Environ Monit Assess (2014) 186:2135–2149
DOI 10.1007/s10661-013-3523-6

Tolerance values of benthic macroinvertebrates for stream biomonitoring: assessment of assumptions underlying scoring systems worldwide

Feng-Hsun Chang · Justin E. Lawrence · Blanca Rios-Touma · Vincent H. Resh

Received: 1 July 2013 / Accepted: 28 October 2013 / Published online: 11 November 2013

Abstract Tolerance values (TVs) based on benthic macroinvertebrates are one of the most widely used tools for monitoring the biological impacts of water pollution, particularly in streams and rivers. We compiled TVs of benthic macroinvertebrates from 29 regions around the world to test 11 basic assumptions about pollution tolerance, that: (1) Arthropoda are < tolerant than non-Arthropoda; (2) Insecta < non-Insecta; (3) non-Oligochaeta < Oligochaeta; (4) other macroinvertebrates < Oligochaeta + Chironomidae; (5) other macroinvertebrate taxa < Isopoda + Gastropoda + Hirudinea; (6) Ephemeroptera + Plecoptera + Trichoptera (EPT) < Odonata + Coleoptera + Heteroptera (OCH); (7) EPT < non-EPT insects; (8) Diptera < Insecta; (9) Bivalvia < Gastropoda; (10) Baetidae < other Ephemeroptera; and (11) Hydropsychidae < other Trichoptera. We found that the first eight of these 11 assumptions were supported despite regional variability. In addition, we examined the effect of Best Professional Judgment (BPJ) and non-independence of TVs among countries by performing all analyses using subsets of the original dataset. These subsets included a group based on those systems using TVs that were derived from techniques other than BPJ, and groups based on methods used for TV assignment. The results obtained from these subsets and the entire dataset are similar. We also made seven a priori hypotheses about the regional similarity of TVs based on geography. Only one of these was supported. Development of TVs and the reporting of how they are assigned need to be more rigorous and be better described.

Keywords Benthic macroinvertebrates · Biological monitoring · Pollution tolerance · Saprobity · Water pollution

Introduction

Biological monitoring of rivers and streams has its origins over a century ago (Metcalfe 1989; Cairns and Pratt 1993; Dolédec and Statzner 2010) and a wide variety of taxonomic groups have been suggested for use, ranging from viruses and bacteria to plants, macroinvertebrates, and fish (Hellawell 1984). In examining the advantages

F.-H. Chang
Institute of Ecology and Evolutionary Biology,
National Taiwan University,
Taipei 10617, Taiwan

F.-H. Chang · J. E. Lawrence · V. H. Resh
Department of Environmental Science, Policy,
and Management, University of California,
Berkeley, CA 94720-3114, USA

J. E. Lawrence · V. H. Resh (✉)
Re-inventing the Nation's Urban Water Infrastructure
(ReNUWIt), National Science Foundation Engineering
Research Center, University of California,
Berkeley, CA 94720-3114, USA
e-mail: resh@berkeley.edu

B. Rios-Touma
Institute of Urban and Regional Development,
College of Environmental Design, University of California,
Berkeley, CA 94720, USA

and disadvantages reported for benthic diatoms, zooplankton, macroinvertebrates, and fish, Resh (2008) concluded that of all these potential bioindicators, macroinvertebrates provided the highest return on investment in terms of information gained for research funds spent. Indeed, benthic macroinvertebrates are the most widely used group for biological monitoring of streams and rivers around the world (Bonada et al. 2006).

Most currently used macroinvertebrate analyses for water quality assessments typically involve the use of tolerance values (TVs), often in calculating scores of biotic indices (Carter and Resh 2013). The use of TVs, which describe resistance of organisms to pollution, has a long-standing tradition in aquatic biomonitoring programs. For example, the initial attempts at biological monitoring (i.e., the saprobien system of Kolkwitz and Marsson 1902) were based on the premise that different taxa have different pollution tolerance (Bonada et al. 2006).

The long-standing tradition of using TVs has resulted in a variety of basic assumptions about the tolerance of benthic macroinvertebrates that we believe have become entrenched in aquatic biomonitoring programs and are assumed to be correct. These assumptions include that: (1) Arthropoda are less pollution tolerant than non-Arthropoda; (2) Insecta are less tolerant than non-Insecta; (3) non-Oligochaeta are less tolerant than Oligochaeta; (4) other benthic macroinvertebrate taxa are less tolerant than Oligochaeta + Chironomidae; (5) other benthic macroinvertebrate taxa are less tolerant than Isopoda + Gastropoda + Hirudinea; (6) Ephemeroptera + Plecoptera + Trichoptera (EPT) are less tolerant than Odonata + Coleoptera + Heteroptera (OCH); (7) EPT are less tolerant than non-EPT insects; (8) Diptera are less tolerant than other Insecta; (9) Bivalvia are less tolerant than Gastropoda; (10) Baetidae are more tolerant than other Ephemeroptera; and (11) Hydropsychidae are more tolerant than other Trichoptera. These assumptions were selected because they include the most commonly used biomonitoring metrics reported by state agencies in the US (from Carter and Resh 2013, their Table 4) and from perusing international programs (Resh 2007).

We believe that these assumptions were based on the results of earlier biomonitoring studies conducted in particular regions, such as in Europe (described by Metcalfe 1989), South Africa (Chutter 1972), and North America (Hilsenhoff 1982, 1987). Tolerance was often determined by a group of organisms' response to dissolved-oxygen deficits resulting from sewage inputs. These tolerances generally have been assumed to hold true for regions and pollutants other than for the ones for which they were developed. However, we have not found any reports indicating that the accuracy or robustness of the above-described assumptions has been statistically tested.

The objectives of this study are to: (1) compile available information about how TVs reported for benthic macroinvertebrates are developed in different regions around the world; (2) statistically evaluate the validity of the 11 basic assumptions (described above) concerning the TVs of specific groups of benthic macroinvertebrates, combinations of benthic-macroinvertebrate groups, and of tolerance-based metrics; and (3) assess how TVs vary geographically, within macroinvertebrate groups at different taxonomic levels and the influence of methods used to derive TVs.

Methods

The progression of our methods is as follows; we: (1) compiled information on TVs worldwide; (2) converted the TVs to the same scaling system; (3) tested the validity of the 11 basic assumptions using permutation tests and bootstrapping applied to the worldwide dataset containing all regions; (4) used five subsets of this worldwide dataset to evaluate potential effects of interdependencies among regions on the results; (5) used one subset of this worldwide dataset to evaluate the effect of Best Professional Judgment (BPJ); and (6) used k-medoids clustering to examine the geographic distribution of TVs within the macroinvertebrate groups at different taxonomic levels and the influence of their derivation methods.

Obtaining information on TVs

We collected information on TVs used in 29 regions located on six continents and Oceania (Table 1). We only used information based on numerical scores assigned to individual benthic-macroinvertebrate families, which resulted in the exclusion of metrics based on multi-metric indexes, biological traits, or molecular data. Most TV databases are published in the grey literature rather than in peer-reviewed journals, and some TVs were only able to be obtainable through email contact with researchers directly involved in the score

Table 1 Data sources used in the analyses

Continent	Regions	References
North America	United States (Midwest)	Hilsenhoff (1987, 1988)
	United States (California)	Ode (2003)
	United States (New York) Canada	New York State Department of Environmental Conservation (2012); Mandaville (2002)
South America	Colombia	Roldán (2003)
	Chile	Figueroa (2004)
	High Andes of Ecuador and Peru	Acosta et al. (2009); Rios-Touma et al. (2013)
	Costa Rica	Decreto Presidencial No. 33903-MINAE-0S (2007)
	Bolivia	Rocabado and Gotia (2011)
	Brazil	Junqueira et al. (2010)
Australia and Oceania	Australia	Chessman (1995)
	New Zealand	Stark (1993)
Asia	Thailand	Mustow (2002)
	Mekong	MRC (2007; 2010); Chessman and Giap (2010); Resh et al. (2013)
	India	De Zwart and Trivedi (1994)
	China (Eastern)	B. X. Wang, unpublished data
	China (Yangtze)	Wang and Yang (2004)
Europe	Great Britain	Walley and Hawkes (1996)
	Spain	Zamora et al. (1995)
	Poland	Unpublished document
	France	AFNOR (1992), Verneaux et al. (1982)
	Belgium	De Pauw and Vanhooren (1983)
	Latvia	EU-STAR (2005)
	Germany	Schmidt-Kloiber and Hering (2012)
	Austria	Schmidt-Kloiber and Hering (2012)
	Czech Republic	Schmidt-Kloiber and Hering (2012)
	Slovakia	Schmidt-Kloiber and Hering (2012)
Africa	South Africa	Dickens and Graham (2002)
	Egypt	Fishar and Williams (2008)

development. We typically only used one set of TVs from each country. However, we used three sets from the US (New York, Midwest, California) and two sets from China (Eastern region and Yangtze River) because of their large size.

Scaling scores

In many regions, TVs are based on an 11-point numerical scale that represents a gradient of pollution resistance. For example, in the US, TVs typically range from 0 to 10, with 0 representing highly intolerant organisms and ten highly tolerant (e.g., Hilsenhoff 1987; Lenat 1993). A 10-point scale is used in Europe, but it is in reverse order with 1 representing highly tolerant and ten highly intolerant (e.g., Armitage et al. 1983). Similarly, the eastern region of China uses a system with ten classes, with 1 representing tolerant and 10 intolerant (Wang and Yang 2004). Other regions use different scales (e.g., France uses a scale from 9 to 1; Austria, Czech Republic, Germany, and Slovakia from 0 to 4; Mekong River Basin from 100 to 1; South Africa from 15 to 1; Costa Rica from 9 to 0; and Brazil from 1 to 4; see Table 1 for references). Japan uses only two classes of tolerance, A and B (Tsuda and

Morishita 1974), and was thus excluded from the worldwide analysis because of the limited potential to differentiate TVs among taxa.

To compare the TVs from different regions, we first converted all the scores to a uniform 10-point scale ranging from 1 to 10, where 1 represents least tolerant and 10 most tolerant. We then converted all the original scores into this 10-point scale using linear interpolation, rounding the converted scores to the nearest whole integer.

Statistical comparisons

We restricted our analyses to permutation tests and bootstrapping, which are non-parametric re-sampling techniques that can be used with categorical data (Anderson 2001; Good 2005). These two methods do not require the assumptions of traditional parametric tests (e.g., normal distributions and homogeneity of variance, as for *t*-test) to be met (Collingridge 2013). These tests were the best available choice because the TVs had non-normal distributions resulting from different methods of development or assignment, and from differences in professional knowledge in the different regions (P. deValpine, University of California, Berkeley, personal communication).

We tested: (1) the validity of 11 basic assumptions described in the Introduction; (2) which taxonomic groups above the level of order are significantly different from one other in terms of these assumptions; (3) which aquatic insect orders and biomonitoring metrics (e.g., EPT, OCH, and non-EPT taxa) based on these orders are significantly different from each other; and (4) which regions clustered together, based on their TVs. The False Discovery Rate (Benjamini and Hochberg 1995) correction was applied to all *p* values. All permutation and bootstrapping procedures were conducted in R statistical software.

The validity of 11 basic assumptions regarding TVs

We used permutation tests to evaluate the first nine of the 11 assumptions described above; the final two were evaluating using bootstrapping. Both permutation and bootstrapping tests are non-parametric methods of re-sampling used to test significance (Legendre and Legendre 2012). We performed permutation tests for the first nine of the 11 assumptions because we were examining if the TVs of two distinct metrics were significantly different from each other. In contrast, we performed bootstrapping for the last two assumptions because we were examining if the TV of a certain family is significantly different from the rest of the TVs in its respective order (i.e., the precision of the sample). In testing the assumptions, we first calculated the average TV for each benthic-macroinvertebrate family among all the regions examined, which we refer to as the family average. For the permutation tests, we used these family averages to calculate the pre-permutation value (i.e., the value derived directly from the TV databases) of the metrics analyzed (e.g., EPT and OCH) or of taxonomic groups (e.g., Arthropoda, Insecta, Oligochaeta).

The permutation tests included 10,000 random permutations, each of which included the appropriate number of families specific to the metric(s) or taxonomic group(s) being compared. For example, each permutation comparing Arthropoda (272 families) versus non-Arthropoda (87 families) included a total of 359 family averages (272+87), which were randomly drawn from the pool of family averages contained in the Arthropoda (i.e., 272 were drawn) and non-Arthropoda (i.e., 87 were drawn) groups, and were distributed accordingly. Based on the random combination of family averages that occurs during each permutation, we calculated permutated values (i.e., values representing the average of family averages from the randomization procedure). In each permutation, we calculated the difference between the average of two permutated values, which represent the two metric(s) or taxonomic classification(s). This permutation procedure was repeated 9,999 times to generate a distribution of all the differences between the permutated values. Lastly, we included the difference between the averages of the pre-permutation values as the 10,000th iteration in this distribution. The likelihood of the difference between two pre-permutation values occurring in this distribution provides the *p* value, which indicates whether the difference is statistically significant.

We used bootstrapping to compare the TV of the individual families examined to that of other members of their order (e.g., the TV of the mayfly family Baetidae to that of other taxa in the order Ephemeroptera). For example, we first used the family averages to calculate the average TV of Ephemeroptera. Then, we calculated the difference between the family average of Baetidae and the average TV of Ephemeroptera, which we refer to as the pre-bootstrapping difference. The family average of Baetidae was fixed, whereas the average of

Ephemeroptera was variable because we re-sampled family averages within the order with replacement to obtain a bootstrapped average TV score (e.g., we drew 34 family averages with replacement from the 34 family averages within Ephemeroptera to calculate a bootstrapped average, which varied with each iteration). Subsequently, we calculated the difference between the family average of Baetidae and this bootstrapped average of the order, repeated this procedure for 9,999 iterations, and included the pre-bootstrapping difference as the 10,000th iteration. The likelihood of the pre-bootstrapping difference occurring in this distribution provides the *p* value, which indicates whether the difference is significant. This bootstrapping procedure was also used to compare Hydropsychidae to other Trichoptera.

Accounting for non-independence

The worldwide dataset contains scoring systems that are potentially not independent of one another because one program may use the same (or modified) TVs from another program. However, there presumably is some independent thought applied by researchers in each region to modify TVs, such as by BPJ. To explore this issue of potential non-independence, we repeated all the permutation tests and bootstrapping described above on five subsets of the entire dataset that were grouped according to the scoring system they derived TVs from (Table 1), including the use of a: (1) locally derived methods group, as done in the Mekong River basin, China (Eastern), and South Africa; (2) Hilsenhoff method group, including US (Midwest, California, and New York), Canada, and China (Yangtze); (3) Trent index method group, including France and Belgium; (4) Saprobien System method group, including Germany, Slovakia, Austria, Latvia, Czech Republic, and Brazil; and (5) Biological Monitoring Working Party (BMWP) method group, including all the remaining 13 countries. We then repeated the procedures described above to test the robustness of the results from using the entire dataset.

Accounting for potential effects of BPJ

We formed a final subset because many of the regions examined depend somewhat on BPJ to determine their TVs. Although not documented, we believe that BPJ often relies on some of the 11 basic assumptions that our study aimed to test. To test this circularity issue, we prepared a subset consisting of five least BPJ-related method regions, including the: (1) Mekong River Basin, where TVs are based on environmental condition; (2) China (Eastern), which is diversity-index based; (3) Brazil and Germany, which are Saprobien-system based; and (4) France, which is Trent Index-based. All permutation tests and bootstrapping were performed on this subset. We then repeated the procedures described above to test the robustness of the results from using the entire dataset.

Differences tested among taxonomic groups and among metrics

We tested for differences among ten higher taxonomic groups above the level of order, including Insecta, Arthropoda, non-Oligochaeta, Bivalvia, non-Insecta, Hirudinea, non-Arthropod, Isopoda, Gastropoda, and Oligochaeta. For each of these groups, the lower and upper 95 % confidence intervals (CI) of the average were determined using bootstrapping. Non-overlap in the 95 % CI indicated significant difference between the two averages.

We tested for differences among aquatic Insecta orders (Plecoptera, Trichoptera, Ephemeroptera, Odonata, Diptera, Heteroptera, Lepidoptera), and commonly used metrics (EPT, OCH, and non-EPT insects) for benthic-macroinvertebrate biomonitoring. Orders with TVs for <4 families, which included Neuroptera (3) and Megaloptera (2), were excluded because of their small sample size. For each of the groups examined, the lower and upper 95 % CI of the average were determined using permutation tests. Non-overlap in the CI indicates significant difference between the two averages.

Differences among regions

In terms of regional similarities, we based our expectations solely according to geographical proximity and climate type, although we were aware that other related factors, such as altitude and cultural/historical ties, could affect the similarity of TVs. Consequently, our expectations were that: (1) TVs in the North America (Midwest, California, and New York in USA, and Canada) should differ because of the distance among these regions; (2) in Europe, north-temperate countries should differ from Spain because of climate; (3) Spain should differ from Central Europe because of climate; (4) however, the

Central European countries, including Czech Republic, Poland, Austria, Slovakia, Germany, Belgium, and Latvia should not differ among themselves; (5) Asian countries (Thailand, India, Mekong River basin, China) should differ from other, non-Asian countries because of geographic and climatic differences, and we also expected these Asian countries or regions to be grouped according to their geographic proximity; (6) in South America, Colombia, Bolivia, Ecuador–Peru should not differ because of location and climatic similarities; (7) and lastly, in Latin America, Costa Rica and Chile should differ because of the distances among them.

An alternative hypothesis was that countries would be grouped by the methods used for deriving TVs (Table 1). To test these hypotheses, we first prepared a modified Euclidean-distance matrix among countries to represent the differences among their scoring systems. The modified Euclidean distance is the Euclidean distance divided by the number of families in common between any two given countries. This distance matrix was subsequently used to perform cluster analysis using the k-medoids method (Kaufman and Rousseeuw 1987), which is a non-hierarchical clustering technique that achieves maximum within-cluster homogeneity without relying on hierarchy. Another advantage of k-medoids is that it is based on the Partitioning Around Medoids (PAM) algorithm, which minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean distances (Reynolds et al. 1992). Therefore, the result from k-medoids is robust to noise and outliers. The k-medoid procedures were performed by using the "cluster" package in R (Maechler et al. 2013).

To determine the most appropriate number of groupings for the cluster analysis, we first calculated the Average Silhouette Width (ASW) of each cluster for groups of different sizes. A higher ASW value indicates that the dataset is more highly clustered and each cluster is more likely to be homogeneous (Rousseeuw 1987). Kaufmann and Rousseuw (2005) suggest that an ASW value ≥0.5 indicates the data are structured, 0.25–0.5 indicates the possibility of structure, and <0.25 indicates lack of structure. We defined the most appropriate number of groups to be that number resulting in the highest ASW, with the condition that this ASW must be ≥0.25. Lastly, we superimposed the grouping onto an NMDS (Non-metric Multidimensional Scaling) ordination plot, which is a statistical method widely used to visualize the similarity of objects in large, complex datasets (Legendre and Legendre 2012; Oksanen et al. 2013).

Results

Analysis of the methods used to assign TVs

BPJ is the method most widely used worldwide for assigning TVs, and most regional TVs are usually modified from those used in other geographical areas. For example, 72 % of the regions examined reported using BPJ, at least in part, to develop TVs (Table 2; methods 2, 3, 4, and 5). Over two-thirds (79 %) of programs drew scores from one region and modified them to fit another region (Table 2; methods 3, 4, 5, and 8, excluding Germany where Saprobien TVs were originally developed). The next-most-used method of TV assignment was based on the Saprobien System, which uses the frequency of occurrence of species or taxonomic units at different saprobity (pollution) levels to assign the TV (or saprobity score) to each taxon. Other approaches used in a single region or country are based on either using a mathematical equation that relies on site disturbance scores assigned by BPJ of human impacts to assign TVs (method 6) or using the Shannon–Weiner taxonomic diversity of sites to assign TVs (method 7).

Testing tolerance assumptions

We found that eight out of 11 of the basic assumptions tested were statistically valid (Table 3). The exceptions were that Gastropoda and Bivalvia were not significantly different in their TVs, and that neither Baetidae nor Hydropsychidae were significantly different from Ephemeroptera or Trichoptera, respectively.

The coefficients of variation (CV) of the different orders, combination of orders, and metrics indicated that the variability was generally low (Table 3). The metrics Oligochaeta, and Oligochaeta + Chrionomidae, had the lowest CVs, 3–10 %, whereas the metrics Insecta, Arthropoda, non-Oligochaeta taxa, all taxa except Oligochaeta + Chrionomidae, Isopoda + Gastropoda + Hirudinea, EPT, Diptera, and other Insecta beside Diptera all had CVs in excess of 30 %.

Table 2 Examples of tolerance value (TV) assignment methods used worldwide with a description of how widely used each method is and where it is applied at the regional level

Method	% of entire dataset	Region	Example references
1. Early methods used for TV assignment based on BPJ.		South Africa, UK, USA	Chutter (1972); Hilsenhoff (1982); Armitage et al. (1983)
2. TVs derived from Hilsenhoff or BMWP are reassigned based on the average value of a biological index where each taxon is present.	10	Midwest (US), Canada, Great Britain	Hilsenhoff (1987); Walley and Hawkes (1996)
3. TVs of BMWP (UK) are modified based on BPJ.	41	Spain, Poland, Egypt, Thailand, India, Australia, New Zealand, Costa Rica, Colombia, Ecuador–Peru, Chile, Bolivia	Stark (1993); Zamora-Muñoz et al. (1995)
4. TVs derived from Trent Index (Woodiwiss 1964) are assigned based on a combination of BPJ and ecological profiles.	7	France, Belgium	Verneaux and Tuffery (1967); Vernaux et al. (1982)
5. TVs derived from Hilsenhoff are modified based on BPJ.	14	California (US), New York (US), Canada, China (Eastern)	Mandaville (2002); Ode (2003); Wang and Yang (2004)
6. TVs are calculated using a mathematical approach based on Site Disturbance Scores (SDS).	3	Mekong basin	Chesmann and Giap (2010); Resh et al. (2013)
7. TVs are assigned using a mathematical approach using the frequency of occurrence of each taxon in categories (Excellent, Good, Good–fair, Fair, and Poor) of water quality created based on Shannon–Weiner diversity.	3	China (Yangtze)	B. X. Wang, Nanjing Agricultural University, unpublished data.
8. TVs are assigned based on the Saprobien method, which uses the frequency of occurrence of species in the different saprobic (pollution) levels that are defined by environmental variables.	21	Germany, Austria, Slovakia, Czech, Latvia, Brazil	Wegl (1983); Junqueira (2010)

Methods 3 to 5 use adaptations of earlier TVs (from method 1) to other regions

BPJ Best Professional Judgment, *BMWP* Biological Monitoring Working Party

Differences among higher-level taxonomic groups

The higher taxonomic groups showed statistically significant differences among each other (Fig. 1a). Insecta had the lowest TVs (95 % CI of mean=4.4–4.8), whereas Oligochaeta had the highest (95 % CI of mean=8.5–9.0). Insecta had significantly lower TVs than all other groups examined except Arthropoda, whereas Oligochaeta had significantly higher TVs than all other groups. Except for non-Oligochaeta, Arthropoda had significantly lower TVs than all of the more tolerant groups (i.e., those to the right of Arthropoda in Fig. 1a). Non-Oligochaeta also had significantly lower TVs than all the more tolerant groups to the right in Fig. 1a. Bivalvia had significantly lower TVs than Oligochaeta. Non-Insecta had significantly lower TVs than Hirudinea and Oligochaeta. Non-Arthropoda had significantly lower TVs than Hirudinia and Oligochaeta but not Isopoda. Isopoda were significantly less tolerant than Oligochaeta, but not Gastropoda or non-Arthropoda (Fig. 1).

Differences among aquatic insect orders and metrics based on these orders

The aquatic insect orders and metrics based on these orders also showed statistically significant differences from each other (Fig. 1b). Plecoptera had significantly lower TVs (95 % CI of mean=1.9–2.6) than all the other groups examined (95 % CI=3.3–6.8), whereas Lepidoptera had the highest (95 % CI=5.8–6.8). The EPT orders when combined had a significantly lower

Table 3 Taxonomic comparisons underlying 11 basic assumptions regarding tolerance values (TVs) of benthic macroinvertebrates

Taxonomic comparison	Number of families in dataset	Mean TV	SD	CV (%)	*p* value
Arthropoda	272	4.8	1.6	32	0.0001
Non-Arthropoda	87	6.2	1.7	27	
Insecta	213	4.6	1.6	35	0.0001
Non-Insecta	146	6.0	1.5	24	
Oligochaeta	11	8.8	0.4	5	0.0001
Non-Oligochaeta	348	5.0	1.6	31	
Oligochaeta + Chironomidae	12	8.7	0.6	6	0.0001
All other taxa	347	5.0	1.6	31	
Isopoda + Gastropoda + Hirudinea	41	6.2	0.7	11	0.0001
All other taxa	318	5.0	1.7	35	
EPT	95	3.6	1.2	33	0.0001
OCH	75	5.4	1.0	18	
EPT	95	3.6	1.2	33	0.0001
Non-EPT insecta	118	5.4	1.5	27	
Diptera	30	5.8	2.0	35	0.0002
Other Insecta	183	4.4	1.5	33	
Bivalvia	9	5.9	1.1	18	0.2678
Gastropoda	30	6.1	0.7	11	
Baetidae	1	5.6	–	–	0.4993
Other Ephemeroptera	34	3.9	1.0	27	
Hydropsychidae	1	5.4	–	–	0.5053
Other Trichoptera	45	3.8	1.1	29	

Summary statistics: *SD* standard deviation, *CV* coefficient of variation, *EPT* Ephemeroptera, Plecoptera, Trichoptera, *OCH* Odonata, Coleoptera, Heteroptera

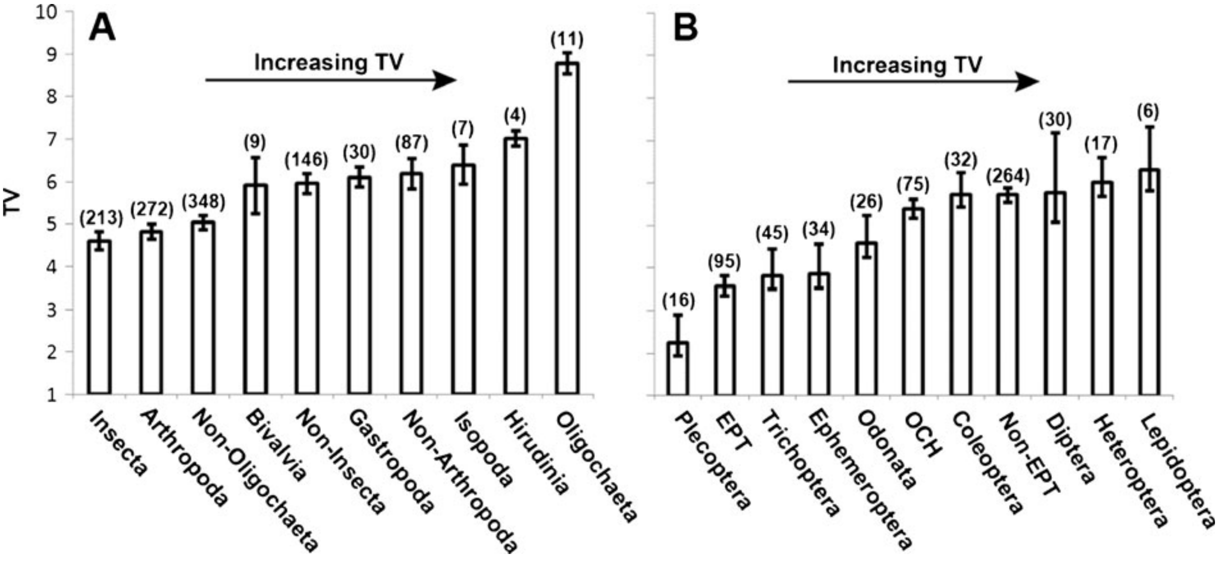

Fig. 1 Comparison of average tolerance values (*TVs*) for **a** higher classifications of aquatic organisms and **b** aquatic insect orders and metrics based on these orders. The error bars represent the 95 % confidence interval, which was determined using permutation tests. Non-overlap of this interval indicates statistical significance. *EPT* Ephemeroptera, Plecoptera, and Trichoptera; *OCH* Odonata, Coleoptera, Heteroptera. The *numbers in parentheses* are the number of families included in the averages

TV than all the more tolerant groups to the right in Fig. 1b, except for Trichoptera and Ephemeroptera. However, EPT was significantly higher than for Plecoptera alone. Trichoptera had significantly lower TVs than all the more tolerant groups except Ephemeroptera, but was significantly higher than Plecoptera. Ephemeroptera had significantly lower TVs than all the more tolerant groups to the right in Fig. 1b. Odonata had significantly lower TVs than all the more tolerant groups. OCH had significantly lower TVs than all the more tolerant groups except Coleoptera and non-EPT Insects. Non-EPT Insects, Diptera, Heteroptera, and Lepidoptera were not significantly different from each other.

Accounting for potential effects of BPJ and non-independence

When we validated our results from our worldwide dataset using a subset of the five regions with non-BPJ-derived methods (e.g., Mekong River Basin, China [Eastern], and Germany), we found similar results as when using the worldwide dataset. However, these five regions showed two contrasting results. TVs of Diptera were not significantly higher than other Insecta, and Bivalvia were significantly lower than Gastropoda. In terms of non-dependence issues, however, differences among these five method groups were not statistically significant and were also comparable to those derived from the entire dataset with only two exceptions. First, when considering only locally developed methods, Insecta TVs were not significantly lower from those of non-Insecta. Second, when considering the Saprobien System users, Diptera TVs were not significantly more tolerant from other Insecta.

Tests of hypotheses about differences among regions

The combination of TVs among the regional scoring systems examined clustered best into five groupings, based on an ASW of 0.26 (indicating the possibility of structure). These groupings ranged in number of regions from one (France) to 14 (Great Britain, Spain, Poland, Thailand, India, Australia, New Zealand, South Africa, Costa Rica, Colombia, Ecuador–Peru, Chile, Bolivia, and Brazil) (Fig. 2). The other three groups had two (Egypt and Mekong River Basin), five (Belgium, Germany, Austria, Slovakia, Latvia, Czech Republic), and six (all four North American and two Chinese) regions. When we reexamined the data with regions arranged into the five clustered groupings, the results were generallyly the same as those from the analysis performed on the worldwide dataset.

We only found support for one of our seven a priori geographic hypotheses regarding the distribution of TVs, which was (hypothesis 6) that Colombia, Bolivia, and Ecuador–Peru should group together. However, even in this case they were also grouped with Great

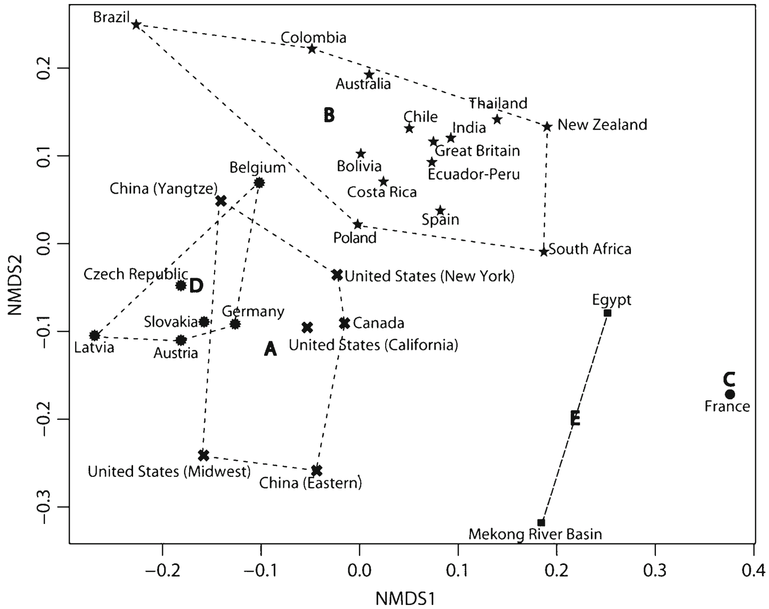

Fig. 2 K-medoid cluster superimposed on 2-D NMSD ordination plot. Groups are indicated by *bold capital letters* (i.e., A, B, C, D, E) and regions within each group share the same symbol. *Dotted lines* indicate the perimeter of each group in *n*-dimensional space

Britain, Spain, Poland, Thailand, India, Australia, New Zealand, South Africa, Costa Rica, Chile, and Brazil. In the US, the Midwest, California, and New York were not different (reject hypothesis 1). In Europe, north temperate countries (UK) were not different from Spain (reject hypothesis 2). Spain was in the same group as Poland, a central European country (reject hypothesis 3). Poland was in a different group than the other central European countries (reject hypothesis 4). Thailand and India were together, but the Mekong and China were in different groups (reject hypothesis 5). Costa Rica and Chile were in the same group (reject hypothesis 7).

When we examined regions to determine if the clustering followed the type of scoring system used, we found some agreement. For example, group B (Fig. 2) comprised all of the regions using BMWP-derived scores. However, Brazil, which does not use this system, was also in group B. In contrast, the two systems used in China clustered together, but they are based on different scoring systems.

In terms of the variability of TVs, the modified Euclidean distance between regions grouped by scoring system used (i.e., locally derived, Hilsenhoff, Trent Index, Saprobien, and BMWP; Fig. 3) indicated that the BMWP-derived scores had the least variability (i.e., the smallest difference between the first and third quartiles) and, not unexpectedly, locally derived scores had the highest variability. Regions that used the Hilsenhoff-derived scores and those that use the Saprobien-derived scores are more similar in their TVs, but with more variation than those regions using BMWP-derived scores. We note that the above pattern could have resulted simply from statistical artifact, because there were more BMWP-derived systems (13) than Hilsenhoff-derived (5) or Saprobien-derived systems (6). The Trent Index-derived TVs have very low variability because they only include Belgium and France.

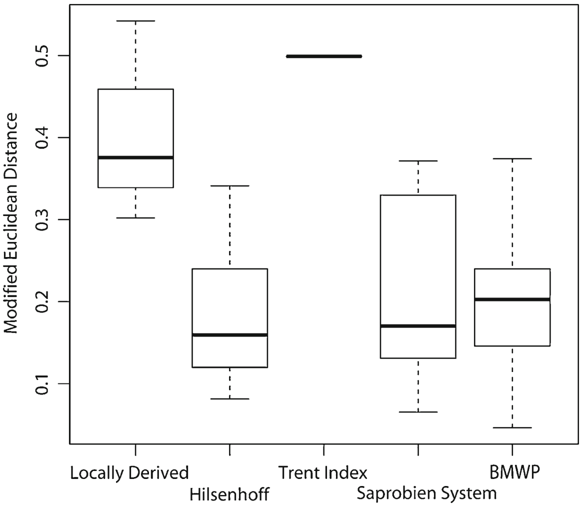

Fig. 3 Box and whisker plot showing tolerance value (*TVs*) variation among regions within the five groups of methods used to calculate TVs. *Bold horizontal lines* represent the median of pairwise distances between regions, lower and upper end of boxes represent the first and third quartiles, respectively, and the lower and upper error bars represent the minimum and maximum, respectively

Discussion

The lack of information provided by programs about how TV scores are developed is surprising given how much weight TVs are given in regulatory decisions, such as those regarding the location and need for wastewater treatment plants and other urban planning issues (Chessman 1995; Purcell et al. 2002). Moreover, we were surprised that most programs use TVs developed by others, albeit with modifications.

Most modified TVs were based on local knowledge and BPJ. The poor performance of multi-metric biotic indices has been attributed in some cases to incorrect TVs for some taxa to heavy metals (Hickey and Clements 1998). Given that innovative projects are underway to reinvent urban water infrastructure for ecosystem benefits, such as streamflow and wetland augmentation with recycled water (Bischel et al. 2013; Halaburka et al. 2013; Lawrence et al. 2013), reliable TVs could play a key role in evaluating the success of such habitat rehabilitation efforts.

Typically, BPJ is informed by both environmental and macroinvertebrate data, relies to a large degree on institutional knowledge, and its programmatic applications have differed among regions and over time (Carter and Resh 2013). For example, in the US, most state-based TVs are based off adaptations of the original Hilsenhoff values, whereas in the UK TVs are assigned by a commission of experts (BMWP, described in Hawkes 1997). Hellawell (1978) found a strong relationship of TVs resulting from the BMWP with taxonomic diversity metrics. However, Washington (1984) questioned the value of diversity indices in terms of their appropriateness to water quality assessments.

Carter and Resh (2013) examined the issue of BPJ and TVs for the programs in the different US states. They found that TVs are generally reported as total tolerance to pollution, tolerance to organic pollution, or tolerance to metals. Most TVs reported by state programs refer to organic loading (87 %), but far fewer referred to total tolerance (13 %) and metal tolerance (28 %). The source for TVs also varied among US programs. Local expertise accounted for 31 % of program choices, which was followed by values from Barbour et al. (1999) (27 %). Values from Hilsenhoff (1982, 1987, 1988, 1998) (13 %) and Lenat (1993) (9 %) were less commonly cited as used.

The lack of geographic specificity, taxonomic resolution, and stressor specificity of TVs represent a limitation in their use. However, several US states are developing TVs that are specific to their regions and for specific stressors such as metals, acid mine drainage, and sediments (Carter and Resh 2013). Bonada et al. (2006) suggested that TVs would be more intuitive to apply if they were on a linear scale where, for example, a TV of 10 would represent an organism that has twice the tolerance as an organism with a score of 5.

Most (eight of 11, or 73 %) of the assumptions that were the basis for our hypotheses were supported in our worldwide analysis of TVs using the entire dataset, and the results were generally similar following our use of subsets of these data. There were two exceptions. First, Bivalvia and Gastropoda were not different from each other in their TVs in the analysis of the entire dataset (but they were different in the subsets). However, this failure to detect a significant difference could relate to the relatively low number of families in the original data for Bivalvia and Gastropoda relative to other groups. Bivalvia had nine families and Gastropoda had 30 families. The lower number of families would result in reduced statistical power. In contrast, the average number of families in the other groups for which there were significant differences examined was 128 (SD=117).

Second, neither the Baetidae nor the Hydropsychidae were significantly different from other families in their respective orders (i.e., Ephemeroptera and Trichoptera, respectively). Both are commonly occurring and abundant families that are often found in mildly polluted waters in some regions (Ratia et al. 2012; Xu et al. 2013). In particular, filter-feeding Hydropsychidae tend to benefit from increases in particulate matter, such as from wastewater treatment plant or fish farm effluents (Paul 2011; Guilpart et al. 2012). However, as with all generalizations about the perceived higher tolerance of Baetidae and Hydropsychidae, there are many exceptions, and broad generalizations are difficult at taxonomic levels above species (Lenat and Resh 2001). The lack of any significant difference in our analysis likely reflects the different levels of tolerance seen by species in these families that occur in the different regions of the world and the difficulty of applying family-level TVs.

Plecoptera have long been regarded as the most pollution intolerant of the aquatic insect orders (e.g., Friedrich et al. 1992), and this widely held view was also supported in our results, which showed that Plecoptera had lower TVs than all the other groups, including Ephemeroptera and Trichoptera. It is generally accepted that Plecoptera evolved in cold mountain streams where oxygen stress was minimal (Zwick 2000), and we would expect lower TVs because oxygen depletion accompanies a wide range of human disturbances. However, several studies report Plecoptera occurring in organic polluted streams (Bispo et al. 2002; Tomanova and Tedesco 2007) and metal-polluted acidic streams (Rosemond et al. 1992; Sjøbakk et al. 1997; Ruse and Herrmann 2000). Furthermore, there is a high variability among the species of this order in terms of tolerating metal pollution (Ruse and Herrmann 2000) and low oxygen (Tomanova and Tedesco 2007) concentrations.

The potential interdependence in the scores among the regions examined led us to conduct an internal validation based on a subset of the entire dataset. This subset included only TVs that were derived from methods other than BPJ. We saw general agreement in results between this subset and the entire dataset.

The multi-dimensional illustration of clustering among countries (expressed in two dimensions in Fig. 2) indicates that TVs were not distributed as expected, but that methods of TV assignment and geographic proximity were often factors in explaining these differences. For example, the North America and Central European countries each clustered according to their expected locations (Fig. 2, groups A and D). These two clusters each have their own respective methods of TV assignment, i.e., the US uses the Hilsenhoff system and Central Europe the Saprobien system (Table 2). Likewise, group B countries clustered together because they all use the BMWP system, with Brazil being an exception. Instead, Brazil uses values derived from the Saprobien system and also uses values at the generic or family level whereas Central Europe generally uses species level, which may explain it not being in group

D. However, in group B, Brazil is at the extreme edge of the grouping (Fig. 2). To us, it is unclear why France does not cluster with other regions. TVs coming from the Mekong River Basin and Egypt likely cluster together because they were developed along two large river systems, the Mekong and the Nile. We also note that other factors could lead to the clustering pattern that we found (Fig. 2). For example, altitude could be influential. As the number of macroinvertebrate families decreases with increasing altitude, the sites at higher elevation become more similar to each other than to those at lower altitudes (Prat et al. 2009, Villamarin et al. 2013). This could explain why Costa Rica and Chile were grouped together because the TVs of both countries do not differentiate among high and low altitude areas.

Several methods have been proposed to improve TVs. For example, there have been programs that based TVs entirely on environmental values (e.g., for the Mekong River Basin, MRC 2007; 2010; Resh et al. 2013). In the western US, Whittier and Van Sickle (2010) used multiyear environmental data and principal component analysis to construct a synthetic disturbance variable, which was used with averages and weights of taxa to calculate TVs. Another multivariate approach assumes a Gaussian response of populations to multiple environmental variables (Juggins 1997). Such approach thus calculates, on a taxon-by-taxon basis, both the degree of intolerance using the mode of population abundance (i.e., the environmental optima) and the range of tolerance using the standard deviation of the mode (Bonada et al. 2004). Other new techniques use generalized additive models (GAMs) to define taxa as tolerant, intermediately tolerant, or sensitive to phosphates, pH, suspended solids, or other specific stressors (Yuan 2004). Smith et al. (2007) defined the sensitivity of macroinvertebrates to nutrients and Utz et al. (2009) defined their sensitivity to land-use coverage at the catchment scale.

The globalization of TVs does have clear limitations. Families may be present, for example in tropical regions (Thorne and Williams 1997; Resh 2007) that do not occur in the temperate regions that typically develop these scores. To solve this problem for missing families, scoring systems in Ecuador–Peru used extensive literature studies to determine their TVs (Rios-Touma et al. 2013). However, globalization can provide synergy. An awareness of what methods are used among different regions to develop TVs and which methods are most effective is a clear advantage whereas decreased awareness can delay biomonitoring advances. For example, the US multimetric system and the UK multivariate-based systems developed largely independent of each other (Resh and Yamamoto 1994). It has taken over two decades for the advantages of each approach to be combined in current biomonitoring programs (Carter and Resh 2013). Because TVs are the foundation of many regional multi-metric approaches that are being used or under development, it is crucial for these scores to be as accurate as possible.

There are also local or national influences that can hinder the improvement of TVs in biological monitoring programs. For example, once a monitoring system is shown to be appropriate for a region or country, and is in place, changing it can be quite difficult even when its limitations are demonstrated and potential improvements proposed. Examples of where changes from long-used indices have been made include some countries in Europe (e.g., Italy and France), and these changes have been strongly influenced by the mandate of the European Framework Directive.

In summary, we collected and examined all the literature that we could find reporting family-level TVs. We subsequently standardized those values and applied non-parametric statistical methods (e.g., permutation tests and bootstrapping) to test 11 basic assumptions about the tolerance of benthic macroinvertebrates, and to examine the geographic and climatic relationships of their TVs among regions. Our comprehensive, global-scale study reveals that those basic TV assumptions are generally supported, and suggests the need for new, perhaps more robust methods of TV development and the reporting of how TVs are assigned.

Acknowledgments We thank P. DeValpine for statistical advice, D. Hering and N. Prat for providing data and information, and D. Rosenberg and J. Carter for their comments on an earlier draft of this article. J. Lawrence received support from the National Science Foundation Engineering Research Center for Reinventing the Nation's Urban Water Infrastructure, ReNUWIt (NSF EEC-028968). B. Rios Touma has a postdoctoral fellowship from the Ecuadorian National Secretariat of Higher Education, SENESCYT (contract 20120122).

References

Acosta, R., Rios-Touma, B., Rieradevall, M., & Prat, N. (2009). Proposal for an evaluation protocol of the ecological quality on andean rivers (CERA) and its use in two basins in Ecuador and Peru. *Limnetica, 28*(1), 35–64.

AFNOR. (1992). *Détermination de l'indice biologique global normalisé (IBGN) [Determination of the Standardized Global Biological Index (IBGN)]. Report No. NF T 90350*. Paris: AFNOR.

Anderson, M. J. (2001). Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences, 58*(3), 626–639.

Armitage, P. D., Moss, D., Wright, J. F., & Furse, M. T. (1983). The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. *Water Research, 17*(3), 333–347.

Barbour, M. T., Gerritsen, J., Snyder, B. D., & Stribling, J. B. (1999). *Rapid bioassessment protocols for use in streams and wadable rivers: Periphyton, benthic macroinvertebrates and fish, 2nd edition. EPA 841-B-99-002*. Washington: Environmental Protection Agency, Office of Water.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. *Series B (Methodological), 57*(1), 289–300.

Bischel, H. N., Lawrence, J. E., Halaburka, B. J., Plumlee, M. H., Bawazir, A. S., King, J. P., McCray, J. E., Resh, V. H., & Luthy, R. G. (2013) Renewing urban streams with recycled water for streamflow augmentation: hydrologic, water quality, and ecosystem services management. Environmental Engineering Science 30(8):455–479. doi:10.1089/ees.2012.0201.

Bispo, P. D. C., Froehlich, C. G., & Oliveira, L. G. (2002). Stonefly (Plecoptera) fauna of streams in a mountainous area of Central Brazil: abiotic factors and nymph density. *Revista Brasileira de Zoologia, 19*(1), 325–334.

Bonada, N., Zamora-Muñoz, C., Rieradevall, M., & Prat, N. (2004). Ecological profiles of caddisfly larvae in Mediterranean streams: implications for bioassessment methods. *Environmental Pollution, 132*(3), 509–521.

Bonada, N., Prat, N., Resh, V. H., & Statzner, B. (2006). Developments in aquatic insect biomonitoring: a comparative analysis of recent approaches. *Annual Review of Entomology*. doi:10.1146/annurev.ento.51.110104.151124.

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

Carter, J. L., & Resh, V. H. (2013). *Analytical approaches used in stream benthic macroinvertebrate biomonitoring programs of state agencies in the USA* (pp. 2013–1129). Menlo Park: United States Geological Survey Open-file Report.

Chessman, B. C. (1995). Rapid assessment of rivers using macroinvertebrates: a procedure based on habitat-specific sampling, family level identification and a biotic index. *Australian Journal of Ecology, 20*(1), 122–129.

Chessman, B. C. & Giap, D. H. (2010). Biological metrics calculation. In: MRC (Mekong River Commission). *Biomonitoring methods for the Lower Mekong Basin* (pp. 57–60). Vientiane, Lao P.D.R.: Mekong River Commission.

Chutter, F. M. (1972). An empirical biotic index of the quality of water in South African streams and rivers. *Water Research, 6*(1), 19–30.

Collingridge, D. S. (2013). A primer on quantitized data analysis and permutation testing. *Journal of Mixed Methods Research, 7*(1), 79–95.

De Pauw, N., & Vanhooren, G. (1983). Method for biological quality assessment of watercourses in Belgium. *Hydrobiologia, 100*(1), 153–168.

De Zwart, D., & Trivedi, R. C. (1994). *Manual on integrated water quality evaluation. Report 802023003*. Bilthoven, The Netherlands: National Institute of Public Health and Environmental Potection (RIVM).

Decreto Presidencial Nº 33903-MINAE-S. (2007). Reglamento para la Evaluación y Clasificación de la Calidad de Cuerpos de Agua Superficiales. La Gaceta-Diario Oficial. AÑO CXXIX. #178-7pp.

Dickens, C. W., & Graham, P. (2002). The South African Scoring System (SASS) version 5 rapid bioassessment method for rivers. *African Journal of Aquatic Science, 27*(1), 1–10.

Dolédec, S., & Statzner, B. (2010). Responses of freshwater biota to human disturbances: contribution of J-NABS to developments in ecological integrity assessments. *Journal of the North American Benthological Society, 29*(1), 286–311.

EU-STAR. (2005). Standardization of river classifications. Protocols. Energy, Environment and Sustainable Development Programme. http://www.eu-star.at/pdf/LatvianMacroinvertebrateSamplingProtocol.pdf. Accessed 30 May 2013.

Figueroa, R. (2004). *Calidad Ambiental de la Cuenca Hidrográfica del río Chillán, VIII Región, Chile. Tesis Doctoral*. Malaga: Universidad de Málaga.

Fishar, M. R., & Williams, W. P. (2008). The development of a biotic pollution index for the River Nile in Egypt. *Hydrobiologia, 598*(1), 17–34.

Friedrich, G., Chapman, D., & Beim, A. (1992). The use of biological material. In D. Chapman (Ed.), *Water quality assessments—a guide to using biota, sediments and water in environmental monitoring* (pp. 171–238). London: Chapman and Hall.

Good, P. I. (2005). *Permutation, parametric and bootstrap tests of hypotheses* (3rd ed.). New York: Springer.

Guilpart, A., Roussel, J. M., Aubin, J., Caquet, T., Marle, M., & Le Bris, H. (2012). The use of benthic invertebrate community and water quality analyses to assess ecological consequences of fish farm effluents in rivers. *Ecological Indicators, 23*(1), 356–365.

Halaburka, B. J., Lawrence, J. E., Bischel, H. N., Plumlee, M. H., Hsiao J., Resh, V. H., & Luthy, R. G. 2013. Economic and ecological costs and benefits of streamflow augmentation using recycled water in a California coastal stream. Environmental Science & Technology.

Hawkes, H. A. (1997). Origin and development of the biological monitoring working party score system. *Water Resources, 32*(3), 964–968.

Hellawell, J. M. (1978). *Biological surveillance of rivers*. Stevenage: Water Research Center.

Hellawell, J. M. (1984). *Biological indicators of freshwater pollution and environmental management*. London: Elsevier Applied Science.

Hickey, C. W., & Clements, W. H. (1998). Effects of heavy metals on benthic macroinvertebrate communities in New Zealand streams. *Environmental Toxicology, 17*(11), 2338–2346.

Hilsenhoff, W. L. (1982). *Using a biotic index to evaluate water quality in streams*. Madison: Department of Natural Resources.

Hilsenhoff, W. L. (1987). An improved biotic index of organic stream pollution. *Great Lakes Entomologist, 20*(1), 31–40.

Hilsenhoff, W. L. (1988). Rapid field assessment of organic pollution with a family-level biotic index. *Journal of the North American Benthological Society, 7*(1), 65–68.

Hilsenhoff, W. L. (1998). A modification of the biotic index of organic pollution to remedy problems and permit its use throughout the year. *Great Lakes Entomologist, 31*(1), 1–12.

Juggins, S. (1997). *CALIBRATE version 0.70. A CCC program for analyzing and visualizing species environment relationships and for predicting environmental values from species assemblages, User guide Version 1.0*. Newcastle: Department of Geography.

Junqueira, M. V., Friedrich, G. N., & de Araujo, P. R. P. (2010). A saprobic index for biological assessment of river water quality in Brazil (Minas Gerais and Rio de Janeiro states). *Environmental Monitoring and Assessment, 163*(1–4), 545–554.

Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of Medoids. In Y. Dodge (Ed.), *Statistical data analysis based on the L_1-norm and related methods* (pp. 405–416). North-Holland: Birkhäuser Basel.

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data. An introduction to cluster analysis*. Hoboken: John Wiley & Sons, Inc.

Kolkwitz, R., & Marsson, M. (1902). Grundsafer die biologische Beurteilung des Wassers nach seiner Flora und Fauna. *Mitteilungen Koniglichen Prüfungsanstalt Wasser Abwasser, 1*(1), 3–72.

Lawrence, J. E., Pavia, C. P. W., Kaing, S., Bischel, H. N., Luthy, R. G., & Resh, V. H. 2013. Recycled water for augmenting urban streams in Mediterranean-climate regions: a potential approach for riparian ecosystem enhancement. Hydrological Sciences Journal.

Legendre, P., & Legendre, L. (2012). *Numerical ecology* (3rd ed.). Oxford: Elsevier.

Lenat, D. R. (1993). A biotic index for the southeastern United States: derivation and list of TVs, with criteria for assigning water-quality ratings. *Journal of the North American Benthological Society, 12*(3), 279–290.

Lenat, D. R., & Resh, V. H. (2001). Taxonomy and stream ecology—the benefits of genus- and species-level identifications. *Journal of the North American Benthological Society, 20*(2), 287–298.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2013). Cluster: cluster analysis basics and extensions. *R Package Version, 1*(14), 4.

Mandaville, S. M. (2002). Benthic macroinvertebrates in freshwaters—taxa tolerance, values, metrics, and prospects. Report of Project H-1, Soil & Water Conservation Halifax, Canada: Society of Metro Halifax.

Metcalfe, J. L. (1989). Biological water quality assessment of running waters based on macroinvertebrate communities: history and present status in Europe. *Environmental Pollution, 60*(1–2), 101–139.

MRC (Mekong River Commission) (2007). Biomonitoring of the Lower Mekong River and selected tributaries 2004–2007. MRC Technical Paper No. 20. Vientiane, Lao P.D.R.: Mekong River Commission.

MRC (Mekong River Commission) (2010). *Biomonitoring methods for the Lower Mekong Basin*. Vientiane, Lao P.D.R.: Mekong River Commission.

Mustow, S. E. (2002). Biological monitoring of rivers in Thailand: use and adaptation of the BMWP score. *Hydrobiologia, 479*(1–3), 191–229.

New York State Department of Environmental Conservation. (2012). Standard Operating Procedure: Biological Monitoring of Surface Waters in New York State. Report: NYSDEC SOP 208–12 Stream Biomonitoring Rev. 1.0. Albany, NY: Division of Water.

Ode, P. (2003). List of California macroinvertebrate taxa and standard taxonomic effort. California Department of Fish & Game, Aquatic Bioassessment Laboratory. CAMLnet. http://www.safit.org/Docs/CABW_std_taxonomic_effort.pdf.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B, Simpson, G. L., Solymos, P., Stevens, M. H. H., & Wagner, H. (2013). vegan: community ecology package. R package version 2.0-7. http://CRAN.R-project.org/package=vegan.

Paul, J. S. (2011). *Life history and secondary production of Cheumatopsyche lasia Ross (Trichoptera: hydrophyschidae) with respect to a wastewater treatment facility in a north central Texas urban stream*. Denton: M.S. thesis, University of North Texas.

Prat, N., Ríos, B., Acosta, R., & Rieradevall, M. (2009). Los macroinvertebrados como indicadores de calidad de las aguas. En: E. Domínguez, & H. Fernández (Eds.), *Macroinvertebrados bentónicos sudamericanos*, Primera ed. San miguel de Tucumán Fundación Miguel Lillo. 631–654.

Purcell, A. H., Friedrich, C., & Resh, V. H. (2002). An assessment of a small urban stream restoration project in Northern California. *Restoration Ecology, 10*(4), 685–694.

Ratia, H., Vuori, K. M., & Oikari, A. (2012). Caddis larvae (Trichoptera, Hydropsychidae) indicate delaying recovery of a watercourse polluted by pulp and paper industry. *Ecological Indicators, 15*(1), 217–226.

Resh, V. H. (2007). Multinational, freshwater biomonitoring programs in the developing world: lessons learned from African and Southeast Asian river surveys. *Environmental Management, 29*(5), 737–748.

Resh, V. H. (2008). Which group is best? Attributes of different biological assemblages used in freshwater biomonitoring. *Environmental Monitoring and Assessment, 138*(1–3), 131–138.

Resh, V. H., & Yamamoto, D. (1994). International collaboration in freshwater ecology. *Freshwater Biology, 32*(3), 613–624.

Resh, V. H., Campbell, I. C., & Chessman, B. C. (2013). Human-disturbance tolerance values for littoral and benthic macroinvertebrates, benthic diatoms, and zooplankton of the Lower Mekong River, Southeast Asia. Aquatic Biology, 2013.

Reynolds, A., Richards, G., de la Iglesia, B., & Rayward-Smith, V. (1992). Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms, 5*(4), 475–504.

Rios-Touma, B., Acosta, R., & Prat, N. (2013). The Andean Biotic Index (ABI): revised tolerance to pollution values for macroinvertebrate families and index performance evaluation. Revista de Biología Tropical.

Rocabado, G., & Gotia, E. (2011). *Guía para evaluar la calidad acuática mediante el índice BMWP/Bol*. La Paz: Ministerio del Ambiente y Agua de Bolivia.

Roldán, G. A. (2003). *Bioindicación de la calidad del aguas en Colombia. Uso del método BMWP/Col*. Medellin: Editorial Universidad de Antioquia.

Rosemond, A. D., Reice, S. R., Elwood, J. W., & Mulholland, P. J. (1992). The effects of stream acidity on benthic invertebrate communities in the south-eastern United States. *Freshwater Biology, 27*(2), 193–209.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics, 20*(1), 53–65.

Ruse, L., & Herrmann, S. (2000). Plecoptera and Trichoptera species distribution related to environmental characteristics of the metal-polluted Arkansas River, Colorado. *Western North American Naturalist, 60*(1), 57–65.

Schmidt-Kloiber, A. & Hering, D. (eds.) (2012). The taxa and autecology database for freshwater organisms, version 5.0. http://www.freshwaterecology.info (accessed on 20 March 2013).

Sjøbakk, T. E., Almli, B., & Steinnes, E. (1997). Heavy metal monitoring in contaminated river systems using mayfly larvae. *Journal of Geochemical Exploration, 58*(2–3), 203–207.

Smith, A. J., Bode, R. W., & Kleppel, G. S. (2007). A nutrient biotic index (NBI) for use with benthic macroinvertebrate communities. *Ecological Indicators, 7*(2), 371–386.

Stark, J. D. (1993). Performance of the macroinvertebrate community index: effects of sampling method, sample replication, water depth, current velocity, and substratum on index values. *New Zealand Journal of Marine and Freshwater Research, 27*(4), 463–478.

Thorne, R. S. J., & Williams, W. P. (1997). The response of benthic macroinvertebrates to pollution in developing countries: a multimetric system of bioassessment. *Freshwater Biology, 37*(3), 671–686.

Tomanova, S., & Tedesco, P. (2007). Body size, ecological tolerance and potential for water quality bioindication in the genus Anacroneuria (Plecoptera: Perlidae) from South America. *Revista de Biología Tropical, 55*(1), 67–81.

Tsuda, M., & Morishita, I. (1974). *Methods of the water quality monitoring by organisms (translated from Japanese "生物による水質調査法"*). Tokyo: Sankaido.

Utz, R. M., Hilderbrand, R. H., & Boward, D. M. (2009). Identifying regional differences in threshold responses of aquatic invertebrates to land cover gradients. *Ecological Indicators, 9*(3), 556–567.

Verneaux, J., & Tuffery, G. (1967). Une méthode zoologique practique de détermination de la qualité biologique des eaux courantes. Indices biotiques. *Annales Scientifiques de l'Universite de Besançon, Zoologique, 3*(1), 79–90.

Verneaux, J., Galmiche, P., Jannier, F., & Monnot, A. (1982). Une nouvelle méthode pratique d'évaluation de la qualité des eaux courantes. Un indice biologique de qualité générale (I. B. G.). *Annales Scientifiques de l'Université de Franche Comté, 4*(1), 11–21.

Villamarín, C., Rieradevall, M., Paul, M. J., Barbour, M. T., & Prat, N. (2013). A tool to assess the ecological condition of tropical high Andean streams in Ecuador and Peru: The IMEERA index. *Ecological Indicators, 29*, 79–92.

Walley, W., & Hawkes, H. (1996). A computer-based reappraisal of the Biological Monitoring Working Party scores using data from the 1990 river quality survey of England and Wales. *Water Research, 30*(9), 2086–2094.

Wang, B. X., & Yang, L. F. (2004). A study on tolerance values of benthic macroinvertebrate taxa in Eastern China. *Acta Ecologica Sinica, 24*(1), 2768–2775.

Washington, H. G. (1984). Diversity, biotic, and similarity indices: a review with a special relevance to aquatic ecosystems. *Water Research, 18*(6), 653–694.

Wegl, R. (1983). Index fu r die Limnosaprobität. *Wasser und Abwasser, Band 26*, 175 S.

Whittier, T. R., & Van Sickle, J. (2010). Macroinvertebrate tolerance values and an assemblage tolerance index (ATI) for western USA streams and rivers. *Journal of the North American Benthological Society, 29*(3), 852–866.

Woodiwiss, F. (1964). The biological system of stream classification used by the Trent River Board. *Chemistry and Industry, 14*(1), 443–447.

Xu, M., Wang, Z., Duan, X., & Pan, B. (2013). Effects of pollution on macroinvertebrates and water quality bio-assessment. *Hydrobiologia*. doi:10.1007/s10750-013-1504-y.

Yuan, L. L. (2004). Assigning macroinvertebrate tolerance classifications using general additive models. *Freshwater Biology, 49*(5), 662–667.

Zamora-Muñoz, C., Sáinz-Cantero, C. E., Sánchez-Ortega, A., & Alba-Tercedor, J. (1995). Are biological indices BMPW' and ASPT' and their significance regarding water quality seasonally dependent? Factors explaining their variations. *Water Research, 29*(1), 285–290.

Zwick, P. (2000). Phylogenetic system and zoogeography of the Plecoptera. *Annual Review of Entomology, 45*(1), 709–746.