator taxa richness		
		Highly sensitive taxa richness		
		Ephemeroptera taxa richness		
		Plecoptera taxa richness		
B	MLR	MMI-A	log C-Area (+), granitic (+), metamorphic (+), DOY (+), FFD (−)	0.15
C	MLR	RA Ephemeroptera	log C-Area (+), Grad (−), granitic (+), metamorphic (+), XAP (+), WD (+)	0.27
		RA Plecoptera	log C-Area (−), ER 11 (−), ER 41 (−), metamorphic (+), XAP (+), WD (+)	0.26
		RA collector-filterers	Lat (−), Long (+), metamorphic (−), XMiAT (+), XAP (−), WD (+)	0.26
		Scraper taxa richness	log C-Area (+), Alt (−), Grad (−), ER 11 (−), XAP (+), WD (−)	0.25
		Ephemeroptera taxa richness	Long (−), log C-Area (+), ER 11 (−), Metamorphic (+), volcanic (−), XMxAT (−), WD (−)	0.23
		Collector-filterer taxa richness	Metamorphic (−), DOY (−), XMiAT (−), XAP (−), WD (+)	0.20
		RA non-insects	Long (+), log C-Area (−), Grad (+), Alt (−), ER 11 (+), ER 16 (+)	0.19
		Predator taxa richness	log C-Area (+), ER 15 (+), metamorphic (−), volcanic (−)	0.13
		RA shredders	None	
		RA predators	None	
		RA highly sensitive taxa	None	
		RA top 3 dominant taxa	None	
D	CART	Highly sensitive taxa richness	C-Area, FFD, DOY	0.49
		RA predators	XAAT, C-Area	0.33
		Plecoptera taxa richness	WD, XMiAT, Lat	0.32
		Predator taxa richness	C-Area, DOY, granitic	0.27
		RA Plectoptera	None	
		RA highly sensitive taxa	None	
		RA Diptera	None	
		RA collector-filterers	None	
E	RF	RA highly sensitive taxa	C-Area, Alt, Lat, granitic, DOY	0.16
		Highly sensitive taxa richness	C-Area, DOY, Long, XAP, XMxAT	0.13
		RA Ephemeroptera	Lat, C-Area, Long, WD, FFD	0.12
		RA predators	None	
		RA collectors-filterers	None	
		RA Diptera	None	
		RA Plecoptera	None	
		Predator taxa richness	None	
		Collector-filterer taxa richness	None	
		Ephemeroptera taxa richness	None	

TV = Hilsenhoff's tolerance value. Highly sensitive taxa (TV = 0–2). Tolerant taxa (TV = 8–10). RA = relative abundance. Plus (+) and minus (−) symbols indicate direction of response of the metric to increasing values of the predictor variables. Abbreviations for predictor variables as in Table 1 except ecoregion = ER. Bedrock type is spelt out.

Table 4 Comparison of index performance in terms of: precision (SD = standard deviation of reference site values), the 10th percentile of reference site values (PER10), systematic bias (SB = % of variation among reference site values associated with natural environmental features after modelling), responsiveness (expressed as mean value for samples from managed sites as well as the Student's *t* value for the comparison between reference and managed site index values) and sensitivity (S = per cent of sites rated as in non-reference condition). Indices are ranked in terms of their sensitivity in assessing managed sites as different from reference. Lower and upper 95% confidence limits are included for estimates of index precision (CI of SD) and responsiveness (CI of mean). Indices with the same letter in the CI overlaps columns have CIs that overlap

		Reference site samples							Managed site samples						
Index	Model type	Mean	SD	L95CI	U95CI	CI Overlap	PER10	SB	Mean	SD	L95CI	U95CI	CI Overlap	Student's *t*	S
OE-RF	RF	1.05	0.11	0.09	0.13	A	0.94	0	0.88	0.22	0.85	0.90	BC	9.93	0.55
MMI-D	CART	1.00	0.14	0.12	0.16	A	0.80	0	0.80	0.19	0.77	0.82	A	11.23	0.49
OE-LDM	LDM	1.01	0.13	0.11	0.15	A	0.85	0	0.83	0.21	0.80	0.85	AB	9.90	0.48
MMI-C	MLR	1.00	0.11	0.09	0.13	A	0.88	0	0.90	0.15	0.88	0.91	BC	6.97	0.47
MMI-E	RF	1.00	0.12	0.10	0.14	A	0.84	0	0.88	0.17	0.86	0.90	BC	7.17	0.43
MMI-A	Null	1.00	0.14	0.12	0.16	A	0.79	3.2	0.87	0.18	0.85	0.89	B	6.92	0.36
OE-Null	Null	1.00	0.17	0.14	0.20	B	0.76	10	0.83	0.23	0.81	0.86	AB	7.37	0.33
MMI-B	MLR	1.00	0.13	0.11	0.15	A	0.82	0	0.92	0.19	0.90	0.94	C	4.46	0.32

RF-based O/E index was generally most sensitive (55% of sites) followed by the CART-based MMI, LDM-based O/E, MLR-based (metrics) MMI and RF-based MMI and null O/E index (43–49% of sites). The null MMI, null O/E and MMI-B (i.e. the MLR-adjusted MMI-A) assessed the fewest sites as in NRC (32-36%).

Both PER10 and index responsiveness (as measured by Student's *t*) influenced the sensitivity of these indices (Fig. 2, $S = -0.52 + 0.90 \cdot \text{PER10} + 0.026 \cdot t$, adjusted $R^2 = 0.93$, $P < 0.001$ for both PER10 and *t*). These two attributes varied independently from one another ($r = 0.21$), and both attributes of index performance contributed approximately equally to index sensitivity (standardised regression coefficients = 0.66 for Student's *t* and 0.59 for PER10).

Index Comparability

The strength of pairwise correlations between metric values at managed sites varied markedly ($0.63 \le r \le 0.92$). The correlations within index types were less variable ($0.80 \le r \le 0.92$ for the five MMIs and $0.80 \le r \le 0.89$ for the three O/E indices) implying that these two types of indices assess somewhat different aspects of ecological alteration.

Index performance: simulated data

Both measures of ecological impairment (% true taxa loss and BC dissimilarity) increased with simulated stress, but responses were site specific (Fig. 3). At the highest stress levels, simulated taxa loss varied between 56 and 80% among sites, and BC distance from reference varied between 0.45 and 0.65. Note that the sites became increasingly dissimilar in terms of both per cent of taxa loss and magnitude of BC dissimilarity to the unaltered sample as stress increased. The NMDS ordination of the simulated data showed that the 13 different reference sites differed substantially in assemblage structure prior

Table 5 Results of pairwise McNemar tests for differences in index sensitivities. *P*-values are adjusted for false discovery rates

Index	MMI-D	OE-LDM	MMI-C	MMI-E	MMI-A	OE-Null	MMI-B
OE-RF	0.139	0.021	0.025	0.000	0.000	0.000	0.000
MMI-D		0.729	0.449	0.025	0.000	0.016	0.000
OE-LDM			0.807	0.181	0.000	0.000	0.000
MMI-C				0.234	0.000	0.158	0.000
MMI-E					0.001	0.693	0.000
MMI-A						0.054	0.000
OE-Null							0.133

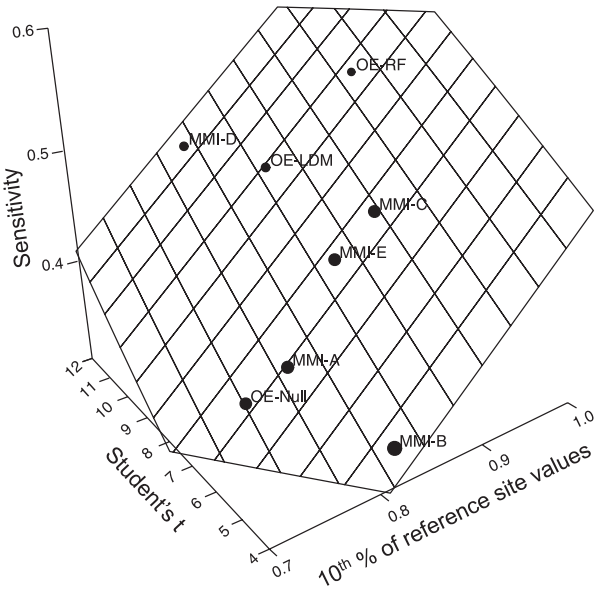

Fig. 2 Relationship between index sensitivity and index precision (as measured by the 10th percentile of reference site values, PER10) and index responsiveness (as measured by the Student's *t* statistic for the difference in mean index values between reference and managed sites).

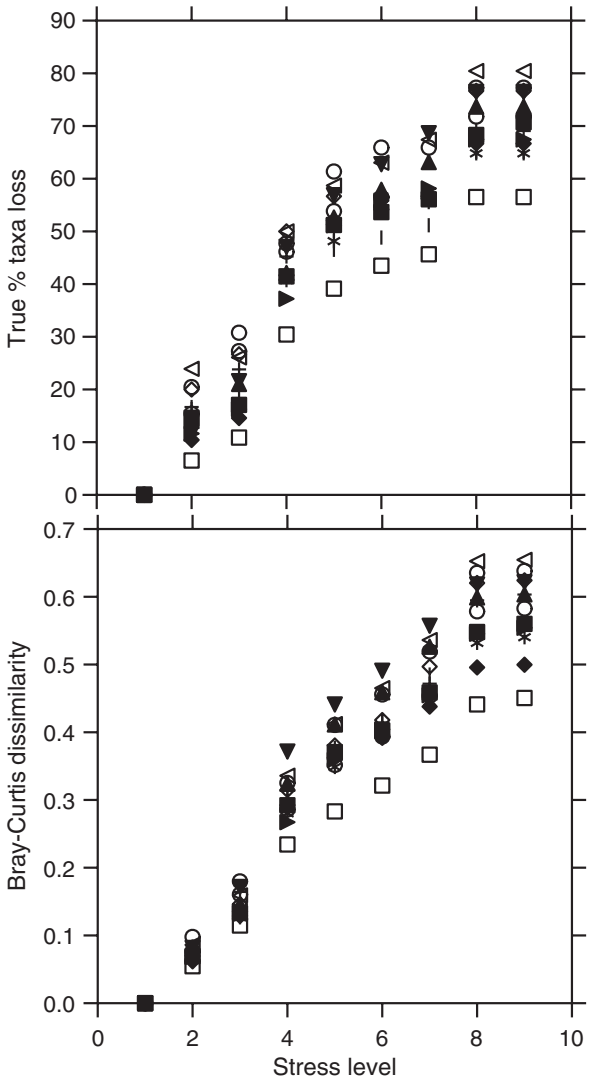

Fig. 3 Results of simulated impairment of 11 different reference sites showing true per cent loss of taxa (top panel) and Bray-Curtis dissimilarity from the original reference condition (bottom panel) in relation to increasing simulated stress (stress level 1 = no stress). Different symbols represent different streams. Note the increasing divergence in both taxa loss and BC dissimilarity with increasing stress.

to stress and that the structure of these assemblages changed markedly and generally in the same direction with increasing stress (Fig. 4). However, the stress-related trajectories for each site were also different enough that these assemblages became increasingly different from one another with increasing stress (cf. Fig. 3). Furthermore, partial alteration of the assemblages at some sites resulted in these samples resembling the unstressed samples of other sites more than their own unstressed samples. Our direct calculations of BC distances for all 78 possible pairs of 13 sites confirmed that assemblages became both increasingly different with increasing stress on average (BC dissimilarity increased from 0.57 to 0.64) and that the variance in among-site dissimilarities increased (variance increased from 0.0037 to 0.0078).

Index response to simulated alteration generally fell into two categories: indices with nearly linear responses across the entire impairment gradient and indices that responded linearly up to a threshold and then showed less marked response to further impairment (Fig. 5). All three of the O/E indices showed a generally linear response to simulated assemblage alteration, whereas all of the MMIs exhibited the latter behaviour. Trend lines for all MMIs were very similar to one another (Fig. 5). The standardised MMI values did not directly track either true taxa loss or the BC measure of assemblage alteration. In all cases, 80% taxa loss or a BC value of 0.6 dissimilarity units resulted in MMI values of *c.* 0.6 (i.e. 0.4 units of alteration). The O/E indices exhibited a more linear response across the entire range of simulated alteration as measured by both taxa loss and BC

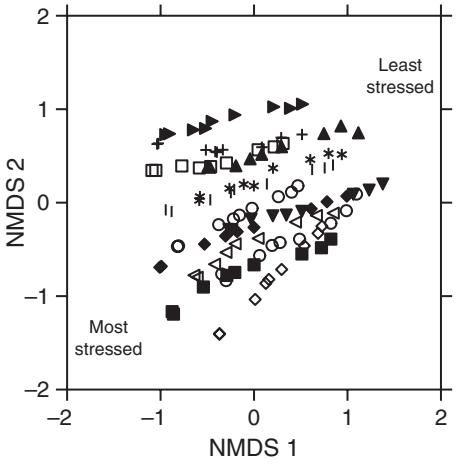

Fig. 4 Non-metric multidimensional scaling (NMDS) ordination showing the trajectories of change in assemblage structure ($\log_{10}$ abundances) with increasing stress. Different symbols represent different streams, but to aid in visualising differences among streams symbols are not the same as in Fig. 3. Note the progressive divergence in assemblage structure with increasing stress.

dissimilarity. However, the O/E indices were projected to imply total loss of predicted taxa at approximately 80–90% true taxa loss and approximately 0.7–0.8 BC dissimilarity.

Discussion

Relevancy and reliability of ecological indices

Considerable effort has been spent developing ecological indices, especially those designed to assess the status of freshwater biota. However, much less effort has been directed towards assessing whether these indices perform as expected or in generally interpretable ways (Yuan & Norton, 2003). Two general criteria for judging index performance are (i) relevancy to management- or policy-defined goals and (ii) reliability. Relevancy pertains to whether the index actually assesses the valued ecological attribute(s) it was meant to assess. Reliability pertains to whether the index possesses appropriate statistical properties and whether it behaves as expected in response to human-caused stress. We designed this study to address how methods used to predict the biota under reference conditions affect the general performance of two types of indices. Because most assessments rely on comparisons of observed index values with those expected under reference conditions, the method used for estimating reference condition biota may affect index performance.

Relevancy generally refers to what an index measures. With respect to the U.S. Clean Water Act, indices should provide information on the biological integrity of surface waters. As first defined by Frey (1977), biological integrity refers to ''the capability of supporting and maintaining a balanced, integrated, adaptive community of organisms having a composition and diversity comparable to that of the natural habitats of the region''. Karr & Dudley (1981) subsequently added 'functional organisation' to Frey's definition. Reliability generally refers to whether an index responds in a predictable and interpretable manner to human-caused stressors and can distinguish responses to those stressors (signal) from the effects of naturally occurring environmental factors (noise). In general, an ideal index would respond linearly across a gradient of known ecological alteration. Furthermore, we would like to detect relatively small changes in conditions, and errors in estimating index values should be truly random and not associated with systematic biases caused by sampling or prediction errors.

One type of index (MMIs/indices of biological integrity) was specifically designed to assess the degree to which waterbodies achieve biological integrity by integrating information from several individual metrics, each of which characterises different ecological attributes. In general, MMIs should have high relevancy to the concept of biological integrity. As a measure of departure from expected overall assemblage structure (abundances of component taxa), the BC index is a good candidate as a comprehensive measure of alteration in overall biological integrity of an assemblage as defined. If our simulation results generally represent the behaviour of MMIs, a direct scaling between MMI values and biological integrity may be difficult. It follows that interpreting the ecological difference between bands or tiers of ecological condition in which the range of index values is typically divided at equal intervals into a few categories (e.g. excellent, good, fair, poor) will be difficult (see Davies & Jackson, 2006) as will mapping between different indices to allow comparison between regions or political entities (Birk & Hering, 2006). In our example, the magnitude of true difference between two higher quality tiers

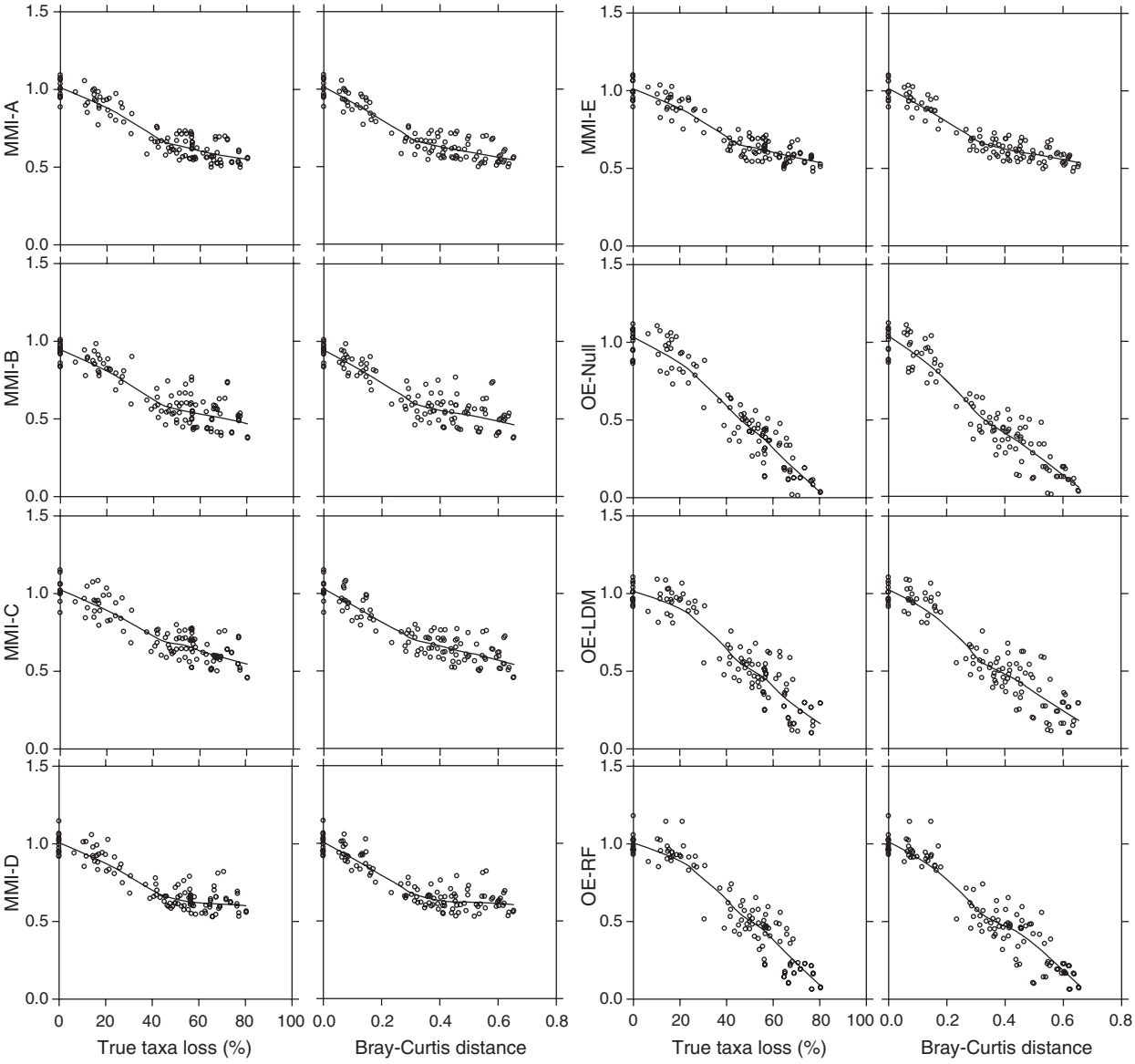

Fig. 5 Responses of five MMIs and three O/E indices to simulated changes in true taxa loss and increasing assemblage dissimilarity associated with increasing stress. Values for all 13 streams are plotted in each graph to show general patterns. Trend lines were fit by locally weighted scatter plot smoothing (LOWESS).

would not be the same as that between two lower quality tiers.

Most other indices are based on less comprehensive concepts and are thus less ambitious in their intent and interpretation. O/E indices (Moss *et al.*, 1987; Hawkins, 2006) quantify taxa loss or changes in taxonomic completeness, measures that primarily quantify aspects of taxa-level biodiversity. To the extent that protection of native biodiversity is important, O/E indices should also have high relevancy to policy and management. Because the response of the O/E index was more linear across the complete range of assemblage alteration, it can be more easily and directly translated into statements of true biodiversity condition (but see further discussion later). Moreover, given its linear tracking of BC distances based on $\log_{10}$ abundances, O/E also appears to be a good index of overall biological integrity.

Index performance and the challenge of adjusting for natural variability

Estimating the appropriate reference condition against which test sites should be assessed is probably the most vexing technical challenge facing the bioassessment community (e.g. Stoddard *et al.*, 2006; Herlihy *et al.*, 2008). Our results showed that inferences regarding ecological condition can be substantially influenced by how different models adjust for natural factors when estimating reference conditions – i.e. ecological health is, to some extent, in the 'eye of the modeller'. It is therefore imperative that bioassessment practitioners understand that how they choose to predict the reference condition will influence their assessments. There appears to be growing consensus that direct modelling of natural environmental gradients is superior to, or can improve on, landscape classifications when predicting local assemblage composition (Hawkins *et al.*, 2000a; Heino *et al.*, 2002; Sandin & Johnson, 2004; Davy-Bowker *et al.*, 2006; Hawkins, 2006), but little information exists regarding specific modelling approaches that yield the most reliable results. Our results showed that performance attributes that characterised precision and responsiveness varied independently of one another, and understanding how to improve index sensitivity will require consideration of both factors. In general, use of either linear (i.e. LDM for O/E, MLR for MMI) and tree-based approaches (CART and RF for MMI and O/E) can result in substantial improvements in index performance.

Bioassessment practitioners have largely focused on aspects of accuracy and precision when evaluating the performance of ecological indices (e.g. Diamond, Barbour & Stribling, 1996; Reynoldson *et al.*, 1997; Barbour *et al.*, 1999; Reynoldson, Rosenberg & Resh, 2001; Herbst & Sildorff, 2006; Mazor *et al.*, 2006). However, neither of these measures can be directly interpreted in context of the aspects of index performance that are perhaps ecologically and technically most important to resource managers. The responsiveness of an index is one such measure, which we defined as a measure of how much of an effect a given stress will have on an assemblage. Because the eight indices we examined clearly varied in their responsiveness to the same stress, they must therefore characterise the same stressed assemblages in different ways.

Our results also clearly showed that index sensitivity (here expressed as the proportion of managed sites inferred to be in NRC) is a function of both precision and responsiveness. Applied to the same sites, the eight indices varied nearly two-fold (32–54%) in the per cent of sites that were flagged as being in NRC. Such differences would have a profound effect on management decisions and the economic and ecological costs associated with those decisions. In our view, assessment of index performance should focus on understanding the factors that affect index precision, the ecological reasons indices differ in their responsiveness and how precision and responsiveness interact to influence index sensitivity.

Simulations of ecological impairment

Simulation has been widely used in many types of ecological studies (e.g. Minchin, 1987; Canham, Cole & Lauenroth, 2003; Elith & Graham, 2009). However, its application to testing ecological indices has been limited (e.g. Cao & Hawkins, 2005; Mazor *et al.*, 2006). We suspect its limited use is mainly associated with the fact that there are few historical applications of simulation of assemblage change to bioassessment issues that would have served to stimulate additional work. One potential criticism of simulation is that if it lacks realism, results cannot be trusted to apply to the real world. Unfortunately, field studies can also suffer from equally problematic issues that affect the reliability of inferences drawn from data. For example, in the real world, we use data from sites that are assumed to be biologically stressed by degraded environmental conditions to calibrate MMIs (see below). We further assume these sites are generally similar with respect to the magnitude of disturbance affecting biota. Given the wide range of ecological index values typically observed at such sites, this assumption is suspect, and hence it is unclear what exactly the index is being calibrated against. Furthermore, natural gradients are often strongly confounded with human-caused alterations to landscapes and waterways. It can therefore be difficult, if not impossible, to separate the effects of natural and stressor-related factors on index behaviour. Simulations provide a way of controlling for these issues, and when used in conjunction with field studies can provide additional independent lines of evidence from which to draw robust conclusions regarding index behaviour.

Our simulations identified two poorly understood aspects of index and ecological response to stress: (i) MMI responses to stress appear to saturate at intermediate levels of stress, whereas O/E indices do not, and (ii) stream assemblages appear to diverge in assemblage structure with increasing stress rather than converge to a homogeneous state consisting of a few tolerant taxa. Both of these observations have significant implications for the application and interpretation of ecological indices and will hopefully stimulate additional work designed to better understand the nature of and ecological significance of community-wide responses to stress.

We suspect the reasons that MMIs exhibited reduced response beyond intermediate levels of stress is related to how MMIs are usually calibrated. MMIs' are calibrated against both reference sites and degraded sites, whereas O/E indices are calibrated against only reference sites. For MMI calibration, degraded sites are defined by the existence of stressors (direct measures of in-stream conditions) or the potential of exposure to stressors (e.g. amount of land use in a catchment). Because calibration protocol dictates that actual biological condition should not be considered when selecting degraded sites (e.g. see Barbour *et al.*, 1999), there is no guarantee that the biota at the selected sites are also degraded. Furthermore, the calibration site with the poorest biological quality may not be the poorest in the region. The actual biological quality that exists at degraded calibration sites can often be highly variable (e.g. see Fig. 9-6 in Barbour *et al.*, 1999), which would tend to anchor the calibration of MMIs at intermediate levels of biological degradation. Our use of the range of each metric, instead of 95th and 5th percentiles (cf. Blocksom, 2003), should have helped minimise this problem, but it seems clear that if MMIs are calibrated to respond between minimal and moderate degradation, we should not expect their behaviour to be similar outside of those conditions. Unfortunately, solving this problem is not straightforward. For example, if the biologically poorest sites (i.e. completely degraded biology) were used to calibrate MMIs, then almost any biological metric would distinguish reference from degraded sites. In this case, distinguishing those metrics that were most responsive to moderate levels of stress from all others would be difficult if not impossible. The best solution to MMI calibration would seem to involve quantification of metric responses across the full range of possible stress (e.g. Yuan & Norton, 2003) as advocated and practiced by Fore & Grafe (2002) and Fore *et al.* (2007), with subsequent selection of that suite of metrics that, in combination, produce a continuous and near linear MMI response.

The simulations also produced some intriguing results unrelated to our primary goal of better understanding the statistical properties that affect index performance. The fact that as reference site assemblages were increasingly stressed, they became increasingly different from one another is of inherent ecological interest. There is growing concerns that increasing stress in stream ecosystems is leading to homogenisation of biotic assemblages (e.g. Olden & Poff, 2004; Poff *et al.*, 2007). Our results suggest a different pattern, at least with respect to assemblage similarity, and we do not think our results are artefacts of the way we simulated stress. In fact, our results are consistent with patterns reported in at least one recent empirical study (Pollard & Yuan, 2006). In our study and that by Pollard & Yuan (2006), stress appears to exclude sensitive taxa, resulting in only tolerant taxa remaining. However, the specific tolerant taxa that remain differ from site to site, presumably as a function of their other habitat-specific requirements with respect to natural differences between sites in temperature, substratum, etc. This interpretation leads to the view that heavily stressed systems, while tending to be depauperate, will be dominated by taxa somewhat unique to each stream.

Concluding remarks

Considering the financial and ecological costs and benefits that are associated with ecological assessment of freshwater ecosystems and associated management actions, we cannot afford to let ecological truth be defined by uninformed choice of how we predict reference condition biota. It is imperative that researchers and practitioners alike recognise that development and application of ecological indices without critical assessment of their behaviour, performance and ecological meaning has inherent dangers. Like the 'Blind Men and the Elephant' parable, without such information we could greatly err in how we characterise true ecological condition. No single ecological index developed to date completely captures all the varied ecological responses that

natural assemblages exhibit to human-caused stress (e.g. von der Ohe *et al.*, 2007), but probing index behaviour in as many ways as possible should go far towards allowing practitioners to see the whole 'elephant' of ecological condition and promote the selection of indices based on well-understood ecological and statistical properties.

Acknowledgments

This study was supported by the USDA Forest Service and the USDI Bureau of Land Management. Data were collected by the Pacfish/Infish Biological Effectiveness Monitoring Program. We are grateful to Rick Henderson for organising the environmental data and John Olson and Ryan Hill for deriving climate data. Completion of the work was facilitated by support from EPA Science To Achieve Results (STAR) grants R-82863701 and R-82863701. John Van Sickle and a second reviewer made recommendations that greatly improved the quality of this manuscript.

References

Baker E.A., Wehrly K.E., Seelbach P.W., Wang L., Wiley M.J. & Simon T. (2005) A multimetric assessment of stream condition in the Northern Lakes and Forests ecoregion using spatially explicit statistical modeling and regional normalization. *Transactions of the American Fisheries Society*, **134**, 697–710.

Barbour M.T., Gerritsen J., Snyder B.D. & Stribling J.B. (1999) *Rapid Bioassessment Protocols for Use in Streams and Rivers: Periphyton, Benthic Macroinvertebrates, and Fish*. EPA 841-B-99-002. United States Environmental Protection Agency, Office of Water, Washington, DC, U.S.A.

Birk S. & Hering D. (2006) Direct comparison of assessment methods using benthic macroinvertebrates: a contribution to the EU Water Framework Directive intercalibration exercise. *Hydrobiologia*, **566**, 401–415.

Blocksom K.A. (2003) A performance comparison of metric scoring methods for a multimetric index for mid-Atlantic highlands streams. *Environmental Management*, **31**, 670–682.

Breiman L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Breiman L., Friedman J.H., Olshen R.A. & Stone C.G. (1984) *Classification and Regression Trees*. Chapman and Hall, New York, NY, U.S.A.

Cairns J. Jr & Pratt J.R. (1993) A history of biological monitoring using benthic macroinvertebrates. In: *Freshwater Biomonitoring and Benthic Macroinvertebrates* (Eds D.M. Rosenberg & V.H. Resh), pp. 10–27. Chapman & Hall, New York, NY, U.S.A.

Canham C.D., Cole J.J. & Lauenroth W.K. (2003) *Models in Ecosystem Science*. Princeton University Press, Princeton, NJ, U.S.A.

Cao Y. & Hawkins C.P. (2005) Simulating biological impairment to evaluate the accuracy of ecological indicators. *Journal of Applied Ecology*, **42**, 954–965.

Cao Y., Hawkins C.P., Olson J. & Kosterman M.A. (2007) Modelling natural environmental gradients improves the accuracy and precision of diatom-based indicators for Idaho streams. *Journal of the North American Benthological Society*, **26**, 566–585.

Clarke R.T. & Murphy J.F. (2006) Effects of locally rare taxa on the precision and sensitivity of RIVPACS bioassessment of freshwaters. *Freshwater Biology* **51**, 1924–1940.

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

Davies S.P. & Jackson S.K. (2006) The biological condition gradient: a descriptive model for interpreting change in aquatic ecosystems. *Ecological Applications*, **16**, 1251–1266.

Davy-Bowker J., Clarke R.T., Johnson R.K., Kokes J., Murphy J.F. & Zahrádková S. (2006) A comparison of the European Water Framework Directive physical typology and RIVPACS-type models as alternative methods of establishing reference conditions for benthic macroinvertebrates. *Hydrobiologia* **566**, 91–105.

Diamond J.M., Barbour M.T. & Stribling J.B. (1996) Characterizing and comparing bioassessment methods and their results: a perspective. *Journal North American Benthological Society*, **15**, 713–727.

Draper N. & Smith H. (1981) *Applied Regression Analysis*. John Wiley and Sons, New York, NY, U.S.A.

Elith J. & Graham C.H. (2009) Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, **32**, 66–77.

Fore L.S. & Grafe C. (2002) Using diatoms to assess the biological condition of large rivers in Idaho (U.S.A.). *Freshwater Biology*, **47**, 2015–2037.

Fore L.S., Frydenborg R., Miller D., Frick T., Whiting D., Espy J. & Wolfe L. (2007) *Development and Testing of Biomonitoring Tools for Macroinvertebrates in Florida Streams (Stream Condition Index and Biorecon)*. Florida Department of Environmental Protection, Tallahassee, FL, U.S.A.

Frey D.G. (1977) Biological integrity of water- an historical approach. In: *The Integrity of Water: Proceedings of a Symposium, Washington, DC, March 10–12, 1975* (Eds R.K. Ballantine & L.J. Guarraia), pp. 127–140. U.S. Environmental Protection Agency, Washington, DC, U.S.A.

Hawkins C.P. (2006) Quantifying biological integrity by taxonomic completeness: its utility in regional and global assessments. *Ecological Applications*, **16**, 1277–1294.

Hawkins C.P., Norris R.H., Gerritsen J., Hughes R.M., Jackson S.K., Johnson R.K. & Stevenson R.J. (2000a) Evaluation of the use of landscape classifications for the prediction of freshwater biota: synthesis and recommendations. *Journal North American Benthological Society*, **19**, 541–556.

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

Heino J. & Mykra H. (2006) Assessing physical surrogates for biodiversity: do tributary and stream type classifications reflect macroinvertebrate assemblage diversity in running waters? *Biological Conservation*, **129**, 418–426.

Heino J., Muotka T., Paavola R., Hämäläinen H. & Koskenniemi E. (2002) Correspondence between regional delineations and spatial patterns in macroinvertebrate assemblages of boreal headwater streams. *Journal North American Benthological Society*, **21**, 397–413.

Henderson R.C., Archer E.K., Bouwes B.A., Coles-Richie M.C. & Kershner J.L. (2005) *PACFISH/*INFISH Biological Opinion (PIBO): Effectiveness Monitoring Program Seven-Year Status Report 1998 Through 2004. General Technical Report RMRS-GTR-162. United States Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO, U.S.A.

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

Herlihy A.T., Hughes R.M. & Sifneos J.C. (2006) National clusters of fish species assemblages in the conterminous United States and their relationship to existing landscape classification schemes. In: *Influences of Landscapes on Stream Habitats and Ecological Assemblages* (Eds R.M. Hughes, L. Wang & P.W. Seelbach), pp. 87–112. Vol. 48. American Fisheries Society, Bethesda, MD, U.S.A.

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

Hilsenhoff W.L. (1987) An improved biotic index of organic stream pollution. *The Great Lakes Entomologist*, **20**, 31–39.

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

Johnson R.K., Wiederholm T. & Rosenberg D.M. (1993) Freshwater biomonitoring using individual organisms, populations, and species assemblages of benthic macroinvertebrates. In: *Freshwater Biomonitoring and Benthic Macroinvertebrates* (Eds D.M. Rosenberg & V.H. Resh), pp. 40–158. Chapman & Hall, New York, NY, U.S.A.

Karr J.R. & Chu E.W. (1998) *Restoring Life in Running Waters: Better Ecological Monitoring*. Island Press, Washington, DC, U.S.A.

Karr J.R. & Dudley D.R. (1981) Ecological perspective on water quality goals. *Environmental Management*, **5**, 55–68.

Kershner J.L., Archer E.K., Coles-Ritchie M.S., Cowley E.R., Henderson R.C., Kratz K., Quimby C.M., Turner D.L., Ulmer L.C. & Vinson M.R. (2004) *Guide to Effective Monitoring of Aquatic and Riparian Resources. General Technical Report*. RMRS-GTR-121. United States Department of Agriculture, Forest Service, Rocky Mountain Experiment Station, Fort Collins, CO, U.S.A.

Legendre P. & Legendre L. (1998) *Numerical Ecology*. Elsevier, New York, NY, U.S.A.

Liaw A. & Wiener M. (2002) Classification and Regression by randomForest. *R News*, **2**, 18–22.

Lumley T.. (2004) The leaps package. cran.r-project.org/doc/packages/leaps.pdf.

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

Miller A.. (2002) *Subset Selection in Regression*, 2nd edn. Chapman & Hall, London.

Minchin P.R. (1987) An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio*, **69**, 89–107.

Montgomery D.R. & Macdonald L.H. (2002) Diagnostic approach to stream channel assessment and

monitoring. *Journal of the American Water Resources Association*, **38**, 1–16.

Moss D., Furse M.T., Wright J.F. & Armitage P.D. (1987) The prediction of the macro-invertebrate fauna of unpolluted running-water sites in Great Britain using environmental data. *Freshwater Biology*, **17**, 41–52.

von der Ohe P.C., Prüβ A., Schäfer R.B., Liess M., Deckere De. & Brack W. (2007) Water quality indices across Europe – a comparison of the good ecological status of the five river basins. *Journal of Environmental Monitoring*, **9**, 970–978.

Olden J.D. & Poff N.L. (2004) Ecological processes driving biotic homogenization: testing a mechanistic model using fish faunas. *Ecology*, **85**, 1867–1875.

Omernik J.M. (1987) Ecoregions of the conterminous United States. *Annals of the Association of American Geographers*, **77**, 118–125.

Omernik J.M. (1995) Ecoregions: a spatial framework for environmental management. In: *Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making* (Eds W.S. Davis & T.P. Simon), pp. 49–62. Lewis Publishers, Boca Raton, FL, U.S.A.

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

Plafkin J.L., Barbour M.T., Porter K.D., Gross S.K. & Hughes R.M. (1989) *Rapid Bioassessment Protocols for Use in Streams and Rivers: Benthic Macroinvertebrates and Fish*. EPA 440-4-89-001. United States Environmental Protection Agency, Office of Water Regulations and Standards, Washington, DC, U.S.A.

Poff N.L., Olden J.D., Merritt D.M. & Pepin D.M. (2007) Homogenization of regional river dynamics by dams and global biodiversity implications. *Proceedings of the National Academy of Science*, **104**, 5732–5737.

Pollard A.I. & Yuan L. (2006) Community response patterns: evaluating benthic invertebrate composition in metal-polluted streams. *Ecological Applications*, **16**, 645–655.

Pont D., Hugueny B., Beier U., Goffaux D., Melcher A., Noble R., Rogers C., Roset N. & Schmutz S. (2006) Assessing river biotic condition at a continental scale: a European approach using functional metrics and fish assemblages. *Journal of Applied Ecology*, **43**, 70–80.

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

Quigley T.M., Haynes R.W. & Graham R.T. (1996) *Integrated Scientific Assessment for Ecosystem Management in the Interior Columbia Basin, and Portions of the Klamath and Great Basins*. General Technical Report PNW-GTR-382. United States Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland, OR, U.S.A.

Resh V.H., Norris R.H. & Barbour M.T. (1995) Design and implementation of rapid assessment approaches for water resource monitoring using benthic macroinvertebrates. *Australian Journal of Ecology*, **20**, 108–121.

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

Reynoldson T.B., Rosenberg D.M. & Resh V.H. (2001) Comparison of models predicting invertebrate assemblages for biomonitoring in the Fraser River catchment, British Columbia. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 1395–1410.

Rosenberg D.M. & Resh V.H. (1993) Introduction to freshwater biomonitoring and benthic macro-invertebrates. In: *Freshwater Biomonitoring and Benthic Macroinvertebrates* (Eds D.M. Rosenberg & V.H. Resh), pp. 1–9. Chapman & Hall, New York, NY, U.S.A.

Sandin L. & Johnson R.K. (2004) Local, landscape and regional factors structuring benthic macroinvertebrate assemblages in Swedish streams. *Landscape Ecology*, **19**, 501–514.

Stevens D.L. & Olsen A.R. (1999) Spatially restricted surveys over time for aquatic resources. *Journal of Agricultural, Biological, and Environmental Statistics*, **4**, 415–428.

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

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

United States Environmental Protection Agency (2006) *Wadeable Streams Assessment: A Collaborative Survey of the Nation's Streams*. EPA 841-B-06-002. United States Environmental Protection Agency, Washington, DC, U.S.A.

Van Sickle J., Hawkins C.P., Larsen D.P. & Herlihy A.T. (2005) A null model for the expected macro-invertebrate assemblages in streams. *Journal North American Benthological Society*, **24**, 178–191.

Van Sickle J., Huff D.D. & Hawkins C.P. (2006) Selecting discriminant function models for predicting the expected richness of aquatic macroinvertebrates. *Freshwater Biology*, **51**, 359–372.

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

Waite T.A. & Cambell L.G. (2006) Controlling the false discovery rate and increasing statistical power in ecological studies. *Ecoscience*, **13**, 439–442.

Whittingham M.J., Stephens P.A., Bradbury R.B. & Freckleton R.P. (2006) Why do we still use stepwise modelling in ecology and behavior? *Journal of Animal Ecology*, **75**, 1182–1189.

Wright J.F., Sutliffe D.W. & Furse M.T. (Eds) (2000) *Assessing the Biological Quality of Fresh Water*. Freshwater Biological Association, Ambleside, Cumbria, U.K.

Yuan L.L. & Norton S.B. (2003) Comparing responses of macroinvertebrate metrics to increasing stress. *Journal North American Benthological Society*, **22**, 308–322.

Ziglio G., Flaim G. & Siligardi M. (2006) *Ecological Monitoring of Rivers (Water Quality Measurements)*. John Wiley and Sons, New York, NY, U.S.A.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Macroinvertebrate metrics used in several other studies that developed multimetric indices (MMIs) and the 37 specific metrics tested (*) in this study. More than one variant of a metric (number in parentheses) was sometimes calculated, i.e. % dominance based on different numbers of taxa or use of different tolerance value thresholds to define intolerant or tolerant taxa. We also tested a few other metrics that were not listed in the cited sources but that we considered to have potential for use in MMIs (−). RA = relative abundance of individuals. EPT = Ephemeroptera, Plecoptera and Trichoptera.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copyedited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

(Manuscript accepted 27 October 2009)