United States
Environmental Protection
Agency

The Evaluation of Methods for Creating Defensible, Repeatable, Objective and Accurate Tolerance Values for Aquatic Taxa

RESEARCH AND DEVELOPMENT

EPA 600/R-06/045
May 2006
www.epa.gov

The Evaluation of Methods for Creating Defensible, Repeatable, Objective and Accurate Tolerance Values for Aquatic Taxa

Karen A. Blocksom
National Exposure Research Laboratory

Lori Winters
Oak Ridge Institute for Science and Education

U.S. Environmental Protection Agency
Office of Research and Development
National Exposure Research Laboratory
Cincinnati, OH 45268

U.S. Environmental Protection Agency
Office of Research and Development
Washington, DC 20460

Notice

The information in this document has been funded in part by the United States Environmental Protection Agency under Interagency Agreement DW89938954 to Oak Ridge Institute for Science and Education. It has been subjected to the Agency's peer and administrative review and has been approved for publication as an EPA document.

Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Executive Summary

Introduction

In the field of bioassessment, tolerance has traditionally referred to the degree to which organisms can withstand environmental degradation. This concept has been around for many years and its use is widespread. In numerous cases, tolerance values (TVs) have been assigned to individual taxa or groups of taxa to represent their tolerance to pollution. The TVs are then often combined into metrics which describe characteristics of aquatic communities. Perhaps the most familiar example is the Hilsenhoff Biotic Index (HBI) (Hilsenhoff, 1977), an index that has been incorporated into many bioassessment programs. The HBI is typically very useful in distinguishing among sites of higher and lower water quality. To calculate the HBI, each environmental agency or organization typically uses its own set of tolerance values. However, the origins of these values, and rationales for their selection, are often obscure and unverifiable. Available methods for deriving TVs more objectively vary substantially in approach and complexity. Therefore, this study conducted systematic comparisons of existing lists of macroinvertebrate TVs and their resulting HBI scores. It also compared several objective TV derivation approaches, as well as bioassessment metrics derived from each, to determine their repeatability and sensitivity to disturbance. All analyses were run at the family and genus levels.

Comparison of Tolerance Values

Existing lists of macroinvertebrate TVs from Delaware, Kentucky, Maryland, Massachusetts, and a U.S. Environmental Protection Agency lab were assembled into a single database for the purpose of direct comparisons. At the family level, there were not systematic differences in TVs, and correlations were high among lists. However, at the genus level, the Kentucky list differed significantly from all other lists, although correlations among lists were only slightly lower overall. In both cases, considerable variation was observed in bivariate plots of TVs from all possible pairs of lists. This variation was somewhat muted when TVs were incorporated into HBI scores, particularly at the genus level, but systematic differences among lists were more obvious for family level HBI scores. In both cases, the lowest correlations for HBI scores occurred between the EPA list and other TV lists. As much as HBI scores and TVs varied among lists, the processes used to develop TVs by individual organizations also seem to vary and often depend on professional judgment. Although such processes may provide effective tools for bioassessment, they are not repeatable, and confidence in the TVs assigned to individual taxa is therefore limited.

Comparison of Methods to Derive Tolerance Values

Several more objective approaches to developing TVs have been proposed, and this study compared and evaluated several for repeatability. To compare objective TV

development approaches, data from a single stream study in the mid-Atlantic highlands were divided randomly into calibration and validation sets, and each approach was carried out on both data sets. All but one approach consisted of first defining a disturbance gradient and then using one of two procedures to calculate the TV. One approach used EPT richness to describe the disturbance level, another used a principal components analysis (PCA), and a third used PCA with generalized additive modeling (GAM) to more precisely model the relationship between the probability of taxon presence and disturbance. A fourth approach relied on the observed frequency of a taxon compared to its expected frequency using predictive modeling. For the EPT, PCA, and GAM approaches, two procedures each for calculating TVs were examined, one based on a single value from the gradient defined and the other a weighted average. The mechanics of carrying out each approach are covered in detail and examples are provided. The TVs generated with each approach were compared between the calibration and validation data sets. Three tolerance-based macroinvertebrate metrics based on these two sets of TVs were compared over the same set of sites. In addition, the ability of these metrics to distinguish between reference and impaired sites was evaluated for each method.

Results of analysis varied across the four broad approaches that were evaluated. The EPT approach resulted in strong correlations between data sets, but there were also significant shifts in TVs from one data set to the other. In addition, the EPT approach was the least defensible approach because of its circular nature. The PCA approach tended to differ less between data sets, but correlations were not as strong. The predictive modeling approach as applied in this study exhibited high correlations between data sets and allowed calculation of TVs for the largest number of taxa. However, this approach only produced TVs for taxa that would be expected to be found at reference sites. The GAM approach provided TVs for a much more limited number of taxa and required more observations than any of the other methods. In this way, the GAM approach “selected” those taxa that showed a relatively strong relationship with the stressor gradient and excluded those without a strong association of some kind. In general, weighted procedures for calculating TVs from the defined disturbance gradient resulted in higher repeatability than procedures that identified a specific value from a distribution.

Although absolute metric values varied widely across approaches, general patterns tended to be similar among most methods. The distinction between reference and impaired sites typically improved from family- to genus-level data. Among the metrics evaluated, intolerant taxa richness was most useful overall because it consistently distinguished most strongly between reference and impaired and was repeatable across most approaches. HBI scores also discriminated well between reference and impaired sites but was slightly more variable than intolerant taxa richness. Percent intolerant individuals exhibited the most variability in values when calculated for the same set of sites using TVs based on calibration and validation data sets.

Recommendations

All of the approaches evaluated exhibited some degree of repeatability but varied with respect to the type of data and the degree of statistical experience or training required. Ultimately, these characteristics are most important in choosing an approach. When abiotic data are available that adequately characterize the gradient of disturbance in the region of interest, employing an approach that incorporates these data is more desirable. The PCA and GAM approaches both utilize extensive abiotic data to characterize sites having corresponding macroinvertebrate data. In addition, the GAM approach benefits from additional site information (e.g., watershed area, elevation, etc.) to account for natural variability. These approaches more directly relate taxonomic occurrence or abundance with the level of disturbance, and this may make the resulting TVs more defensible. For both the PCA and GAM approaches, the procedures for calculating TVs that used all available count data (i.e., weighted and weighted average, respectively), rather than just an optimum or percentile, produced more consistent results. Of the statistical techniques employed by these two approaches, PCA is simpler to perform and is available widely in statistical software packages, but GAMs may more precisely describe the relationship between individual taxa and environmental gradients. However, neither approach will be as valuable if the abiotic data lack variables that are important in describing the disturbance gradient.

If biotic and certain abiotic data are available on a set of samples identified as representing reference condition, the predictive modeling approach may be most useful. This is particularly true if the variables included in the abiotic data can be used to characterize natural classes of samples. Predictive modeling is particularly attractive in situations where limited or no data are available to describe the disturbance gradient itself, as the gradient is dealt with indirectly in this approach. However, this approach involves several steps that require the use of potentially complex multivariate statistical techniques. The techniques required can be found in many statistical software packages, but a lot of movement of data among different programs may be necessary to complete the development of TVs. In addition, some specialized statistical training or experience may be required to carry out the necessary techniques and interpret the results.

The EPT approach is the least desirable in terms of defensibility because it results in a somewhat circular process. If no other approach is feasible, using EPT as the disturbance gradient could be considered a last resort. Still, there must be confidence in how well EPT values represent the full range of conditions occurring in the region. In addition, there must be a rationale for defining the disturbance gradient using EPT for the specific region of interest. Although this approach appears to be the simplest and most straightforward one, it has many drawbacks and limitations in practice.

No approach described in this report can be selected and carried out blindly. All require careful evaluation of the data available and the statistical techniques involved. The data set to be used in developing TVs is often the most limiting factor in terms of the choice of approach. Typically, data has already been collected, and the variables in that data

set may or may not include those that are necessary to carry out a particular approach. The statistical techniques necessary for a particular methodology can also limit the choices available. Not only must the user have access to and familiarity with the appropriate software package to run analyses, but he/she must also be able to understand and interpret the results obtained. If suitable attention is given to these issues, an approach to developing TVs can be identified that is defensible and appropriate

Table of Contents

List of Tables

List of Figures

List of Boxes

Acknowledgments

We thank Margaret Passmore, Greg Pond, and Louis Reynolds, all of U.S. EPA Region 3, Wheeling, WV, as well as Phil Larsen and Lester Yuan, U.S. EPA Office of Research and Development, for thorough reviews and valuable comments and suggestions on earlier drafts of this report. We appreciate the careful scrutiny of a late draft of this report by Thom Whittier, Oregon State University. The assistance of Alicia Shelton in assembling a database of tolerance values from numerous sources was invaluable to the success of this study. Her work was supported through contract 68D01048 with SoBran, Inc, c/o U.S. EPA, Cincinnati, OH. The participation in this work by Lori Winters was supported under a fellowship with the Oak Ridge Institute for Science and Education (ORISE) and partially funded through the Regionally Applied Research Effort (RARE) program.

1 Introduction

The concept of *tolerance* is a cornerstone of the field of bioassessment. Within this field, tolerance has traditionally meant the degree to which organisms can withstand environmental degradation. On a gradient of environmental impairment, very tolerant organisms are most common at the degraded end, whereas very intolerant organisms are expected to be most common at the pristine, natural end of the gradient. The terminology can be semantically confusing because *tolerance* means similar, yet different things in bioassessment and ecology. In ecology, a tolerant organism can withstand a wide range of environmental conditions. Tolerance refers to the breadth of the occurrence of an organism along an environmental gradient (Putman and Wratten 1984). Within bioassessment, it is the position on this scale at which the organism is most likely to occur that defines its tolerance (Johnson et al. 1993). Since the word *optimum* more accurately reflects the meaning of this idea and there is no potentially confusing homonym in ecology, the term is being furthered as a replacement for *tolerance* in bioassessment. However, for the purposes of this paper, we will assume that a taxon's tolerance value (TV) represents its optimal position on a gradient of disturbance rather than the breadth of its occurrence.

The use of such information on the tolerances of individual taxa to determine water quality is not a recent concept. The Saprobien System of the early 1900's is the first documentation of an empirical approach that evaluated the condition of a water body by the resident assemblages, incorporating assemblages from algae to fish (Kolkwitz & Marsson 1909). Since then, the concept has evolved and diversified to eventually become the basis of many indispensable tools of modern bioassessment. In 1972, Chutter developed an index in South Africa that assigned values to species on a scale of 0-10. A zero was given to those species in the cleanest streams, while a 10 was given to species found in the most polluted streams. The value assigned to a species was then multiplied by the number of individuals in that taxon found in a stream. The product was summed across all taxa, and the sum was then divided by the total number of individuals in the stream. Hilsenhoff adopted Chutter's approach to create the Hilsenhoff Biotic Index (HBI) for North American streams but subjectively assigned index values on a scale of 0-5 based on "previous experience and knowledge" and then adjusted the values when the HBI score didn't correlate well with physical and chemical parameters (Hilsenhoff 1977). Later, Hilsenhoff (1982) used data collected from over 1000 Wisconsin streams to add new tolerance values for additional species and to refine the tolerance values developed in Hilsenhoff (1977) to a 0-10 scale. Furthermore, Hilsenhoff coined the common term "tolerance value" to quantitatively represent a taxon's tolerance. In the last 25 years, the HBI and other similar indices have become a staple of stream biotic evaluation (Flotemersch et al. 2001). The HBI has been incorporated into the U.S. Environmental Protection Agency's Rapid Bioassessment Protocols, and has become ubiquitous in national, state and regional monitoring programs and multi-metric indices used to assess the integrity of streams and lakes (Barbour et al. 1999, Lewis et al 2001, Blocksom et al. 2002).

Due to the widespread use of the HBI in bioassessment, it is important to examine the components that define the Index. With some minor variations, the HBI is generally calculated by summing the product of the proportion of individuals of each taxon in a sample by its assigned pollution tolerance value. Although some contend that the HBI is defined by the tolerance values originally developed by Hilsenhoff, similar indices derived using region-specific TVs are also commonly referred to as the HBI. For the purposes of this report, we use the term *HBI* to represent a modified HBI in which the general formula of the original HBI is applied to any set of TVs. The accuracy of this index relies on the tolerance values assigned to each taxon. Although many methods exist for independently deriving tolerance values (e.g., best professional judgment or literature research of life histories) many regulatory agencies adopt or modify values from a few sources of published tolerance values such as Hilsenhoff (1977, 1982, 1987, 1988a), Green (1990), Lenat (1993) and Bode (1996), or they rely on those values that are in use by another agency. Such practices for developing lists of tolerance values are widely used and accepted, although many concerns exist.

The tolerance value of any particular taxon may vary depending on the list consulted and may even incorporate completely different scales to score pollution tolerance. Tolerance values have been shown to return HBI scores that correlate well with other measures of stream quality (Klemm et al. 2002), but the element of subjectivity is a weakness that subjects the HBI to scrutiny and liability. Also, many values are used for purposes other than their original published purpose. Hilsenhoff's values (1977, 1982, 1987), which were created for identifying organic pollution in Wisconsin, are used widely throughout the country to identify general disturbance regardless of source. Even when regional refinements are made, the tolerance values may range widely within a region, or may differ greatly depending on the list from which the value was originally derived. For example, recent respective lists of genus-level TVs used in the states of Delaware, Maryland, Vermont, Massachusetts, and in EPA's Environmental Monitoring and Assessment Program (EMAP) Mid-Atlantic Highlands Assessment (MAHA) differ by as many as 8 points for the same taxon (Table 1). Of these lists of TVs, about 40% of genera and families have values that differ by at least two points across lists.

Table 1. Genus-level tolerance values from several lists for taxa with a difference across lists of 2 points or greater. Only genera occurring in all 5 lists are included.

Genus	Delaware	Maryland	Vermont	Massachusetts	EMAP-MAHA
Ablabesmyia	7	8	8	8	5.3
Acroneuria	0	0	0	0	2.7
Agapetus	0	2	0	0	3.0
Ameletus	0	0	0	0	3.7
Antocha	3	5	3	3	3.7
Argia	6	8	6	6	5.0
Atherix	2	2	2	4	3.7
Baetis	6	6	6	6	2.7
Brachycentrus	1	1	1	1	4.5
Brillia	5	5	5	5	3.0
Caenis	7	7	7	6	3.8
Callibaetis	9	9	9	9	4.0

Chrysops	7	7	6	5	6.0
Cladotanytarsus	7	7	6	5	5.5
Clinotanypus	8	8	8	8	5.0
Corynoneura	7	7	4	4	6.7
Crangonyx	4	4	8	6	6.0
Cryptochironomus	8	8	8	8	6.0
Cryptotendipes	6	8	6	6	5.2
Culicoides	10	10	10	10	7.0
Dicrotendipes	8	10	8	8	6.7
Diplocladius	7	7	8	8	3.7
Dixa	1	4	1	1	6.0
Dolophilodes	0	0	0	0	3.3
Endochironomus	10	10	10	10	6.0
Ephemerella	1	2	4	1	2.7
Eukiefferiella	8	8	6	6	5.0
Eurylophella	4	4	2	2	3.3
Glossosoma	0	0	0	0	3.3
Glyptotendipes	10	10	10	10	6.7
Hexatoma	2	4	2	2	5.3
Hydrobaenus	8	8	8	8	4.7
Hydropsyche	4	6	5	4	4.3
Kiefferulus	10	10	10	10	5.3
Lepidostoma	1	3	1	1	3.0
Leucrocuta	1	1	1	1	3.0
Leuctra	0	0	0	0	2.3
Micropsectra	7	7	6	7	5.0
Molanna	6	6	6	6	3.0
Nanocladius	3	3	3	7	3.5
Natarsia	8	8	8	8	5.3
Nigronia	2	0	0	0	3.3
Oecetis	8	8	4	5	4.3
Orthocladius	6	6	7	6	5.0
Parachaetocladius	2	2	2	2	4.3
Paratendipes	8	8	8	6	4.7
Phaenopsectra	7	7	7	7	4.5
Pisidium	8	8	8	6	8.0
Potthastia	2	2	2	2	4.0
Procladius	9	9	9	9	6.3
Prosimulium	2	7	2	2	5.3
Psectrocladius	8	8	8	8	5.7
Rheotanytarsus	6	6	6	6	4.0
Rhyacophila	1	1	1	1	3.7
Sialis	4	4	6	4	7.0
Simulium	6	7	5	5	5.6
Somatochlora	1	1	1	9	4.3
Sphaerium	8	8	8	6	8.0
Stempellinella	4	4	4	2	4.7
Stenacron	4	4	7	7	4.0
Sympotthastia	2	2	3	2	4.7
Tanypus	10	10	10	10	5.0

Tipula	4	4	6	6	5.7
Tribelos	5	5	5	7	4.3
Zavrelimyia	8	8	8	8	5.0

There is also evidence showing that seasonality affects HBI scores and that seasonal adjustments are made in some cases (Hilsenhoff 1987, Lenat 1993). This occurs, in part, from seasonal shifts in community structure due to phenological events in taxa life histories. HBI scores that change with season may also reflect changes in environmental stress stemming from seasonal factors such as stream flow and water temperature. For example, the spring season generally affords macroinvertebrates with cooler and more oxygenated water conditions and greater dilution of pollutants. By contrast, the summer low-flow period often concentrates pollutants while higher stream temperatures tend to cause oxygen stress and harsher overall conditions for macroinvertebrates. Both Lenat (1993) and Hilsenhoff (1988b) found that indices based on tolerance values may be higher during summer months, and both used index-level correction factors to adjust for this difference among seasons. Furthermore, the notion that a taxon's sensitivity to stress may vary with these seasonal environmental factors could also confound our perception of a taxon's tolerance. Agency monitoring programs that sample communities throughout the year should evaluate whether biotic index scores vary significantly by season. In this report, we do not explicitly examine differences in TVs due to seasonal effects, although this may be an important factor in our comparisons among lists of TVs.

The purpose of this report is twofold. The first objective is to compare and characterize differences among various lists of TVs for macroinvertebrates currently in use. Our comparison is limited to those TVs intended to represent tolerance to a general disturbance gradient. As an extension of this analysis, we will compare the potential effects of the differences among lists when used in a biotic index, such as the HBI, for bioassessment. The second objective of this report is to describe and evaluate several methods that have been developed for empirically deriving TVs. The pros and cons of each method will be presented, with guidance on selecting an appropriate method for developing tolerance values for a particular region or state.

2 Comparison of tolerance values and HBI scores

2.1 Methods

For the purposes of making a useful comparison across several lists of tolerance values, we selected a subset of five available lists. Four of the lists are used at the state level in Delaware (DE), Kentucky (KY), Maryland (MD), and Massachusetts (MA). The fifth list is one maintained by a USEPA lab (EPA) that has processed large numbers of macroinvertebrate samples, primarily from the eastern United States. The TVs in the EPA list were based largely on available literature and best professional judgment. The various lists were compiled into a single list including all taxa found in

any of the five lists. Tolerance values were examined separately for family and genus taxonomic levels, and only TVs already at the family or genus level were used in analyses. No calculation or estimation of TVs was made based on those available at higher levels of taxonomic resolution (e.g., species level). Across the five TV lists, there were a total of 119 family-level values and 467 genus-level values.

To compare lists at each taxonomic level (family and genus), we performed an unbalanced repeated measures ANOVA with Tukey pairwise comparisons of least squares means, using TV list source (i.e., DE, KY, MD, MA, EPA) as the factor. Each pairwise comparison was based on all of the taxa occurring in both lists of the pair, resulting in a different number of observations for each pairwise comparison. To ensure approximate normality and equal variance in the data, we examined plots of residuals, including histograms, normal probability plots, and scatter plots. Lacking a consistent directional difference, we would not expect a significant difference between TV lists. To examine the variability in the relationship between TV lists, we ran Pearson correlations and created scatter plots of values from all possible pairs of lists.

After the comparison of tolerance values, we calculated HBI scores for two sets of samples in the West Virginia Department of Environmental Protection (WVDEP) wadeable stream macroinvertebrate assemblage database. In the past, WVDEP identified macroinvertebrates only to family level or above. However, more recent samples have been identified to the genus level. We used one set of samples to reflect family-level data and another data set to reflect genus-level data. For family level data, we used all benthic samples collected between April and October during 1997-1999. For genus level data, we used samples collected during the same months of 2002-2003. In each case, we used only taxa that occurred in all five lists to calculate HBI scores. Although this means the HBI scores calculated do not reflect true sample or site condition, we are comparing a consistent sample across the five TV lists and getting a more accurate representation of the differences among lists. For family level analyses, there were 51 of a total of 117 taxa matching across the five lists over 498 samples. In genus level analyses, in which family level observations were included, there were 127 of a total of 251 taxa matching across lists and 134 samples. We excluded observations for which taxonomic resolution was less than family from HBI calculations. We then ran a repeated measures ANOVA with Tukey pairwise comparisons of least squares means, again with list source as the factor. We examined residuals for large deviations from assumptions of normality and homogeneous variance. Finally, variability between HBI scores among list sources was examined using Pearson correlations and scatter plots of HBI scores from all possible pairs of lists.

2.2 Results

Although the overall test for differences among TV lists was marginally significant ($F=2.27$, $df=4,342$, $p=0.0617$), there were no consistent differences among TV lists at the family level (Table 2). At the genus level, there were highly significant differences

among TV lists (F=6.92, df=4, 1145, p<0.0001), but only the Kentucky list differed significantly from the other lists (Table 3). At the family level, all pairs of lists were highly significantly correlated at 0.70 or above, but the lists for DE, MD, and MA were correlated at about 0.90 and above (Table 2). At the genus level, correlations were again all highly significant, but correlations were above 0.70 only among the DE, MD, and MA lists (Table 3). When tolerance values in different lists were plotted against one another, much more variability was evident (Figures 1 and 2). At the genus level, the variation in TVs from one list to another is even more evident. At both the genus and family levels, the EPA TVs have a tendency toward more central values that any other set of TVs, resulting in higher EPA values at the low end of other lists and lower values at the high end of other lists (Figures 1 and 2). The striations evident in some of the plots of genus level TVs are created by the pairing of the KY or EPA TVs, which have values based on 0.1 increments, with TVs in one of the three other lists, which are limited to integer values.

Table 2. Family level results for multiple comparisons of least squares means of five different tolerance value lists. Tests are based on 342 degrees of freedom.

Difference	Difference Estimate	Standard Error	t-statistic	Adjusted p-value	Pearson r (N)
DE-KY	-0.881	0.404	-2.18	0.189	0.776 (62)
DE-MD	-0.232	0.414	-0.56	0.980	0.900 (61)
DE-MA	0.018	0.424	0.04	1.000	0.942 (57)
DE-EPA	-0.652	0.413	-1.58	0.512	0.767 (62)
KY-MD	0.648	0.357	1.82	0.366	0.737 (94)
KY-MA	0.898	0.369	2.44	0.108	0.837 (85)
KY-EPA	0.228	0.356	0.64	0.968	0.745 (95)
MD-MA	-0.250	0.380	-0.66	0.965	0.903 (82)
MD-EPA	-0.420	0.368	-1.14	0.784	0.721 (85)
MA-EPA	-0.670	0.379	-1.77	0.394	0.801 (74)

Table 3. Genus level results for multiple comparisons of least squares means of five different tolerance value lists. Tests are based on 1145 degrees of freedom.

Difference	Difference Estimate	Standard Error	t-statistic	Adjusted p-value	Pearson correlation, r
DE-KY	-0.680	0.203	-3.35	0.008	0.603 (199)
DE-MD	-0.033	0.225	-0.14	1.000	0.940 (175)
DE-MA	0.198	0.214	0.93	0.887	0.875 (192)
DE-EPA	-0.001	0.199	0.00	1.000	0.620 (221)
KY-MD	0.648	0.199	3.25	0.010	0.623 (213)
KY-MA	0.879	0.187	4.70	<0.001	0.710 (253)
KY-EPA	0.679	0.170	4.01	<0.001	0.695 (378)
MD-MA	-0.231	0.211	-1.10	0.808	0.873 (195)
MD-EPA	0.032	0.195	0.16	1.000	0.639 (224)
MA-EPA	-0.199	0.183	-1.09	0.811	0.680 (289)

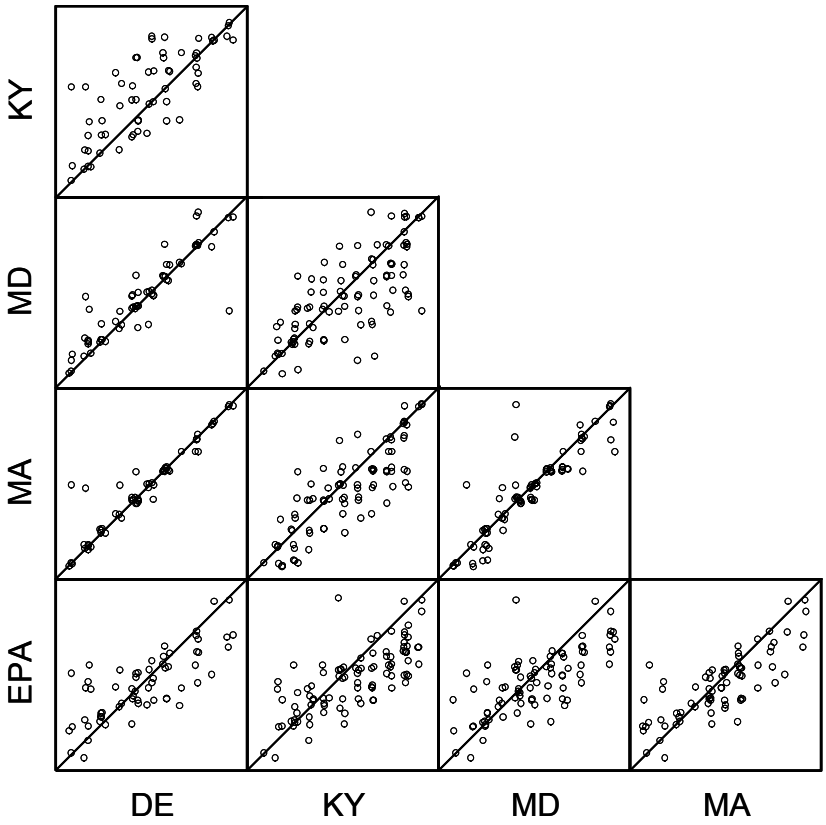


Figure 1 . Comparison of family level tolerance values from five different sources. Random jitter was used in plots so that overlapping points could be seen. The diagonal line represents a one-to-one relationship between TV lists.

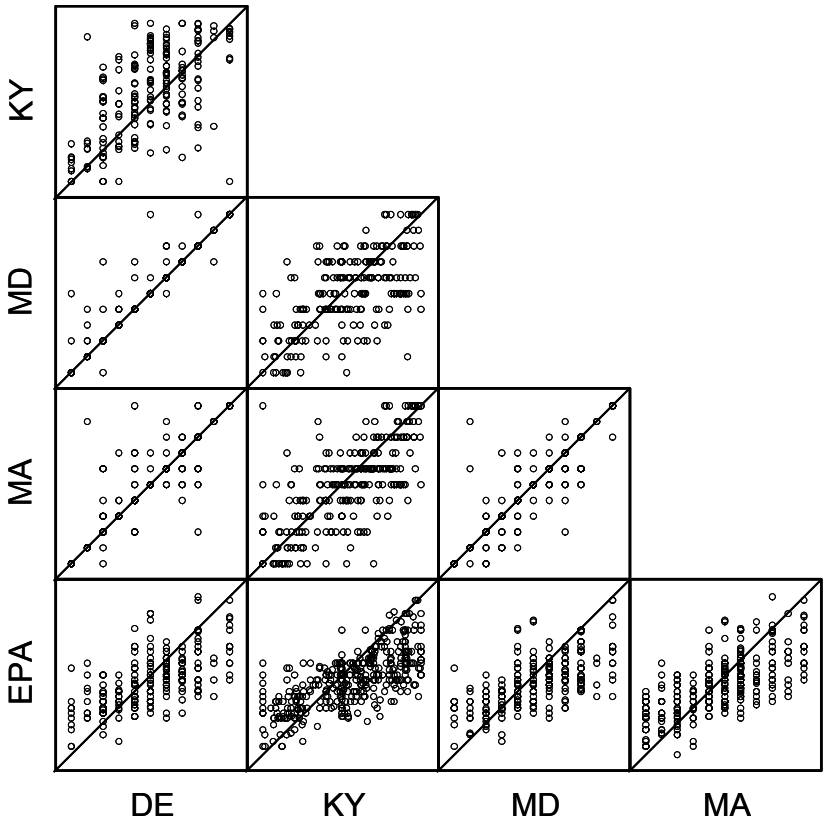


Figure 2. Comparison of genus level tolerance values from five different sources. Random jitter was used in plots so that overlapping points could be seen. The diagonal line represents a one-to-one relationship between TV lists.

HBI scores showed strong differences among most family-level TV lists when applied over a consistent set of taxa to a set of samples ($F=69.15$, $df=4,1988$, $p<0.0001$, Table 4). In addition, the relationships of HBI scores based on EPA TVs with those based on other lists were the most variable (Figure 3), and correlations among HBI scores based on the EPA TV list had the lowest correlations with those based on other lists (Table 4). Based on genus and family level TVs combined, HBI scores did not differ significantly among lists ($F=1.71$, $df=4,532$, $p=0.1469$), and none of the pairwise differences were significant (Table 5). However, HBI scores based on the EPA TV list again varied widely in their relationship to HBI scores based on other lists (Figure 4). Likewise, correlations with other HBI scores were always lowest for those based on the EPA TV list (Table 4).

Table 4. Family level results for multiple comparisons of least squares means of HBI scores based on taxa occurring in all five TV lists.

Difference	Difference Estimate	Standard Error	t-statistics	Adjusted p-value	Pearson r (N=498)
DE-KY	-0.431	0.048	-8.93	<0.001	0.728
DE-MD	-0.479	0.048	-9.94	<0.001	0.898
DE-MA	-0.060	0.048	-1.24	0.730	0.848
DE-EPA	-0.657	0.048	-13.63	<0.001	0.489
KY-MD	-0.048	0.048	-1.00	0.853	0.756
KY-MA	0.371	0.048	7.70	<0.001	0.863
KY-EPA	-0.227	0.048	-4.70	<0.001	0.492
MD-MA	-0.420	0.048	-8.70	<0.001	0.830
MD-EPA	-0.178	0.048	-3.70	0.002	0.565
MA-EPA	-0.598	0.048	-12.40	<0.001	0.608

Table 5. Genus level results for multiple comparisons of least squares means of HBI scores based on taxa occurring in all five TV lists.

Difference	Difference Estimate	Standard Error	t-statistics	Adjusted p-value	Pearson r (N=134)
DE-KY	-0.172	0.151	-1.14	0.785	0.895
DE-MD	-0.256	0.151	-1.70	0.437	0.987
DE-MA	0.002	0.151	0.01	1.000	0.992
DE-EPA	-0.294	0.151	-1.95	0.295	0.709
KY-MD	-0.084	0.151	-0.56	0.981	0.884
KY-MA	0.174	0.151	1.15	0.777	0.913
KY-EPA	-0.122	0.151	-0.80	0.929	0.788
MD-MA	-0.258	0.151	-1.71	0.429	0.977
MD-EPA	-0.038	0.151	-0.25	0.999	0.720
MA-EPA	-0.296	0.151	-1.96	0.288	0.726

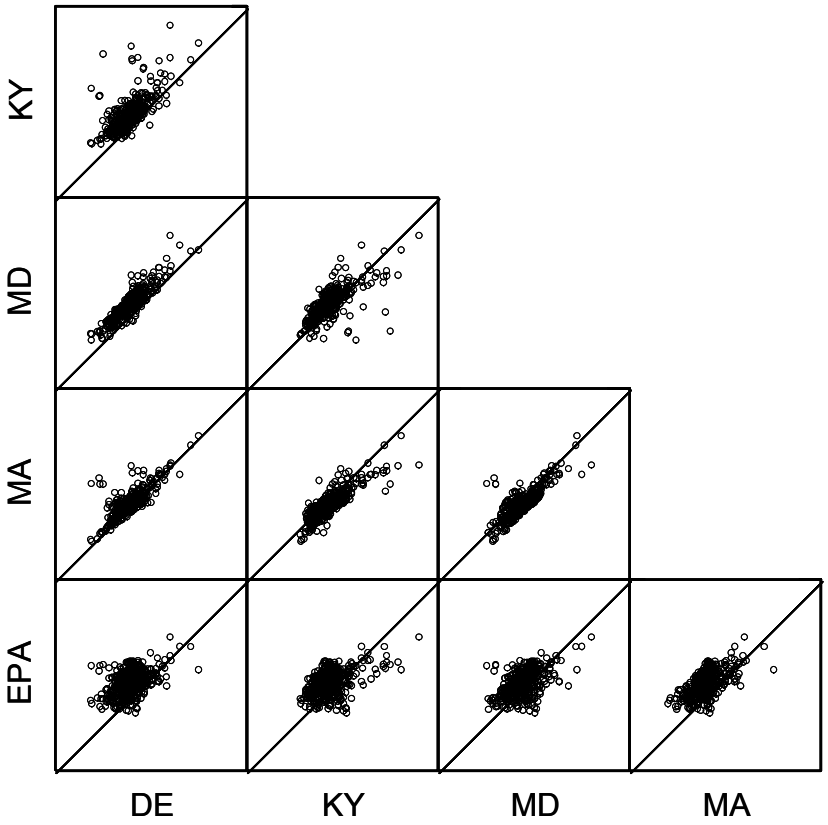


Figure 3 . Comparison of family level HBI scores based on five different TV lists. Random jitter was used in plots so that overlapping points could be seen. The diagonal line represents a one-to-one relationship between HBI scores.

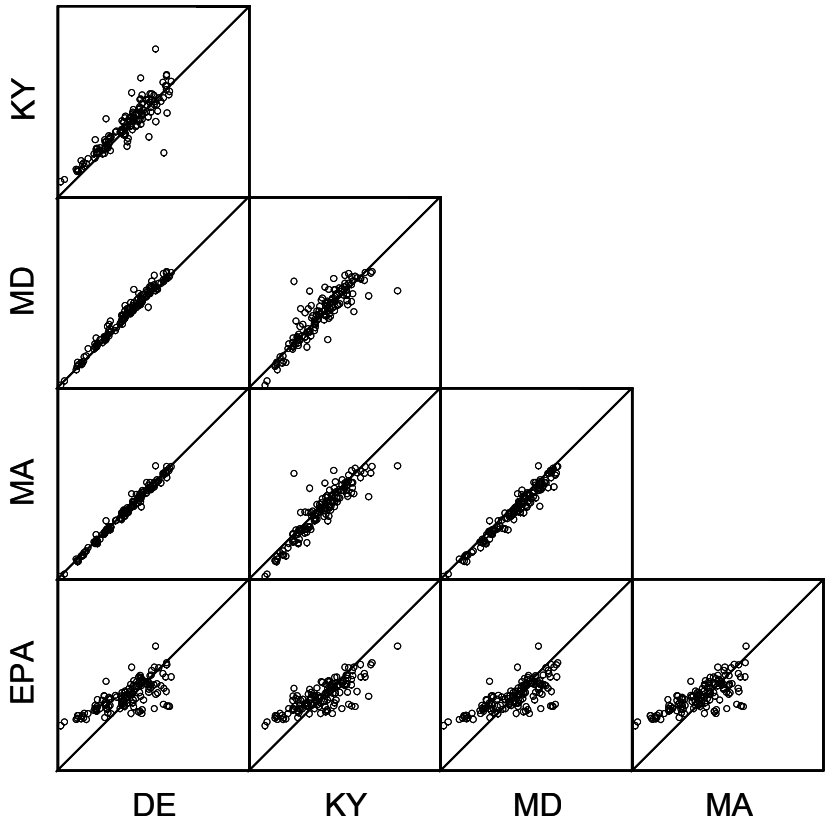


Figure 4 . Comparison of genus level HBI scores based on five different TV lists. Random jitter was used in plots so that overlapping points could be seen. The diagonal line represents a one-to-one relationship between HBI scores.

2.3 Discussion

These results show that tolerance values for a given taxon do vary among different TV lists, and the effect on HBI scores may or may not be important. They show that, in spite of the smaller differences in TVs among lists at the family level, differences in HBI scores among lists were larger at the family level. This may be the result of fewer taxa at the family level, such that small differences in the TVs for common taxa may be extrapolated over a larger proportion of the sample and lead to larger differences in HBI scores. At the genus level, individuals are distributed among many more taxa, and differences in TVs may not result in such large HBI differences. The slight compression of family level TVs for the EPA list (1 to 9.8) relative to other lists (0 to 10) may have contributed slightly to the further compression of HBI scores (3.0 to 7.0) compared to those from other lists (1.8 to 9.0). More likely, the EPA TVs for the more common and abundant taxa were compressed around the middle of the range of values, causing HBI scores to also be concentrated around the middle of the range. A similar pattern of compression of EPA TVs and HBI scores occurred at the genus level as well. Because the EPA TVs were derived mainly from literature and best professional judgment, there may have been a tendency to assign new tolerance values away from the extremes and more towards the middle of the TV range.

The fact that HBI scores may not be affected by differences in TVs is somewhat counterintuitive. We realize that the list of TVs used to calculate the HBI may be highly variable among states or regions. This may be due to differences in the component taxa that comprise that genus or family in different geographic areas or the season in which the data were collected (i.e., taxa TVs may be influenced by the time of year they are collected). This variation among TV lists is also likely due at least in part to the original combination of sources used to compile TVs for a particular agency or organization. However, we also recognize that the HBI is often highly discriminating between reference and impaired sets of sites. It would seem that the differences at the individual taxon level are cancelled out when combined across all taxa at a site and across all sites in a data set. Still, given the wide variability of TVs across lists, our confidence in the tolerance value assigned to a particular taxon is low, even though we have relatively high confidence in the HBI that incorporates that taxon TV. Thus, the development and use of a repeatable, transparent method of deriving TVs is highly desirable because it increases the confidence in the TVs of individual taxa.

Currently, there has been an effort associated with the National Wadeable Streams Assessment (WSA) to create a single list of TVs to apply to streams nationwide. This approach uses an algorithm to calculate a value based on lists from around the U.S. and fills in gaps in data as needed (Michael T. Barbour, Tetra Tech, Inc., personal communication).

3 Comparison of Methods to Derive Tolerance Values

Several approaches have been developed to generate tolerance values for bioassessment. In many cases, TVs have been developed using professional judgment. However, in this section, four approaches which are largely objective are compared for repeatability. The first two examined are really ways to define the disturbance gradient and include two different ways of calculating TVs based on those disturbance gradients. The third method is an approach to calculating TVs but depends on a disturbance gradient that has already been defined. Finally, the fourth method is an approach to generating TVs that does not rely on directly defining the disturbance gradient and includes two procedures for calculating TVs. Thus, this report does not contend to explore all possible combinations of ways to define the disturbance gradient and calculate tolerance values but rather a subset that represents the predominant approaches being proposed currently. The manner of defining the disturbance gradient and procedures to calculate TVs that are included in this report are provided in Table 6.

Table 6. A summary of each approach evaluated in this report with respect to defining the disturbance gradient and calculating tolerance values.

Approach	Disturbance gradient		Calculation of tolerance value for given taxon
	Explicitly defined?	How defined?	
Ephemeroptera, Plecoptera, Trichoptera (EPT)	Yes	EPT richness (representing underlying gradient)	1) 75th percentile of EPT richness based on proportion at each EPT richness value 2) Weighted average based on proportion at each EPT richness value
Principal components analysis (PCA)	Yes	PCA axis	1) 75th percentile of PC axis based on proportion at each axis value 2) Weighted average based on proportion at each axis value
Predictive modeling	No	Conditions in non-reference sites relative to reference sites	Frequency observed across sites relative to expected frequency
Generalized Additive Models (GAM)	Yes	PCA axis	1) Maximum probability of occurrence along PC axis 2) Weighted average of probability along PC axis

3.1 Data Sets

A single data set consisting of macroinvertebrate assemblage samples from the EMAP-MAHA study of wadeable streams in the mid-Atlantic region of the U.S. was divided randomly into calibration and validation data sets of 256 observations each. The data sets were each used to develop tolerance values for the same set of taxa, and correlations between the two sets of values were used to measure the repeatability of each method. Although both riffle and pool macroinvertebrate samples were collected at most sites, only riffle samples were used for analysis. In addition to biological data, physical habitat (using Rapid Bioassessment Protocols approach (Plafkin et al. 1989)) and water chemistry data were also collected at each site. Each approach to developing TVs was carried out on family- and genus-level data. The number of taxa for which TVs could be developed for a given method was dependent on criteria specific to that method, resulting in differing numbers of TVs for each method.

To evaluate the repeatability of a given method, the TVs generated from the calibration and validation data sets were compared using a paired t-test and the Pearson correlation. In addition, tolerance-related metrics were calculated for each sample in the combined data set based on TVs generated from the two data sets and for all methods. Thus, the influence of method on final metric values could be assessed as well.

3.2 Methods

3.2.1 EPT Approach

First, a simple approach based on the richness of Ephemeroptera, Plecoptera, and Trichoptera (EPT) taxa was used to develop TVs. In this approach, which is based on that of Lenat (1993), EPT richness is treated as a surrogate for the disturbance gradient. First, the proportion of each sample represented by each taxon and the number of samples in which each taxon occurred were calculated. Within each data set (calibration and validation), tolerance values were only estimated for those taxa with at least 25 observations. For each site in the data set, the total richness of EPT taxa was calculated using all distinct taxa identified to the lowest possible taxonomic level in these orders. For each value of EPT richness observed, the average proportion of individuals represented by a given taxon was calculated across all samples with that EPT richness. The average proportion values were used as weights for the cumulative distribution of EPT richness.

Lenat (1993) found that the 75th percentile produced the greatest separation of intolerant and tolerant species. Thus, the 75th percentile values generated in the manner described above were identified and then rescaled to a 0-10 range, with 10 as most tolerant and 0 as least tolerant, using the formula:

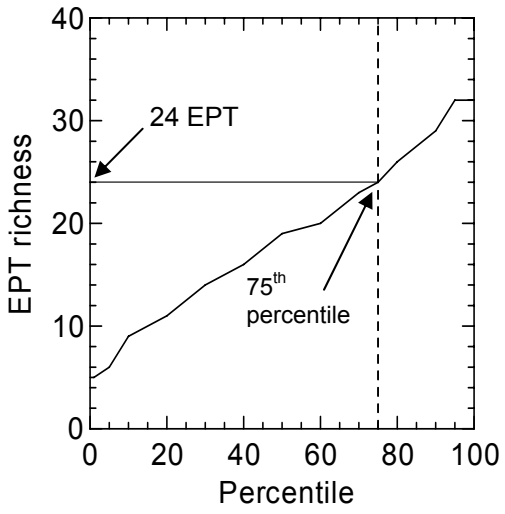
$$TV_{final} = \frac{TV_{\max} - TV_{init}}{TV_{\max} - TV_{\min}} \times 10 \qquad \text{(Equation 1.1)}$$

where TV_{init} is the initial TV of a taxon, calculated in this approach as the 75th percentile of the EPT cumulative distribution, TV_{min} is the minimum TV_{init} value across all taxa, and TV_{max} is the maximum TV_{init} value across all taxa. In other approaches, TV_{min} and TV_{max} are the minimum and maximum TV_{init} values across all taxa based on the particular disturbance gradient used. An example is provided in Box 1.

Box 1. Example of EPT approach with 75th percentile procedure for Perlidae.

Collected at 115 sites with 24 different values for EPT richness

EPT richness value	Mean proportion individuals
5	0.009
6	0.019
7	0.010
9	0.046
10	0.007
11	0.015
12	0.015
13	0.017
14	0.015
15	0.017
16	0.021
17	0.021
18	0.020
19	0.024
20	0.040
21	0.015
22	0.018
23	0.019
24	0.020
26	0.031
27	0.009
28	0.010
29	0.018
32	0.035



75^{th} percentile = 24 EPT taxa = TV_{init}

Across all taxa:
Maximum EPT-based TV_{init} = TV_{max} = 27
Minimum EPT-based TV_{init} = TV_{min} = 5

For Perlidae:
TV_{final} = (27-24)/(27-5)*10 = 1.4

A second procedure for developing TVs based on weighted EPT richness was applied to the data. The initial TV was calculated as:

$$TV_{initial} = \frac{\sum_i (proportion_i \times EPT_i)}{\sum_i proportion_i} \qquad \text{(Equation 1.2)}$$

where $proportion_i$ is the proportion of a given taxon in sample i and EPT_i is the EPT richness at that same site. Then, this initial TV was rescaled to a 10-point range using Equation 1.1. The entire procedure described above was repeated for the validation data set. An example of this procedure is provided in Box 2.

Box 2. Example of EPT approach using weighted procedure for the family Perlidae.

For Perlidae, the calculations were calculated as below (ordered by EPT richenss):

$$TV_{init} = \frac{\left[(5*0.014)+(5*0.004)+...+(29*0.028)+(32*0.035)\right]}{\left[0.014+0.004+...+0.028+0.035\right]} = \frac{46.30}{2.62} = 17.67$$

Across all taxa: TV_{max} = 20.84 TV_{min} = 5.38

For Perlidae: TV_{final} = (20.84 – 17.67)/(20.84 – 5.38)*10 = 2.1

3.2.2 Principal Components Analysis (PCA) Approach

The second approach to developing TVs used a principal components analysis (PCA) on physical habitat and water chemistry data to represent the disturbance gradient. This approach was used in developing TVs for the Mississippi Department of Environmental Quality (MDEQ, 2003). In an effort to include as many sites as possible in the analysis, only abiotic variables collected at most if not all sites were selected for the analysis. This meant that only the Rapid Bioassessment Protocols (RBP) habitat variables were included to represent habitat condition, as many other quantitative habitat variables were only collected at a subset of sites. The variables included in the PCA were conductivity, total phosphorus, total nitrogen, pH, RBP instream cover score, RBP embeddedness score, RBP bank condition score. All four water chemistry variables were transformed with $\log_{10}$ to reduce skewness. The PCA was run using a correlation matrix to account for the different units of the variables. A separate PCA was run for the calibration and validation data sets. All PC 1 scores were rounded to the nearest 0.1 for analyses.

Next, tolerance values were calculated for calibration and validation data sets separately using parallel procedures. As with the EPT approach, TVs were only developed for taxa with at least 25 observations in each data set, and for each taxon used, the proportion of each sample represented by that taxon was calculated. The remainder of the procedure is similar to that for the EPT approach. Tolerance values were calculated using both the 75th percentile and weighted procedures described above, replacing EPT richness with the rescaled PC 1 score. Because higher PC 1 scores were associated with poorer quality conditions for both the calibration and validation data sets, rescaling TVs to a 0-10 range did not require reversing the scale as with the EPT approach. Rescaling was carried out using the equation:

$$TV_{final} = \frac{TV_{init} - TV_{\min}}{TV_{\max} - TV_{\min}} \times 10$$

where TV_{init} was the initial TV calculated from PC 1 axis scores, TV_{max} and TV_{min} were the maximum and minimum TVs, respectively, across taxa. Again, the procedure

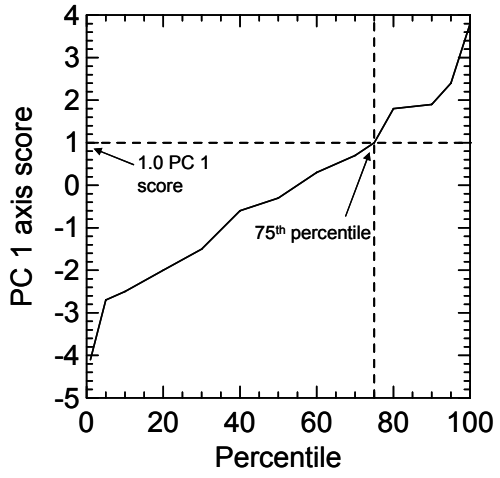
outlined above was rerun using the validation data set. An example of both procedures of this approach for the family Perlidae is provided in Box 3.

Box 3. Example calculation of TVs using the 75th percentile and weighted procedures for the PCA approach for Perlidae.

75th percentile procedure

139 samples with Perlidae

PC 1 axis scores	Mean proportion
-4.1	0.010
-3.4	0.004
-3.2	0.022
-3.0	0.008
-2.7	0.002
-2.5	0.042
-2.4	0.010
.	.
.	.
.	.
2.1	0.020
2.4	0.024
2.5	0.006
2.9	0.008
3.1	0.015
3.4	0.004
3.8	0.004



75th percentile = 1.0 PC 1 score = TV_{init}
Maximum PC 1 score = TV_{max} = 4.9
Minimum PC 1 score = TV_{min} = -0.4

TV_{final} = (1.0 – (-0.4))/(4.9 – (-0.4))*10 = 2.64

Weighted procedure

Calculation for Perlidae, ordered by rounded PC 1 score:

$$TV_{init} = \frac{\left[(-4.1 * 0.010) + (-3.4 * 0.004) + ... + (3.4 * 0.003) + (3.8 * 0.004)\right]}{\left[0.010 + 0.004 + ... + 0.003 + 0.004\right]} = -0.22$$

TV_{max} = 2.94 TV_{min} = -1.39

TV_{final} = (-0.22 - (-1.39))/(2.94 - (-1.39))*10 = 2.70

3.2.3 Predictive Modeling (O/E) Approach

The predictive modeling approach relies on the development of predictive models to determine the proportion of observed to expected taxa at a given site. These models have been largely in use in Great Britain (Moss et al. 1987, Wright 1995) and Australia (Turak et al. 1999), but they have been developed for parts of the U.S. as well (Hawkins et al. 2000). For this study, the procedures in Hawkins et al. (2000) were generally followed to develop predictive models from which TVs could be estimated.

The first step of the approach is to develop a predictive model, from which TVs for individual taxa can then be inferred. The model depends on the use of reference site data to determine the expected taxonomic composition under natural conditions. A subset of all sites in the mid-Atlantic data set were identified as reference based on water chemistry and habitat criteria developed previously (Klemm et al. 2003). Since there were only 88 reference sites across both the calibration and validation data sets, all reference sites were combined to develop a single predictive model. Then observations from the calibration and validation data sets were used independently to develop TVs.

To develop a predictive model, two major steps are required initially. First, cluster analysis divides sites into natural groupings based on similarity of the macroinvertebrate assemblages. For this step, only reference sites are used so that groupings are based on natural factors and not disturbance gradients. Following common practice for predictive models, clustering was performed on only presence-absence data for each taxon. To avoid clustering based on taxa that are only present in a few samples or those that are ubiquitous, taxa present in 5% or fewer or 95% or more of samples were excluded (Hawkins et al. 2000). Flexible beta clustering was performed in PC-ORD for Windows (version 4.25, MjM Software, Gleneden Beach, Oregon, http://home.centurytel.net/~mjm/) on the data with $\beta = -0.5$ and using the Sorensen distance measure on presence-absence data. The number of clusters used in the next step of analysis was based partly on the percent of information remaining and partly on the size of the smallest cluster.

In the second major step in developing a predictive model, discriminant function analysis (DFA) is used to determine the combination of abiotic variables that can separate sites into different groups most accurately. The abiotic variables were chosen to represent natural factors describing stream reaches that might affect macroinvertebrate composition. Only a limited number of abiotic variables were available at most sites in this data set. These variables included latitude, longitude, watershed area, elevation, approximate distance to the ocean, estimated annual runoff, Julian day, and estimated aspect (direction) of the longest dimension of the stream reach. A few variables were transformed to reduce skewness, including $\log_{10}$ transformation of watershed area and runoff and square root transformation of aspect. Only reference sites were used in the DFA, with the clusters defined in the previous step serving as the grouping variable. A stepwise discriminant analysis was performed

to develop discriminant functions to predict membership of test sites in each of the cluster groups. All DFA runs were performed in SAS (v. 9.1, SAS Institute Inc., Cary, North Carolina).

To determine the expected number of taxa at a particular test site, the probability of capture (P_c) of each taxon must first be calculated. First, for each taxon in the reference site data set, the proportion of sites at which that taxon is observed ($P_{refcluster}$) is calculated for each reference site cluster group. Next, the discriminant functions are used to calculate the probability of membership of the test site in each reference site cluster ($P_{cluster}$). The probability of capture (P_c) of a taxon in a test sample is the product of $P_{cluster}$ and $P_{refcluster}$ for each cluster, then summed across clusters. An example of the P_c is provided in Box 4.

Box 4. Example calculation of probability of capture (P_c) for Perlidae at site MD003.

Using reference site data:

- Three groups of sites were identified using cluster analysis.
- Discriminant function analysis (DFA) was used to develop a model to classify sites based on abiotic variables.
- The probabilities of occurrence of Perlidae among reference sites in clusters 1, 2, and 3 were calculated ($P_{refcluster}$).

Using data from site MD003 (a test site):

- The DFA model was used to predict the probabilities that the site belongs to clusters 1, 2, and 3 ($P_{cluster}$).
- The probability of capture was calculated as the sum of $P_{refcluster}$*$P_{cluster}$ across all clusters.

Cluster	$P_{refcluster}$	$P_{cluster}$	$P_{cluster}$*$P_{refcluster}$
1	0.583	0.404	0.236
2	0.704	0.233	0.164
3	0.826	0.363	0.300

$$P_c = \sum_{cluster} (P_{cluster} * P_{refcluster}) = 0.236 + 0.164 + 0.300 = 0.700$$

The final value calculated for a test site is the ratio of observed to expected taxa richness (O/E). The sum of P_c values across taxa determines the expected number of taxa (E) for that site. Only taxa having $P_c \geq 0.50$ were used in calculations of O and E, as limiting the calculation to these more common taxa has been shown to result in lower variability of O/E values at reference sites (Hawkins et al. 2000). The O/E values were calculated for all sites in the data set with the abiotic variables necessary to apply the discriminant functions.

The approach to estimating the responsiveness of each taxon as a sort of tolerance value was based on Hawkins (2004). For a given taxon, the P_c values are summed across all sites (reference and otherwise) to obtain the predicted number of sites at which that taxon is expected to occur (S_e). The number of sites at which the taxon is actually found is S_o. The ratio of S_o to S_e is an index of responsiveness to stress. All taxa found at reference sites were considered regardless of P_c, but responsiveness was only estimated for those taxa with a S_o or S_e of at least 15. A ratio of less than 1 indicates that the taxon decreases in response to stress and a ratio of greater than 1

indicates an increase in response to stress. By definition, TVs can only be calculated for taxa expected at reference sites, as other taxa lack P_c values. In order to compare them to TVs, these ratios were rescaled to a range of 0-10 using the same method as for the PCA approach. The calculation of the TV for Perlidae is again provided as an example in Box 5.

Box 5. Example calculation of TV for Perlidae using predictive modeling approach.

$S_e = \sum_{allsites} P_c = 196.4$ (expected sites with Perlidae) $S_o = 162$ (observed at 162 sites)

$TV_{init} = S_o/S_e = 162/196.4 = 0.825$ $TV_{min} = 0.448$ $TV_{max} = 8.337$

$TV_{final} = (0.825 – 0.448)/(8.337 – 0.448)*10 = 0.477$

3.2.4 Generalized Additive Model (GAM) Approach

Generalized Additive Models (GAMs) are a generalization of Generalized Linear Models (GLMs) in which some predictors are modeled nonparametrically along with linear and polynomial terms for other predictors (Guisan et al. 2002). The GAM approach allows for nonlinear relationships between responses and multiple predictors. Various types of smoothers can be applied in these models in order to more precisely approximate the relationship between a response and a particular predictor variable, and different types of relationships can be simultaneously described within a single model. For the purposes of developing tolerance values, the process involves first identifying a predictor variable that serves as a disturbance gradient, along with other variables representing natural factors that may influence relationships. Then, a GAM is used to describe the nature of the relationship between an individual taxon and the disturbance gradient. The uses and fitting of GAMs are covered in detail in Hastie and Tibshirani (1990).

The GAM approach followed for this study is based on work by Yuan (2004). However, Yuan (2004) was interested primarily in developing tolerance values for individual stressors, and this report is focused on developing TVs for a general disturbance gradient. Thus, the gradients of phosphorus and sulfate concentrations in Yuan (2004) are replaced here by the first principal component axis of a PCA based on abiotic variables. Only a definable abiotic gradient was of interest for this analysis, so the analysis was not run using an EPT gradient. The same calibration and validation first PC axis scores generated for the PCA approach were used as the gradients in this approach. As with the predictive modeling approach and in Yuan (2004), only presence-absence data were used in the analysis. The model for each taxon was generated in SAS using PROC GAM and a binomial distribution for the response variable (presence or absence: 1 or 0). The binomial nature of the data means that the response modeled was actually the logit of the probability of occurrence (i.e., $\ln(p/(1-p))$, where p is the probability of occurrence). As in Yuan (2004), additional variables were

included in each model as covariates, including latitude, watershed area (log-transformed), elevation, and estimated annual runoff.

In determining the form of the model for each taxon, Yuan (2004) recommends sample sizes which allow at least ten *observations* for each degree of freedom in the model. For most taxa, the term *observation* refers to samples containing the taxon. However, because both presence and absence of a taxon are required to model its probability of occurrence, for very common taxa, the term *observation* refers to samples from which the taxon is absent. For the purposes of this analysis, the minimum of the two values representing the number of sites with and without a taxon can be used as the number of *observations*. For each predictor included in the model, a smoothing (regression) spline with 90% confidence limits was specified with 2 degrees of freedom, such that the minimum degrees of freedom was 2 and the maximum was 10, if all four covariates were included. In order to maximize the number of models for taxa, we grouped taxa by the number of *observations* and developed more complex models for taxa with more *observations* (Yuan 2004). For taxa with 20 to 59 *observations*, only the first PC axis was included in the model. For taxa with 60 to 99 *observations*, the first PC axis, latitude, and watershed area were included in the model, and for taxa with 100 or more *observations*, the model included PC axis 1, latitude, watershed area, elevation, and annual runoff. The additional covariates only served to reduce the unexplained variability in the relationship between PC axis 1, the generalized stressor gradient, and the logit of taxon probability of occurrence, and are not of general interest here. This grouping of taxa was applied at the genus and family levels to both the calibration and validation data sets. At a given taxonomic level, the grouping with the fewest variables in the model was used for both the calibration and validation data sets in order to allow for comparisons of TVs generated from the two data sets.

Two important features were evaluated for each model. First, each model was evaluated for the significance of the relationship between taxon presence and the PC axis 1 scores at the α=0.05 level. In SAS, one degree of freedom is automatically removed to account for the linear portion of the model, and the output provides separate significance tests for linear and nonlinear components of each relationship. Next, the nature of the relationship was evaluated. In Yuan (2004), the significance of unimodality with the PC axis 1 was determined graphically as cases where the maximum predicted logit value was greater than the 90% confidence limits at the minimum and maximum PC axis 1 values. Because of generally smaller sample sizes than Yuan used, unimodality was considered significant if the maximum value for the upper 90% confidence limit, rather than the maximum predicted value itself, was greater than the upper 90% confidence limits on the maximum and minimum PC axis 1 values. If the linear relationship between PC axis 1 and the probability of occurrence was significant but unimodality was not, a monotonic relationship was assumed and the direction was determined using the linear slope estimate. If a taxon was significantly related to PC axis 1 nonlinearly but was not unimodal, TVs were still calculated for that taxon.

The goal of this work was to generate tolerance values, in contrast to Yuan (2004), for which the primary goal was to evaluate the ability of the GAM approach to determine the general type and direction of the relationship. Tolerance values were calculated as both the optimum and the weighted average, based on the predicted probability of occurrence. The predicted probability was calculated as the inverse of the predicted logit. The optimum value was determined as the PC axis 1 score at which the predicted probability of occurrence of that taxon was highest. The weighted average was calculated using presence and absence data from all sites as

$$TV_{wtd} = \frac{\sum p \times PC1}{\sum p}$$

where *PC1* is the PC axis 1 score for a given site and *p* is the predicted probability of occurrence at that value. Each TV was rescaled to a 0 to 10 range using the method described for the PCA approach. If no relationships were significant between taxon occurrence and PC axis 1, no TV was calculated for that taxon. Likewise, if the relationship was U-shaped, no TV was calculated. This process was repeated for the validation data set, but only taxa with significant relationships for calibration data were evaluated. Box 6 contains an example calculation of both the optimum and the weighted average for the GAM approach with Perlidae.

Box 6. Example calculation of TV for Perlidae using GAM optimum and weighted average approaches.

- Perlidae was observed in 139 of 256 samples.
- In a GAM that included PC 1 axis scores, log(watershed area), latitude, elevation, and annual runoff, the logit of the probability of occurrence was significantly negatively related to PC 1 axis scores.
- For each site in the analysis, a predicted probability of occurrence was determined using the inverse of the predicted logit value (i.e., ln[p/(1-p)]).
- **Optimum:** The maximum predicted probability was 0.818, and the PC 1 axis score corresponding to the site with this logit value was -3.409 (=TV_{init}).
  - TV_{max} = 4.825 TV_{min} = -4.146
  - **TV_{final} = (-3.409 – (-4.146))/(4.825 – (-4.146))*10 = 0.822**
- **Weighted average:** The sum of the product of the predicted probability X PC axis 1 scores for Perlidae was -45.224, and the sum of the predicted probabilities was 139.0, resulting in a TV_{init} of -0.325.
  - TV_{max} = 1.249 TV_{min} = -1.021
  - **TV_{final} = (-0.325 – (-1.021))/(1.249 – (-1.021))*10 = 3.063**

3.2.5 Comparison of Tolerance Metrics

Recognizing that the TVs generated eventually would be used to derive macroinvertebrate metrics, a comparison of three metrics -- HBI, number of intolerant taxa (TV ≤ 3), and percent intolerant individuals -- was performed. First, all sites in the validation and calibration data sets were combined. Then, data were aggregated to the family and genus levels. The three metrics were calculated for each site using TVs generated from each approach based on both the calibration and validation data sets. Only taxa having TVs for all of the methods were used in calculations to ensure an equitable comparison of metric values. At the very large sample sizes of over 500 sites, even small mean differences in metric values would be significant in simple paired t-tests because of the extremely large degrees of freedom. Thus, comparisons between the metrics calculated from calibration- and validation-based TVs were assessed qualitatively using bivariate scatter plots. Next, sites were identified that had been previously classified as reference or impaired using water chemistry and habitat variables (Klemm et al. 2003). Of interest were differences in metric values between reference and impaired sites, to determine which metric discriminates the two types of sites most strongly. Comparisons among approaches were made by examining the overlap of the interquartile ranges (25^{th} to 75^{th} percentiles) of box plots to determine the approach that best separates reference and impaired sites. For these plots, only metric values calculated from calibration-based TVs were included. All comparisons were carried out separately for genus- and family-level data.

3.3 Results

3.3.1 EPT Approach

The range of EPT richness observed for both the calibration and validation data sets was 0 to 32 taxa. The distributions of TVs from both the 75^{th} percentile and weighted procedures are provided in Table 7. The average TVs were higher for the weighted adjustment procedure than for the 75^{th} percentile procedure. When comparing the TVs created from the calibration and validation data sets, differences were significant regardless of taxonomic level or scoring procedure (Table 8), but correlations were also highly significant. Correlations between values at the genus level were lower than for the family level (Table 8). From scatter plots of the validation against the calibration TVs (Figure 5), there was more variability in the genus level data, but the plots did not differ obviously between the 75^{th} percentile and weighted procedures. The differences detected in the paired t-tests were evident but not strong in these bivariate plots.

Table 7. Comparison of distributions between the calibration and validation data sets for family and genus, as well as both the 75th percentile and weighted procedures. Values are based on the EPT approach.

		Calibration		Validation	
Taxonomic level	Scoring procedure	Mean	Standard deviation	Mean	Standard deviation
Family (N=44)	75th	3.39	2.34	3.93	2.21
	Weighted	4.29	2.40	3.82	2.38
Genus (N=116)	75th	3.79	1.95	4.65	1.75
	Weighted	3.94	1.92	4.45	1.94

Table 8. Differences and correlations between calibration and validation tolerance values at the family and genus levels for the 75th percentile and weighted procedures using the EPT approach. Values for *t* are from the paired t-test for differences.

Taxonomic level	Scoring procedure	Mean difference	t-statistic (p-value)	Pearson r (p-value)
Family (df=43)	75th	-0.53	-3.33 (0.0018)	0.893 (<0.0001)
	Weighted	0.47	3.53 (0.0010)	0.931 (<0.0001)
Genus (df=115)	75th	-0.86	-5.81 (<0.0001)	0.775 (<0.0001)
	Weighted	-0.51	-4.19 (<0.0001)	0.629 (<0.0001)

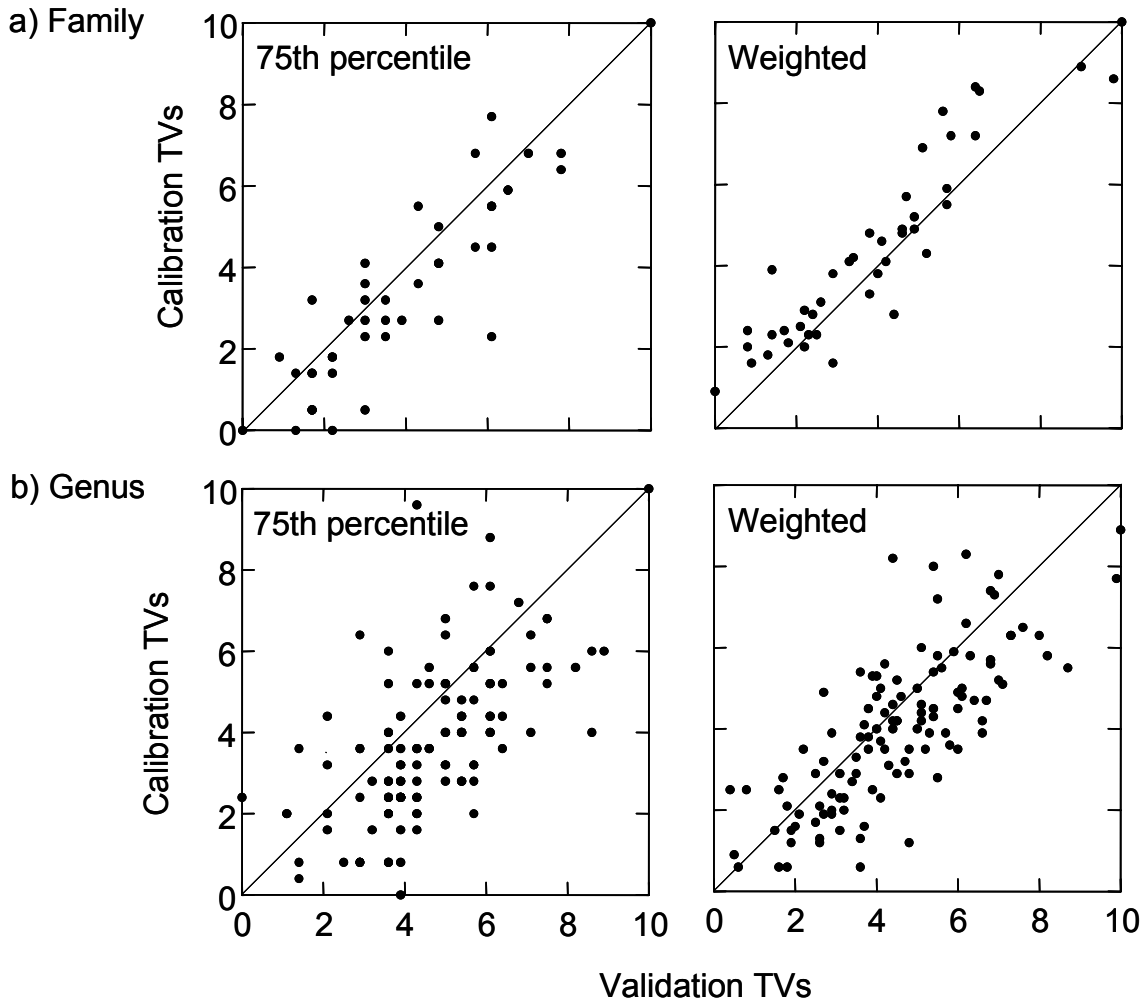


Figure 5. Calibration and validation tolerance values, matched by taxon for a) family and b) genus levels, based on the EPT approach using the 75th percentile and weighted procedures. The diagonal line represents the same TV for a taxon in both the validation and calibration data sets.

3.3.2 PCA Approach

The PCA resulted in two axes with eigenvalues larger than 1. The first axis of the calibration PCA explained about 41% of variation and the second about 20%. For the validation PCA, the first explained 41% and the second 18%. All of the variables were significantly correlated with the first principal component (PC1), but pH and the RBP bank condition score had weaker correlations than the other variables (Table 9). Because all variables were correlated with the first principal component, the first axis was used to represent a general disturbance gradient across sites.

Table 9. Spearman rank correlations between abiotic variables used in PCA and principal component axes.

	Calibration		Validation	
Variable	PC1	PC2	PC1	PC2
Conductivity	0.68	0.36	0.71	0.38
Total Phosphorus	0.69	0.32	0.60	0.33
Total Nitrogen	0.73	0.36	0.75	0.33
pH	0.45	0.49	0.52	0.40
RBP bank condition	-0.55	0.58	-0.58	0.37
RBP embeddedness	-0.68	0.46	-0.71	0.46
RBP instream cover	-0.68	0.48	-0.60	0.64

Average TVs for the PCA approach tended to have similar magnitudes and standard deviations to those for the EPT approach (Table 10). The weighted procedure resulted in slightly smaller mean TVs, regardless of taxonomic level. The difference between validation and calibration TVs was significant only for the genus level 75th percentile procedure, although the average difference was still less than 0.5 point (Table 11, Figure 6). Correlations between calibration and validation TVs were lower for the 75th percentile scoring procedure, regardless of taxonomic level (Table 11), and family level correlations were much lower than those seen for the EPT approach.

Table 10. Comparison of distributions between the calibration and validation data sets for family and genus, as well as both the 75th percentile and weighted procedures. Values are based on the PCA approach.

		Calibration		Validation	
Taxonomic level	Scoring procedure	Mean	Standard deviation	Mean	Standard deviation
Family (N=44)	75th	3.34	2.02	3.32	2.40
	Weighted	3.19	1.98	3.16	2.17
Genus (N=116)	75th	4.55	1.58	4.14	1.79
	Weighted	3.97	1.65	4.06	1.81

Table 11. Differences and correlations between calibration and validation tolerance values at the family and genus levels for the 75th percentile and weighted procedures using the PCA approach. Values for *t* are from the paired t-test for differences.

Taxonomic level	Scoring procedure	Mean difference	t-statistic (p-value)	Pearson r (p-value)
Family (df=43)	75th	0.03	0.11 (0.9119)	0.776 (<0.0001)
	Weighted	0.03	0.17 (0.8638)	0.808 (<0.0001)
Genus (df=115)	75th	0.40	3.04 (0.0029)	0.647 (<0.0001)
	Weighted	-0.09	-0.88 (0.3798)	0.789 (<0.0001)

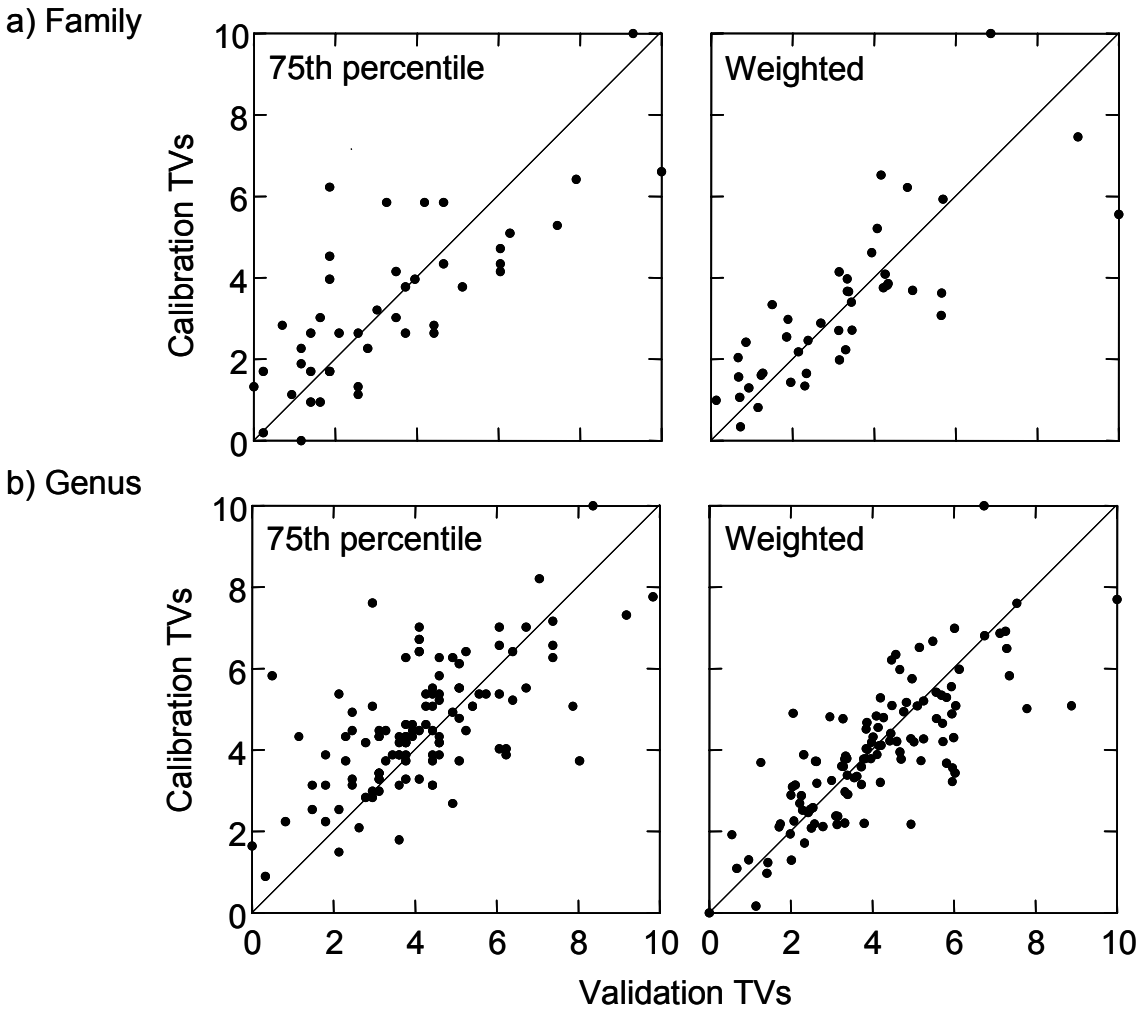


Figure 6. Validation and calibration TVs matched by taxon for a) family and b) genus levels, based on the PCA approach using the 75th percentile and weighted procedures. The diagonal line represents the same TV for a taxon in both the validation and calibration data sets.

3.3.3 Predictive Modeling Approach

The initial step of clustering sites was performed separately for genus- and family-level data. At the family level, 57 taxa were used in clustering, and at the genus level, 172 taxa were included in the cluster analysis. Initial clustering resulted in three sites forming very small groups. Upon further examination, these sites were determined to be in poor condition based on previous assessments of macroinvertebrates. Therefore, the reference data set was reduced by these three outliers to 85 sites for clustering at both the genus and family levels.

At the family level, three clusters left approximately 30% information remaining. The clusters consisted of 12, 27, and 46 sites. Watershed area, longitude, runoff, Julian day, and aspect were selected in stepwise DFA, and a pooled covariance matrix was used. The error rate of the model was 25% and the cross-validation error rate was 39%. The mean O/E score for reference sites was 1.01, with a standard deviation of 0.16. The mean very close to 1 indicates an unbiased estimate of the number of taxa expected to occur at a site, and the level of variation is on par with that found in Hawkins et al. (2000). There were 58 families in the calibration data set and 60 in the validation data set with S_e or S_o of at least 15, and 57 families occurred in both data sets.

At the genus level, four clusters resulted in approximately 30% information remaining. The clusters were made up of 13, 26, 29, and 17 sites. In stepwise DFA, watershed area, latitude, runoff, Julian day, and elevation were selected to best separate sites into clusters, and a pool covariance matrix was again used. The error rate of the model was 29%, with a cross-validation error rate of about 36%. The mean O/E score among reference sites was 1.04, with a standard deviation of 0.18, again indicating unbiased estimates and reasonable variation among reference sites (Hawkins et al. 2000). There were 173 taxa in the calibration data set and 179 in the validation data with S_e or S_o values of at least 15, with 171 taxa overlapping between the two data sets.

The average TVs based on the predictive modeling approach were much smaller than for the EPT or PCA approaches (Table 12). This skewing of values toward the low end is likely due at least in part to the limitation of analyses to taxa found at reference sites. Differences between the calibration and validation data sets were significant but small, and correlations were extremely high (Table 13). From the plots, it is obvious that although the relationships between the calibration and validation TVs were very strong, the slopes of those relationships were far from 1, such that larger values tended to be more different and smaller values tended to be more similar (Figure 7).

Table 12. Comparison of TV distributions between the calibration and validation data sets for family and genus levels, based on the predictive modeling approach.

	Calibration		Validation	
Taxonomic level	Mean	Standard deviation	Mean	Standard deviation
Family (N=57)	0.80	0.92	1.13	1.50
Genus (N=171)	2.08	1.64	1.29	0.98

Table 13. Differences and correlations between calibration and validation tolerance values at the family and genus levels using the predictive modeling approach. Values for *t* are from the paired t-test for differences.

Taxonomic level	Mean difference	t-statistic (p-value)	Pearson r (p-value)
Family (df=56)	-0.33	-3.84 (0.0003)	0.968 (<0.0001)
Genus (df=170)	0.79	12.77 (<0.0001)	0.932 (<0.0001)

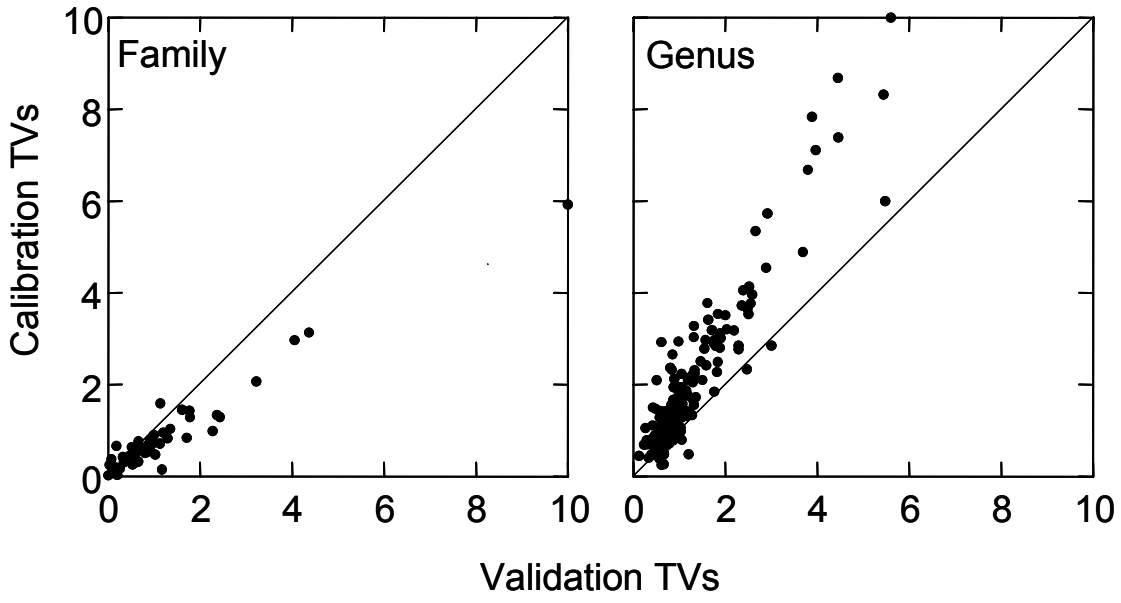


Figure 7. Validation and calibration TVs matched by taxon for family and genus levels, based on the predictive modeling approach. The diagonal line represents the same TV for a taxon in both the validation and calibration data sets.

3.3.4 Generalized Additive Model Approach

Only the family Chironomidae was excluded from analysis because of too many occurrences (i.e., too few *observations*). In the calibration and validation data sets, 68 and 69 families, respectively, had too few occurrences. Of the remaining families, 48 overlapped between the two data sets and were included in analyses. Only 26 taxa exhibited a significant relationship of some kind with the general stressor gradient (PC axis 1) for both the calibration and validation data sets. However, one of these taxa showed a U-shaped relationship with PC axis 1 for both data sets and was excluded from calculation of TVs. There were 35 families in the calibration data set (2 U-shaped) and 31 in the validation data set (1 U-shaped) that showed a relationship between the probability of occurrence (as a logit) and PC axis 1.

At the genus level, 350 (of 496) and 331 (of 476) taxa had too few observations in the calibration and validation data sets, respectively. Of those taxa remaining, 139 were observed in both data sets. Only 83 taxa in the calibration data set and 80 in the validation data set had significant relationships with PC axis 1. Of these, 63 taxa had significant relationships for both data sets.

The mean TVs for this approach were comparable to those of other approaches, but the standard deviations tended to be considerably larger (Table 14), indicating a broader distribution of values than for other approaches. Differences between the calibration and validation TVs were small and nonsignificant, regardless of scoring procedure and taxonomic level (Table 15). Values were highly correlated between the calibration and validation TVs (Table 15), but there was more variability in the relationship for optima than for weighted average (Figure 8).

Table 14. Comparison of TV distributions between the calibration and validation data sets for family and genus levels, based on the GAMs approach.

		Calibration		Validation	
Taxonomic level	Scoring procedure	Mean	Standard deviation	Mean	Standard deviation
Family (N=26)	Optimum	3.35	3.37	3.22	3.56
	Weighted average	3.74	3.13	3.40	2.69
Genus (N=63)	Optimum	4.79	3.58	4.42	3.96
	Weighted average	4.57	2.35	4.41	2.73

Table 15. Differences and correlations between calibration and validation tolerance values at the family and genus levels using the GAMs approach. Values for *t* are from the paired t-test for differences.

Taxonomic level	Scoring procedure	Mean difference	t-statistic (p-value)	Pearson r (p-value)
Family (df=24)	Optimum	0.12	0.37 (0.7142)	0.885 (<0.0001)
	Weighted average	0.34	1.50 (0.1460)	0.936 (<0.0001)
Genus (df=61)	Optimum	0.36	1.77 (0.0817)	0.911 (<0.0001)
	Weighted average	0.17	1.52 (0.1341)	0.952 (<0.0001)

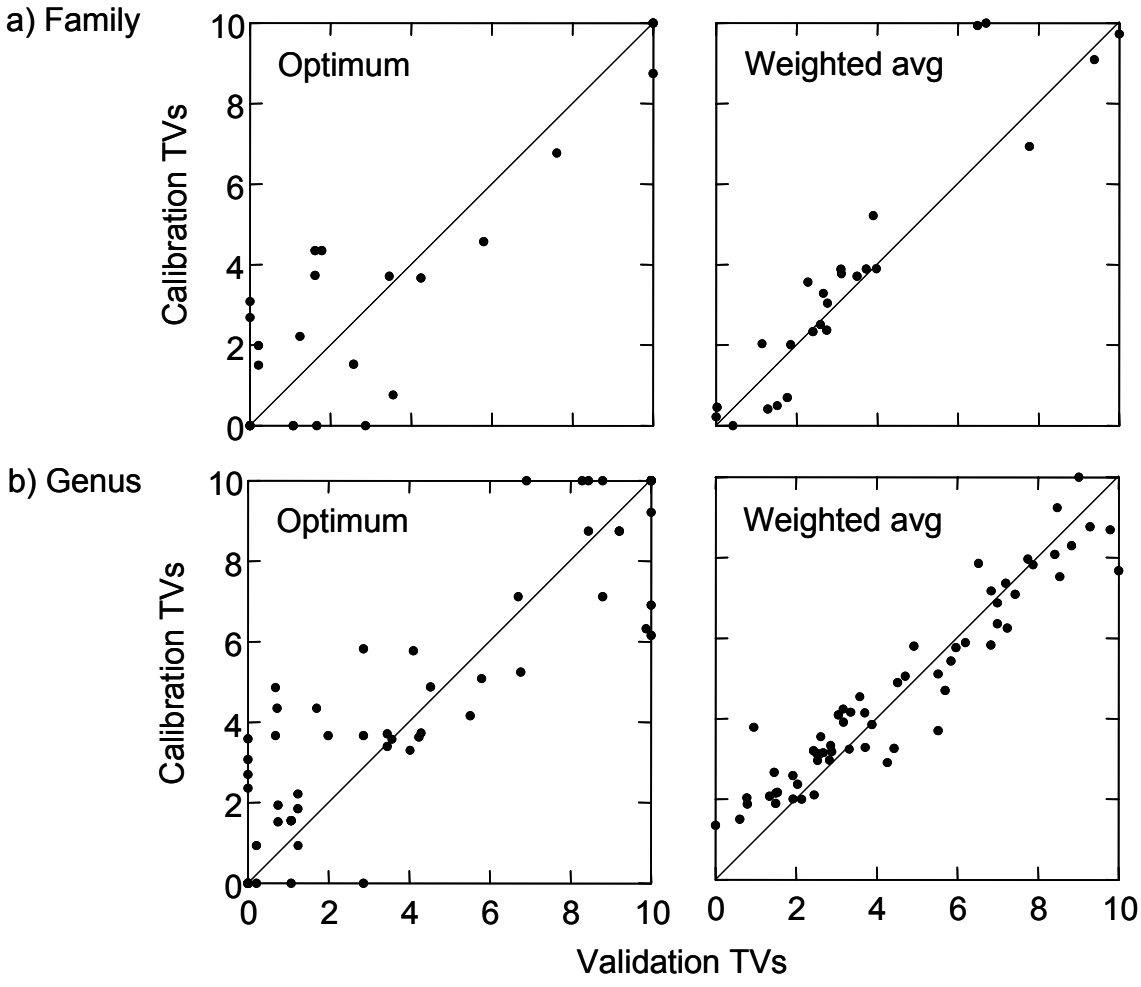


Figure 8. Validation and calibration TVs matched by taxon for a) family and b) genus levels, based on the GAM approach. The diagonal line represents the same TV for a taxon in both the validation and calibration data sets.

3.3.5 Comparison of Tolerance Metrics

At the family level, the results varied greatly across metrics tested. HBI scores based on calibration- and validation-based TVs showed a strong, tight trend with a slight shift in values and good separation between reference and impaired for the EPT approach (Figures 9 and 10). Relationships were also strong for the PCA, predictive modeling and GAM weighted average approaches, but trends were not always linear with a slope of 1 (Figure 9), and discrimination between reference and impaired sites was much weaker or non-existent compared to the EPT approach (Figure 10). Intolerant taxa richness performed most consistently and favorably overall, with little variability between the validation- and calibration-based values (Figure 11) and strong separation between reference and impaired sites (Figure 12), regardless of the approach or scoring procedure used. Percent intolerant individuals was most variable, particularly for both scoring procedures of GAM approach (Figure 13). The EPT weighted approach showed the tightest relationship between the validation- and calibration-based values and was the only approach resulting in a strong separation between reference and impaired sites (Figure 14).

Results were similar for genus level data. Very strong relationships were observed between HBI scores using on calibration- and validation-based TVs for all approaches (Figure 15). There were clear shifts in HBI scores toward slightly higher values using the validation-based TVs for the EPT approaches but toward slightly lower values for the PCA 75th percentile and GAM optima approaches (Figure 15). Separation between reference and impaired sites based on HBI scores was observed for all approaches but strongest for both scoring procedures of the EPT approach (Figure 16). Tight relationships between calibration- and validation-based intolerant taxa richness values were observed for genus level data, but shifts were more evident for all approaches except the PCA weighted approach (Figure 17). Strong separation between reference and impaired sites was observed for all approaches for this metric (Figure 18). Percent intolerant individuals exhibited large variability between calibration- and validation-based values for all approaches but showed the tightest relationship for the EPT weighted approach (Figure 19). Discrimination between reference and impaired sites was strongest for the two EPT approaches but still evident for the PCA 75th percentile, predictive modeling, and GAM weighted average approaches (Figure 20).

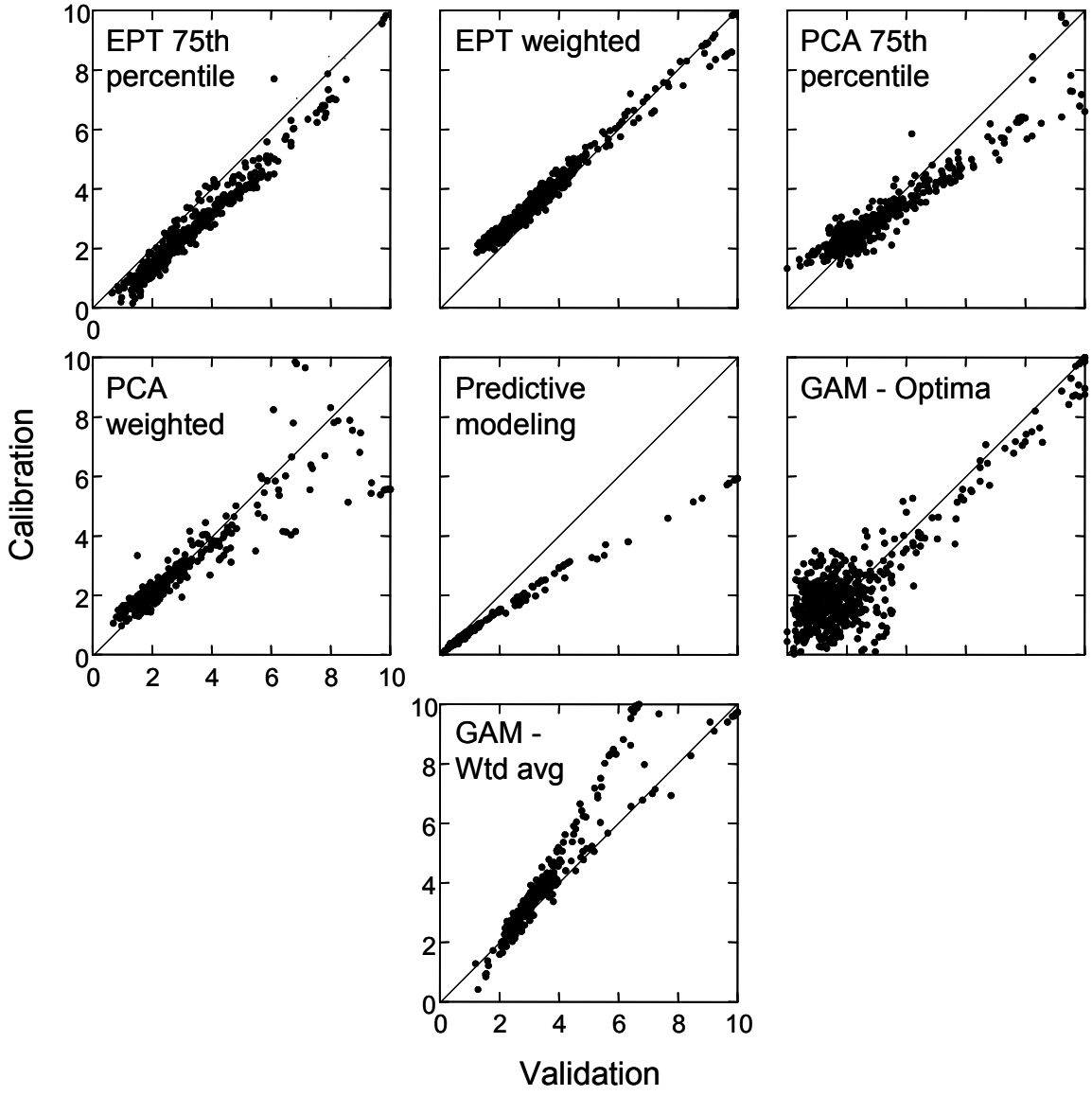


Figure 9. Plots of HBI scores based on TVs generated from calibration and validation data sets for family level data. Diagonal line represents matching values between the calibration and validation scores.

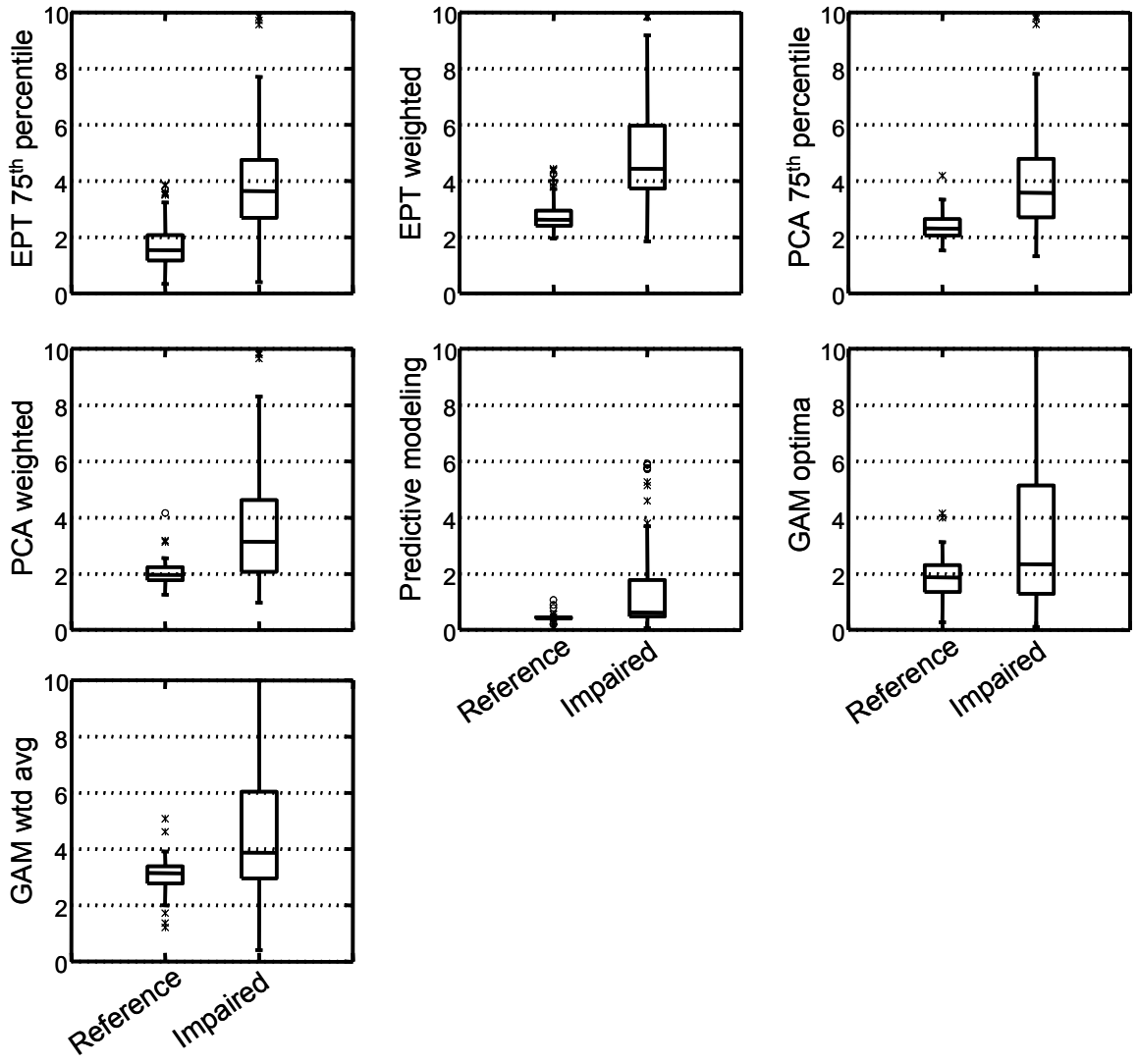


Figure 10. Distributions of family-level HBI scores for each approach using calibration-based TVs for reference and impaired sites.

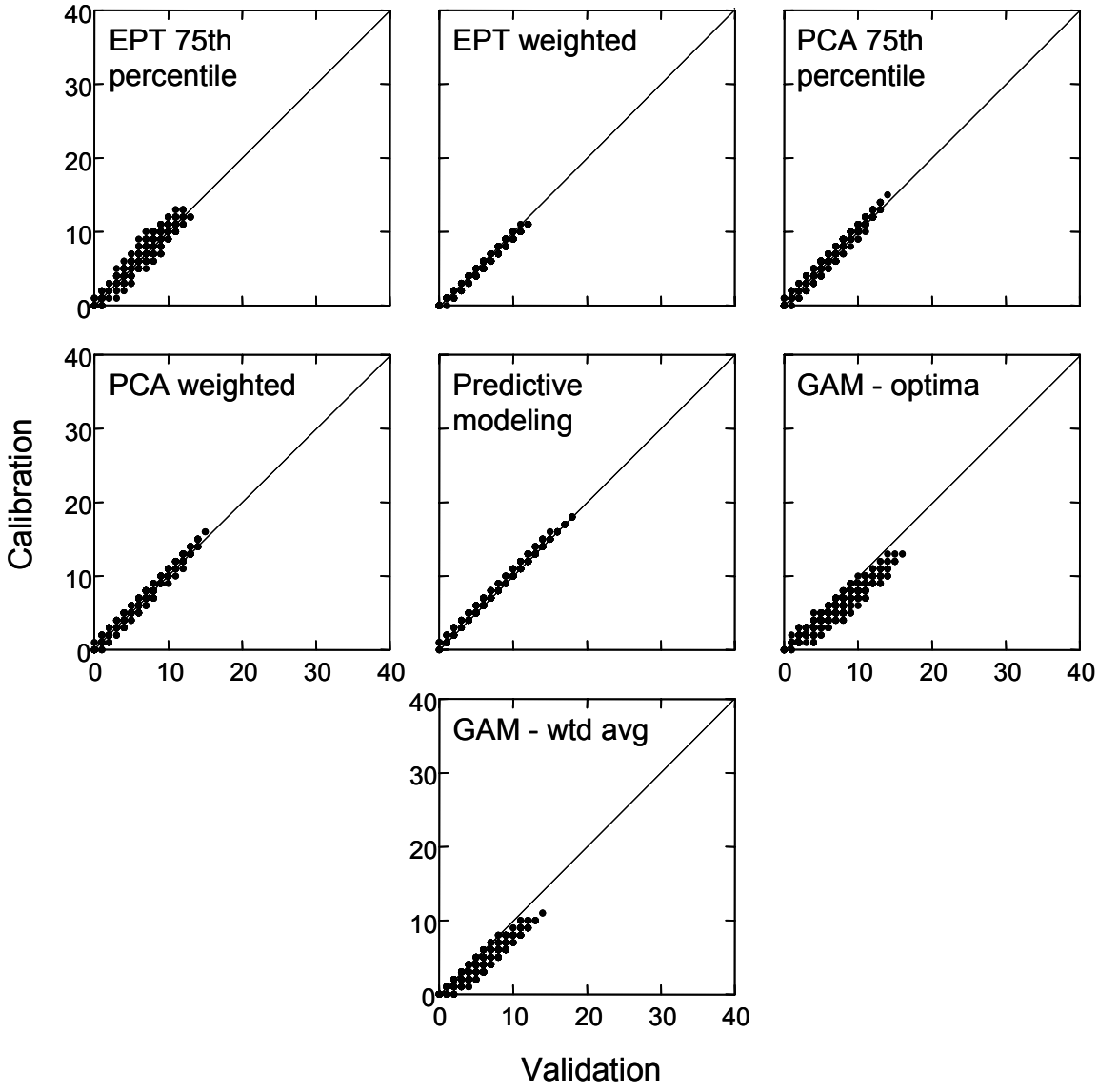


Figure 11. Plots of intolerant taxa richness based on TVs generated from calibration and validation data sets for family level data. Diagonal line represents matching values between the calibration and validation scores.

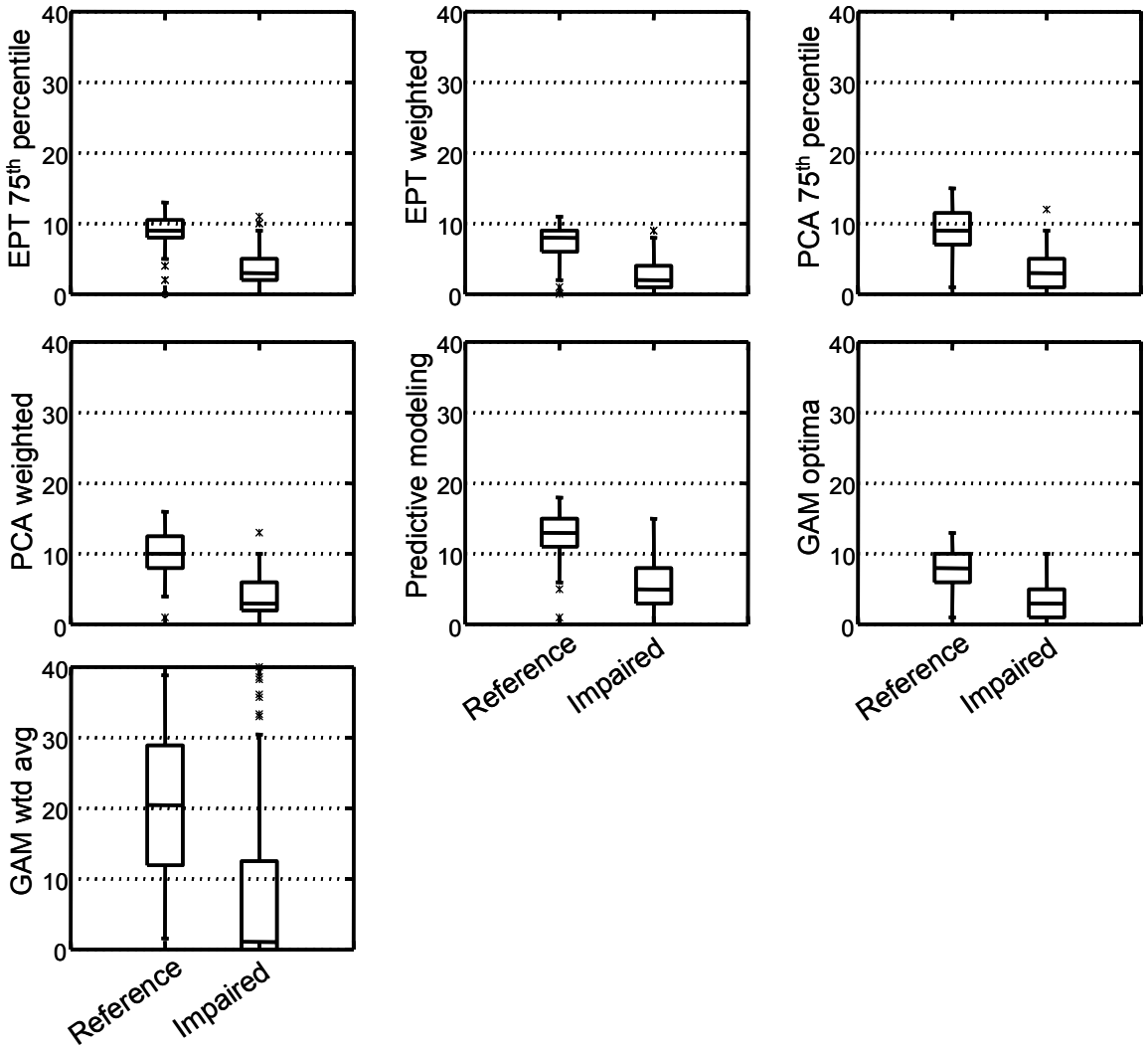


Figure 12. Distributions of family-level intolerant taxa richness for each approach using calibration-based TVs for reference and impaired sites.

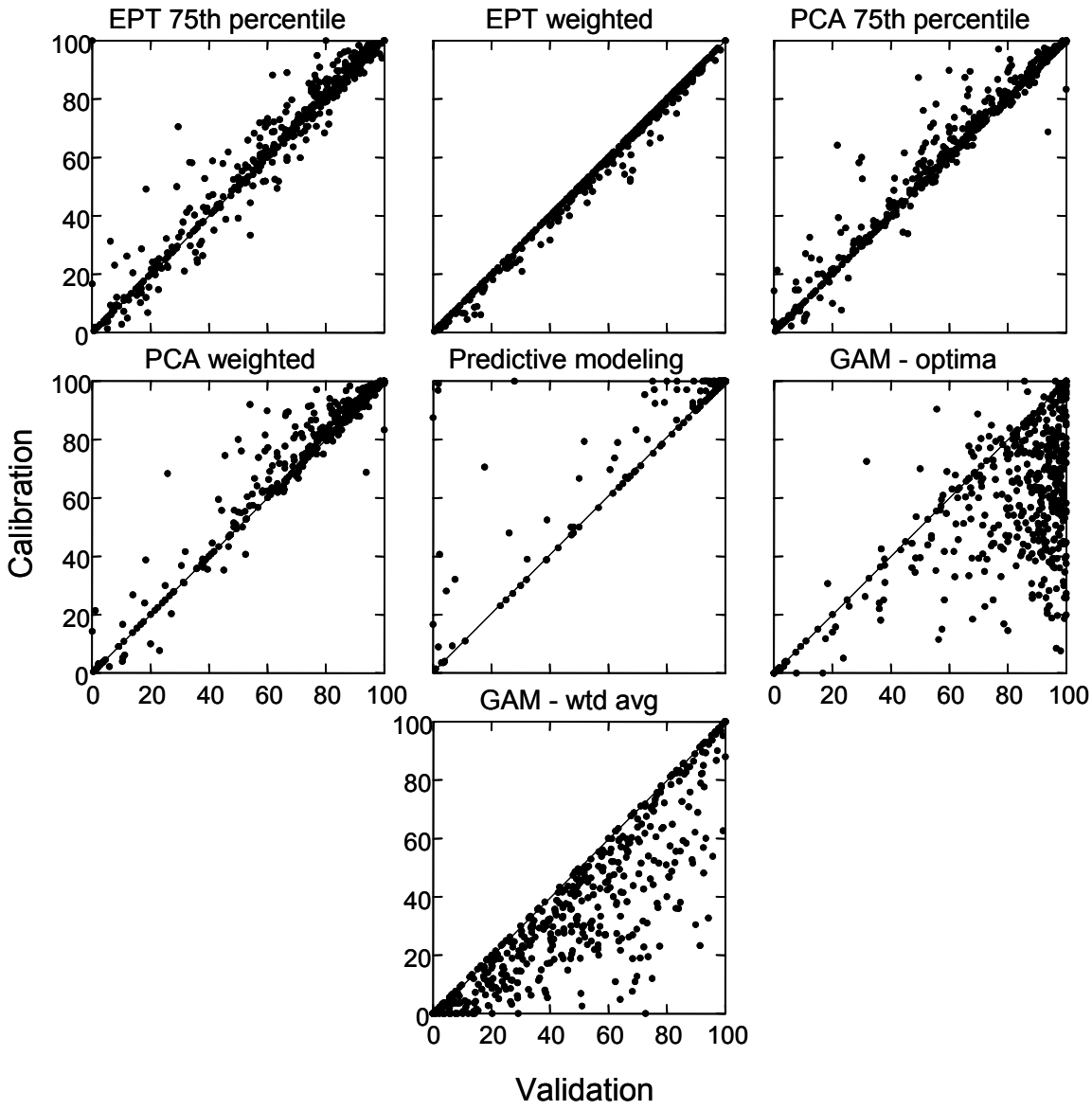


Figure 13. Plots of percent intolerant individuals based on TVs generated from calibration and validation data sets for family level data. Diagonal line represents matching values between the calibration and validation scores.

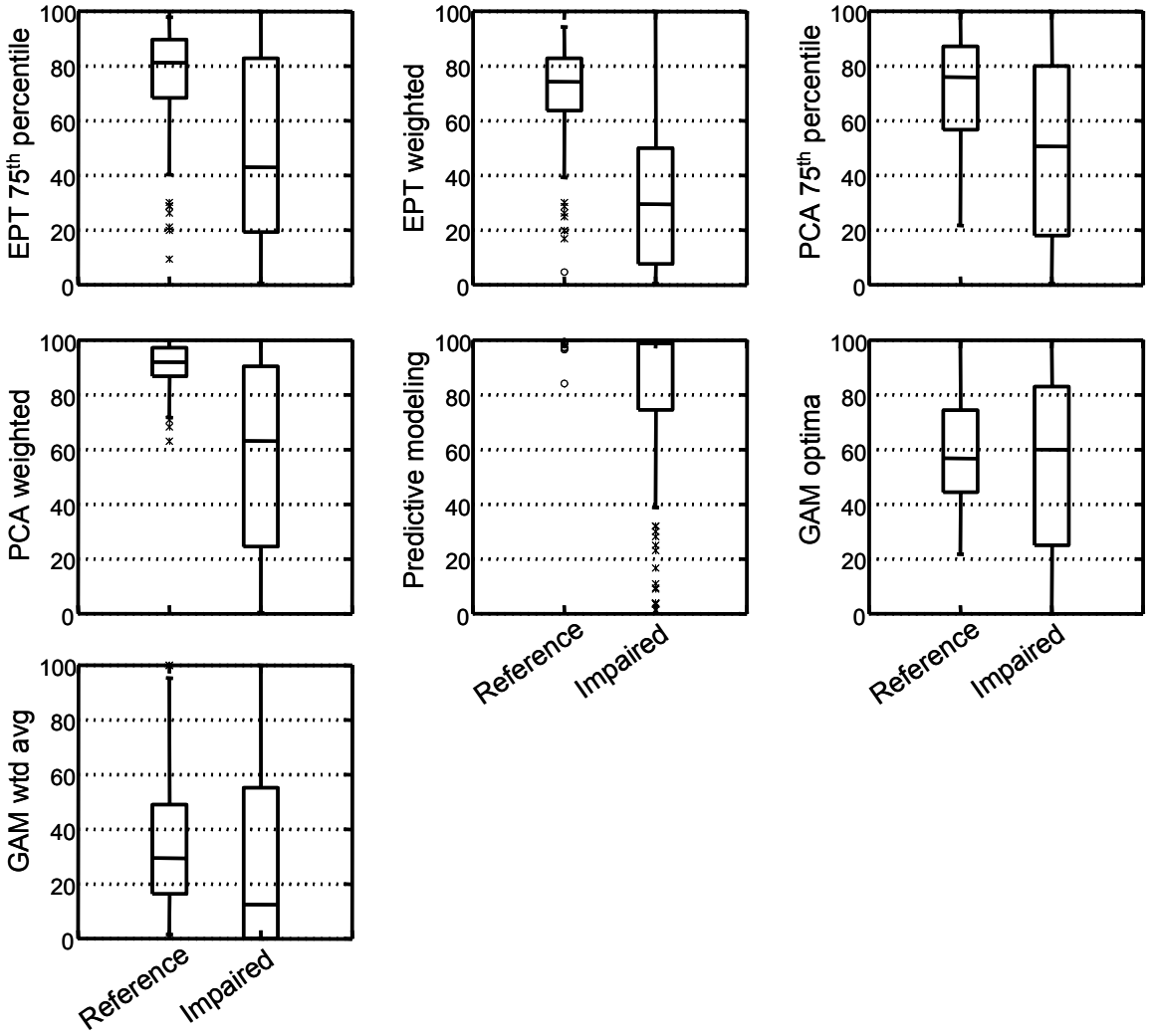


Figure 14. Distributions of family-level percent intolerant individuals for each approach using calibration-based TVs for reference and impaired sites.

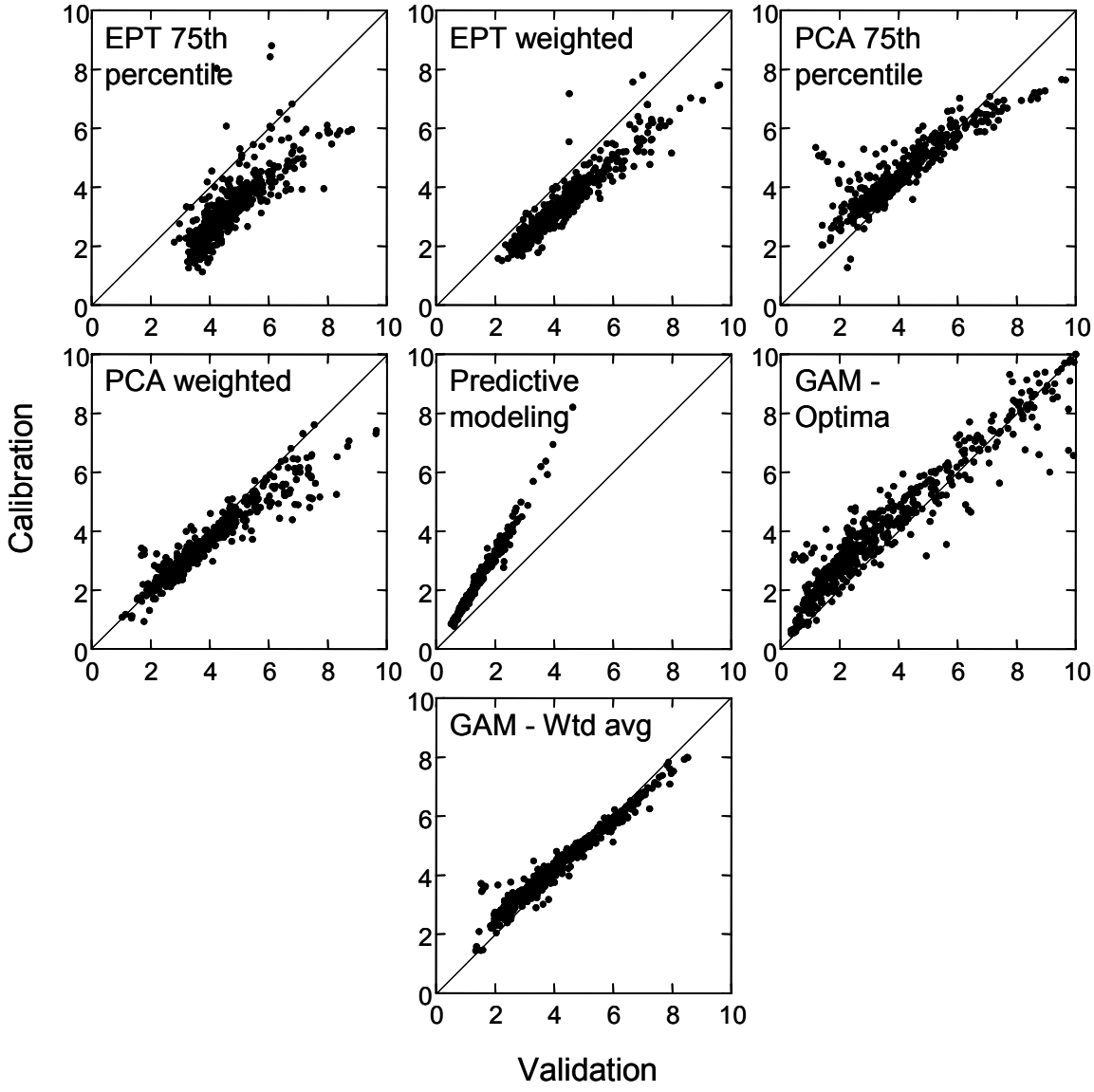


Figure 15. Plots of HBI scores based on TVs generated from calibration and validation data sets for genus level data. Diagonal line represents matching values between the calibration and validation scores.

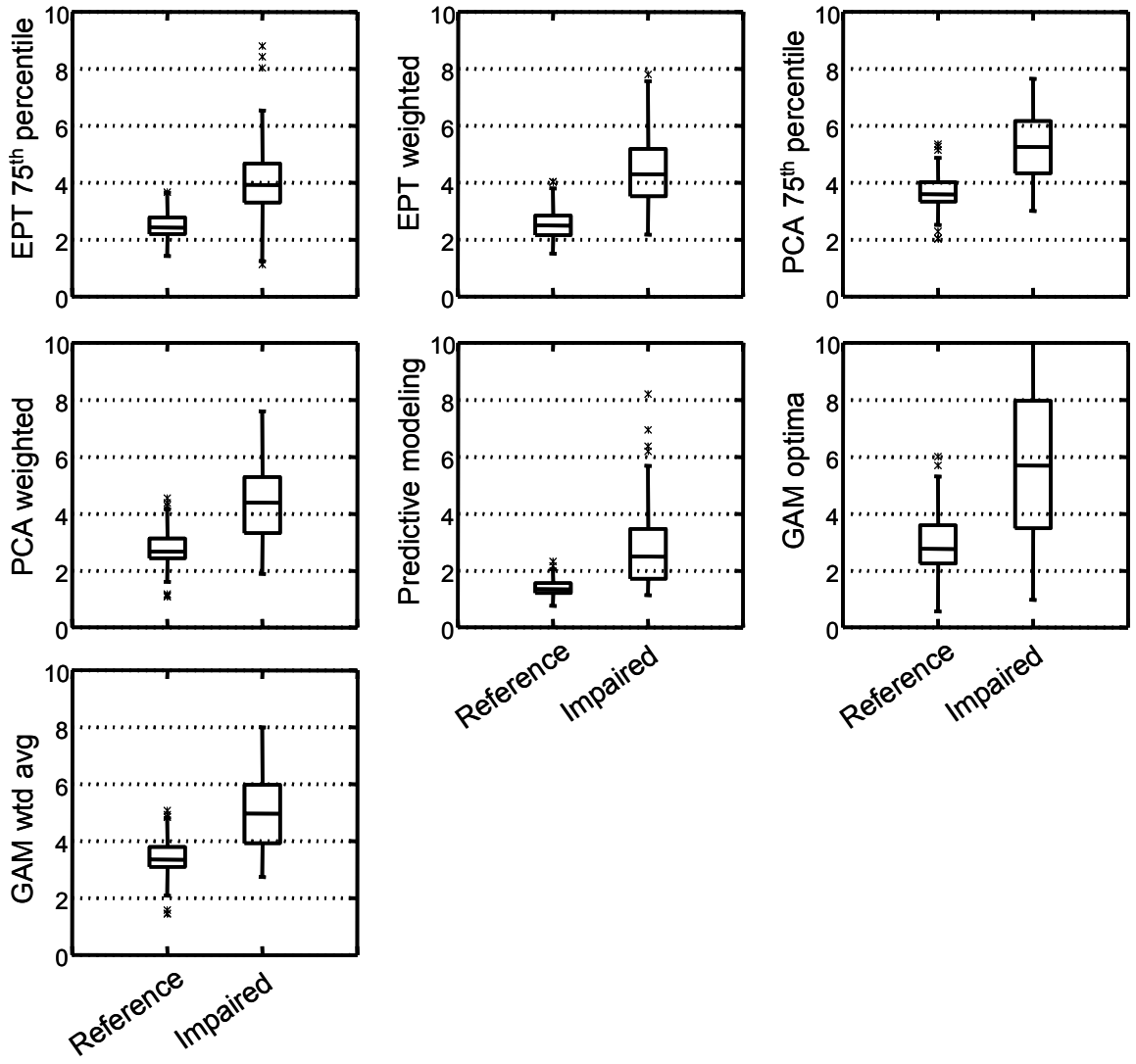


Figure 16. Distributions of genus-level HBI scores for each approach using calibration-based TVs for reference and impaired sites.

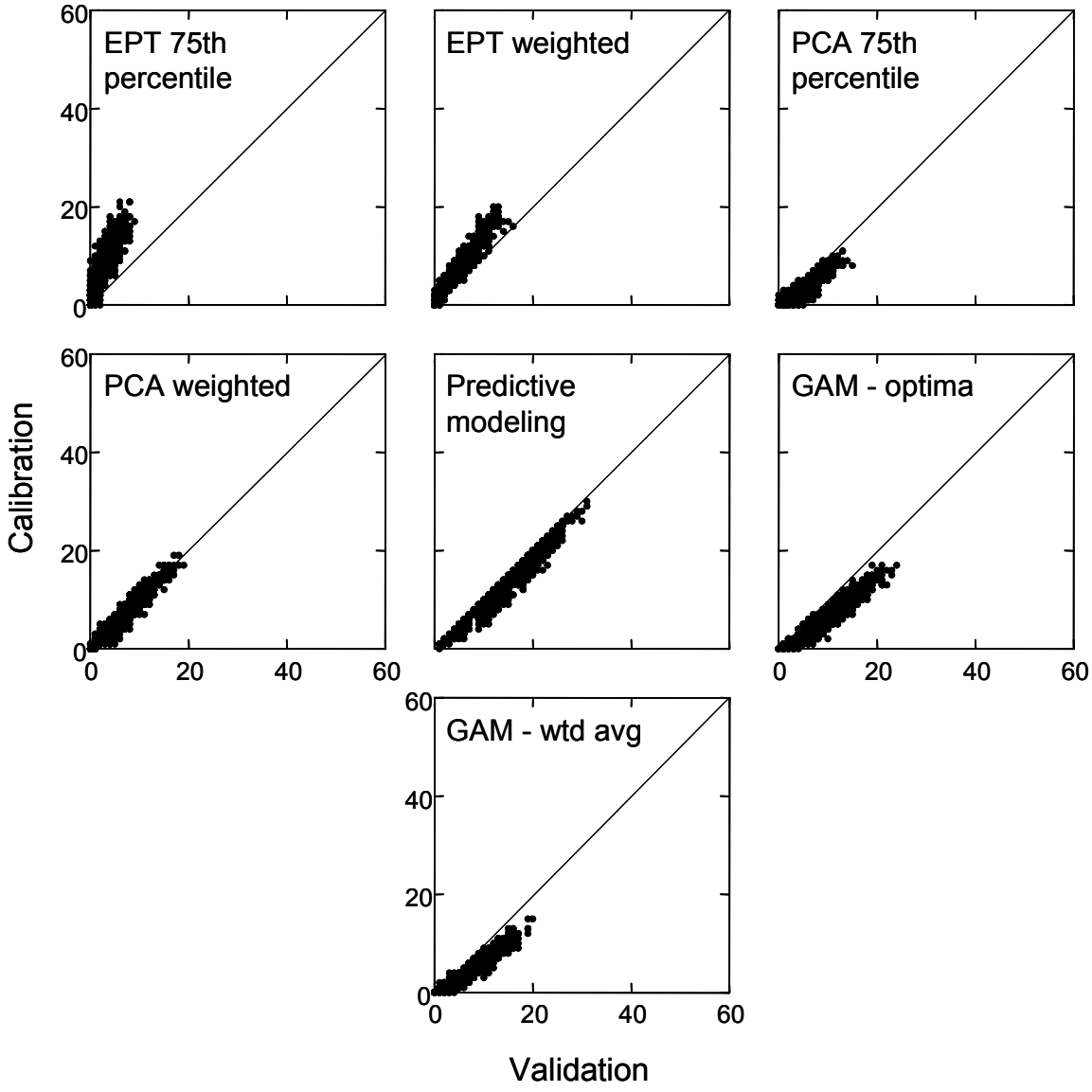


Figure 17. Plots of intolerant taxa richness based on TVs generated from calibration and validation data sets for genus level data. Diagonal line represents matching values between the calibration and validation scores.

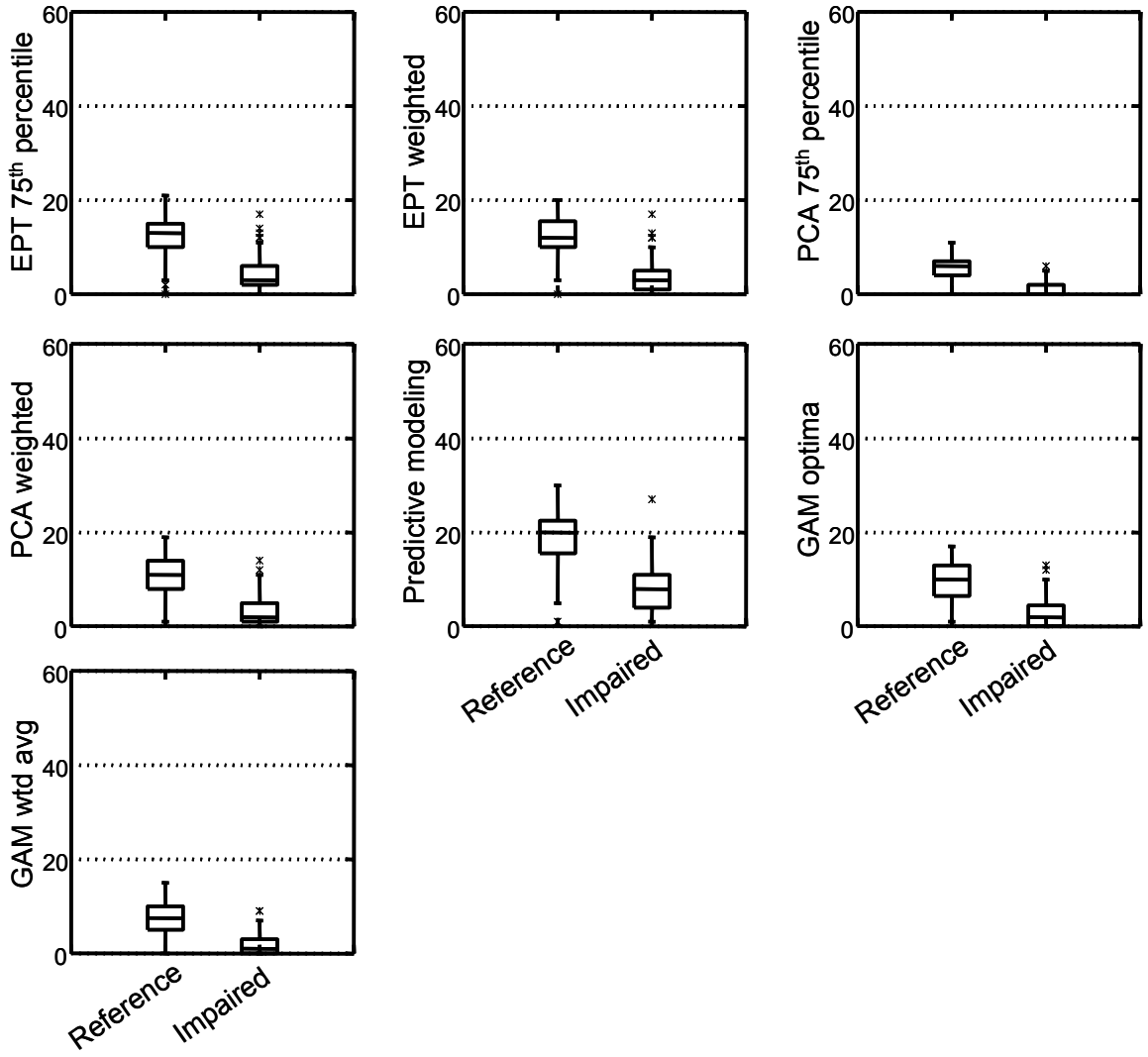


Figure 18. Distributions of genus-level intolerant taxa richness for each approach using calibration-based TVs for reference and impaired sites.

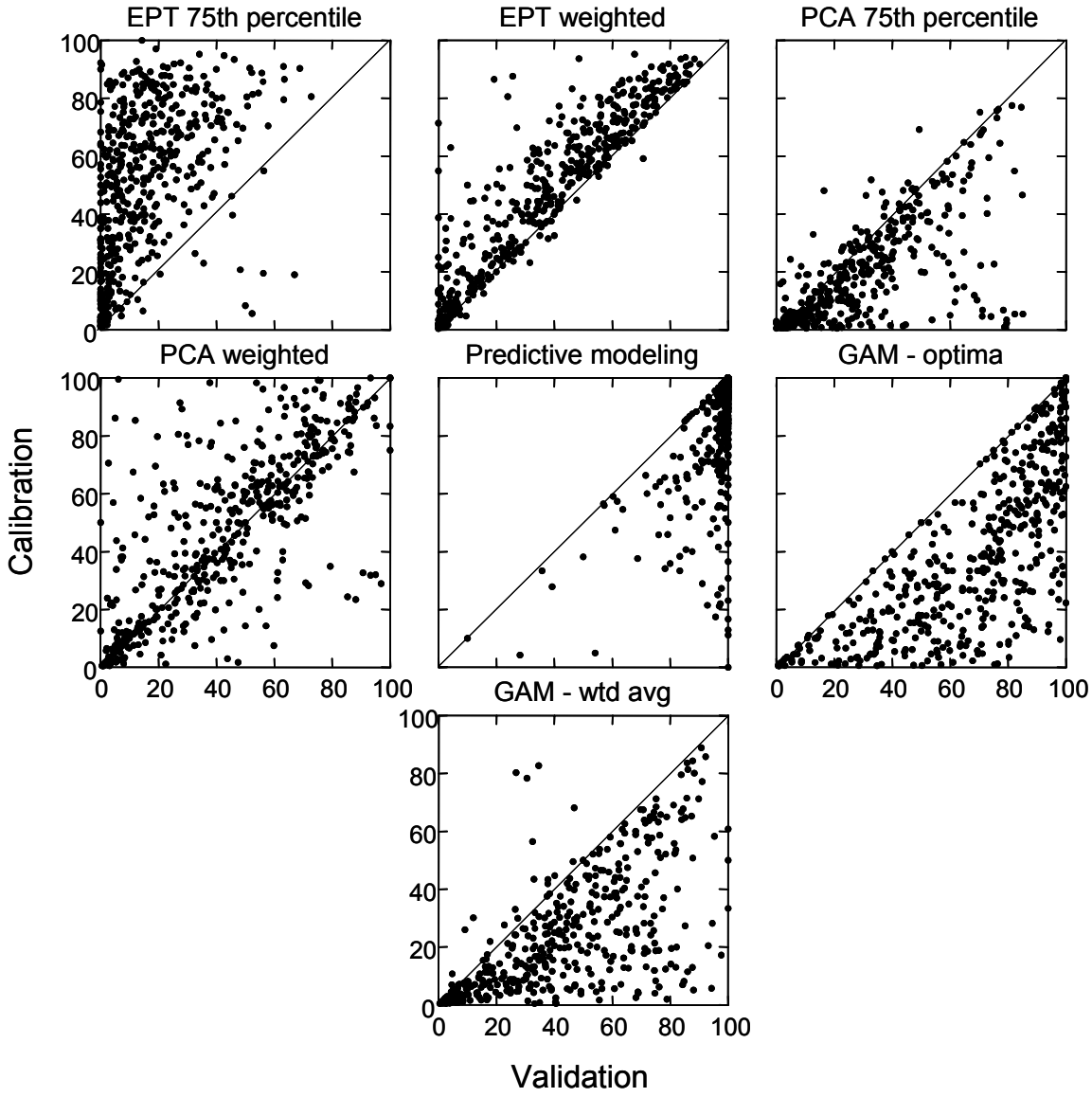


Figure 19. Plots of percent intolerant individuals based on TVs generated from calibration and validation data sets for genus level data. Diagonal line represents matching values between the calibration and validation scores.

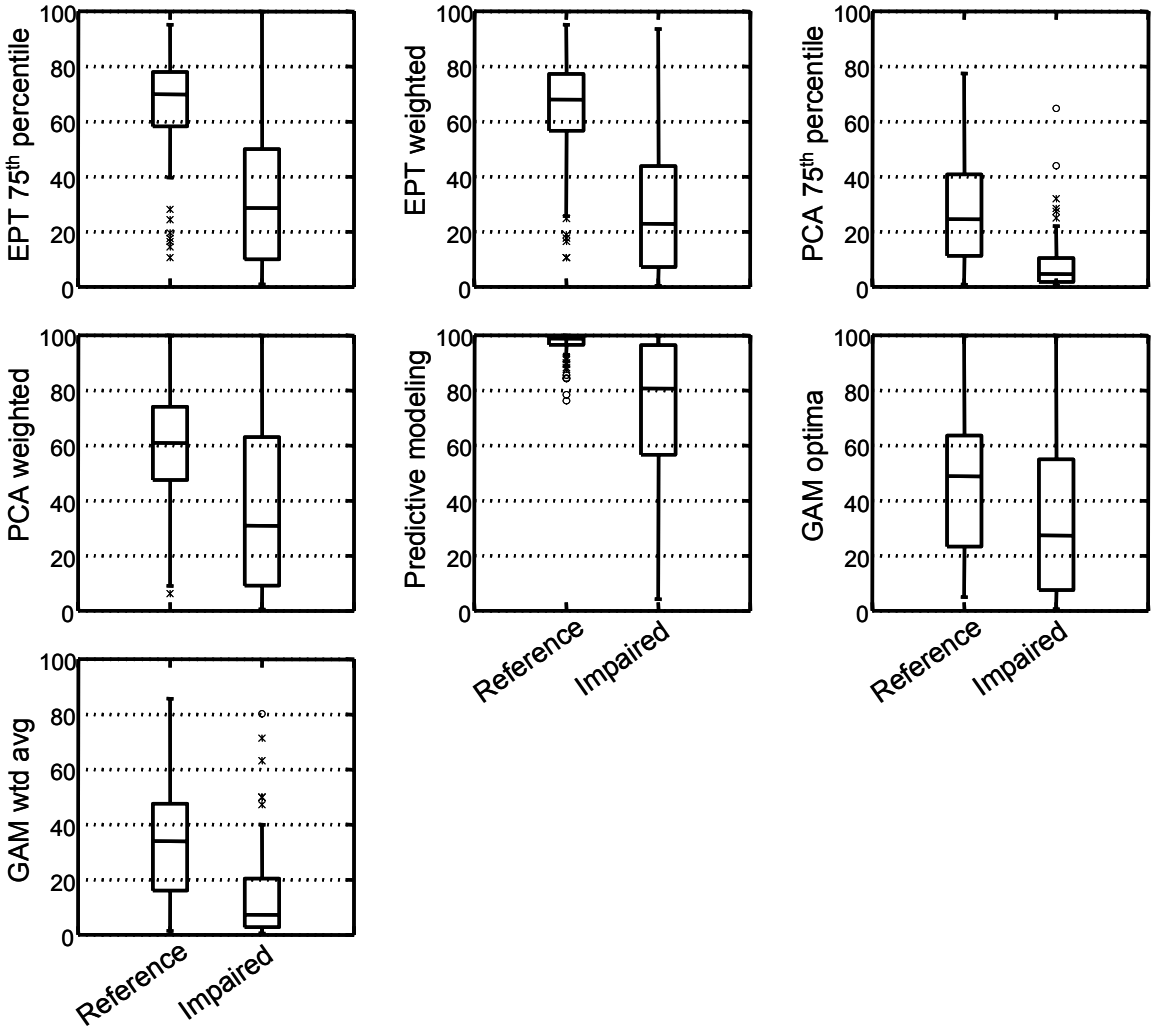


Figure 20. Distributions of genus-level percent intolerant individuals for each approach using calibration-based TVs for reference and impaired sites.

3.4 Discussion

Each of the four approaches evaluated for the development of generalized tolerance values performed a little differently in analyses. The EPT approach resulted in strong correlations between data sets, but there were also significant shifts in TVs from one data set to the other. The PCA approach tended to differ less between data sets, but correlations were not as strong. The predictive modeling approach as applied in this study exhibited high correlations between data sets and allowed calculation of TVs for the largest number of taxa. However, this approach only produced TVs for taxa that would be expected to be found at reference sites. The GAM approach provided TVs for a much more limited number of taxa and required more observations than any of the other methods. In this way, the GAM approach "selected" those taxa that showed a relatively strong relationship with the stressor gradient and excluded those without a strong association of some kind. Each approach has its own strengths and weaknesses, and the utility of a given approach is dependent on the data available and the ultimate use of the TVs produced from it.

The EPT approach is somewhat circular in nature because it does assume that EPT taxa are the ultimate signal of the quality of a water body. Inherently, this approach presumes that there is some underlying abiotic (stressor) gradient that is reflected in the EPT taxa richness. To remove the circular nature, however, one needs to use the underlying abiotic gradient rather than the biological measure associated with it. For this reason, use of the EPT approach must be accompanied by a caveat explaining that the TVs generated in this way are only worthwhile if the actual correspondence between EPT taxa richness and the general stressor gradient is strong. Even with this caveat, this approach may be particularly undesirable in calculating TVs for EPT taxa because it is truly circular to use EPT taxa as both the gradient and the response to that gradient. In general, the EPT method was best at discriminating reference and impaired sites, but there were also obvious, albeit slight, discrepancies between the validation and calibration data sets. Metrics based on TVs generated from the EPT approach produced the most consistent results between TVs based on calibration and validation data sets.

The PCA approach is dependent on existence or collection of potentially large amounts of abiotic data in order to effectively define the generalized stressor gradient existing across water bodies. In addition, these data need to be co-located with the biological data of interest (e.g., macroinvertebrates). This approach combined with the weighted procedure is very similar to the technique of weighted averaging, as described in ter Braak and Looman (1995). One disadvantage for weighted averaging that applies to this approach in general is that absences of a taxon are not taken into account as they are with the GAM approach. Therefore, the approach assumes a homogeneous distribution of samples along the environmental gradient and may generate misleading TVs if samples are unevenly apportioned to part of the gradient (ter Braak and Lookman 1995). As long as a state or tribe believes it has adequate data to effectively describe the general range of biological conditions found across a particular type of water

resource (e.g., wadeable streams) and samples are relatively well-distributed along the gradient, the PCA approach may be a useful one. At the level of HBI scores and intolerant taxa richness, the PCA approach performed almost as well as the EPT approach at separating reference and impaired sites. There was slightly better correspondence of HBI scores using validation- and calibration-based TVs for the PCA approach than for the EPT approach. However, this was not true for other metrics.

A major advantage of the predictive modeling approach is that a disturbance gradient does not need to be defined explicitly. However, the types of sites included in the analysis to represent "reference" condition can affect the results strongly. This approach resulted in TVs for the largest number of taxa overall and effectively reproduced TVs from the calibration to validation data sets, although shifts in values were observed. However, because of the nature of this approach, TVs were developed only for those taxa that tend to occur in reference sites. This approach could likely be modified to include other taxa as well, but it is not clear how well the resulting TVs would reflect conditions. Used in its current form, the HBI and percent intolerant individuals metrics were of very limited ranges, and discrimination between reference and impaired sites was restricted by these features. The intolerant taxa richness metric was much more consistent between calibration- and validation-based TVs, had larger ranges, and discriminated well between reference and impaired sites. In general, effective use of this approach requires extensive data on natural factors that affect the expected fauna at sites, and its use may be limited by the number and quality of reference sites available.

The GAM approach can provide more precise information on the nature of the relationship of a taxon to a gradient, and the number of taxa that could be used with this method was greater than for the EPT and PCA methods. This approach requires many sites in order to effectively determine the type of relationship, if any, that a taxon has with a generalized stressor gradient. Like the PCA approach, this method relies on the collection of additional information at each site in order to characterize the stressor gradient, and this may be a limitation for some data sets. Using the PCA axis as the gradient in the GAM models may not be as effective as using the individual variables in this approach, but using a large number of individual variables to replace the generalized gradient would require far more sites than were available.

The variations on scoring procedures examined for some approaches also showed differences. For both the EPT and PCA approaches, the weighted procedure performed better overall than the 75th percentile procedure. Weighted TVs generally resulted in less variability between the two data sets at the TV and the metric levels of comparison. In addition, metrics based on the weighted procedure tended to discriminate between reference and impaired sites better. Results were similar for the weighted average procedure for the GAM approach in comparison to the optimum procedure for calculating TVs. The weighted scoring procedures may exhibit lower variability because they integrate information over the whole gradient. In contrast, the 75th percentile and optimum procedures rely on identification of a single value, and this process should be associated with more error.

Among the three tolerance-based metrics examined, intolerant taxa richness was the most useful overall. It discriminated well and was generally repeatable across most approaches. This metric is based on grouping taxa into more broad categories, rather than relying on the precision of TVs for individual taxa. This feature likely leads to greater repeatability when compared to HBI scores. Both metrics seemed to discriminate well between reference and impaired sites, but not as strongly as expected or as typically seen for the HBI. This may be due to the fact that only a subset of taxa were used to calculate metrics in order to provide a more equitable comparison across approaches. However, this issue is a general limitation of any approach used to develop TVs because some taxa were excluded from each approach. This results in exclusion from tolerance-related metrics potentially large numbers of taxa that are relatively rare overall. If these excluded taxa are always found in small numbers in samples the effect on the HBI for any given sample might be affected very little. However, if those relatively rare taxa occur in large or even moderate proportions in a given sample, the effect on the HBI could be relatively large. The effect of these exclusions on the assessment of a site must be understood and addressed.

Finally, there was not a clear difference in the precision of family-level TVs when compared to those generated at the genus level. The ultimate ability of metrics to discriminate between reference and impaired sites was somewhat better overall at the genus level. However, tolerance values could be developed for a higher proportion of families than genera because a larger proportion of families had adequate sample sizes. It makes sense that the genus level data would provide more precise TVs, because it is less likely that there will be taxa with vastly different tolerances within a single genus than within a family. When taxa with different tolerances are lumped together into a single group, the breadth of tolerance values may no longer be adequately represented. In such cases, either the tolerances of some taxa are ignored or masked, or an average tolerance which represents the group no longer represents any of the individual taxa in the group. This is a limitation of using family level data to develop tolerance values, but it also applies to a lesser extent to genus level data.

3.5 Recommendations

The choice of approach to use depends largely on the type of data available for development of tolerance values. When abiotic data are available that adequately characterize the gradient of disturbance in the region of interest, employing an approach that incorporates these data is more desirable. The PCA and GAM approaches both utilize extensive abiotic data to characterize sites having corresponding macroinvertebrate data. In addition, the GAM approach benefits from additional site information (e.g., watershed area, elevation, etc.) to account for natural variability. These approaches more directly relate taxonomic occurrence or abundance with the level of disturbance, and this may make the resulting TVs more defensible. For both the PCA and GAM approaches, the procedures for calculating TVs that used all available count data (i.e., weighted and weighted average, respectively), rather than just an

optimum or percentile, produced more consistent results. Thus, if either the PCA or GAM approach is used to develop TVs, the weighted procedure is recommended over the other procedures evaluated. Of the statistical techniques employed by these two approaches, PCA is simpler to perform and is available widely in statistical software packages, but GAMs may more precisely describe the relationship between individual taxa and environmental gradients. However, neither approach will be as valuable if the abiotic data lack variables that are important in describing the disturbance gradient.

If biotic and certain abiotic data are available on a set of samples identified as representing reference condition, the predictive modeling approach may be most useful. This is particularly true if the variables included in the abiotic data can be used to characterize natural classes of samples. Predictive modeling is particularly attractive in situations where limited or no data are available to describe the disturbance gradient itself, as the gradient is dealt with indirectly in this approach. However, this approach involves several steps that require the use of potentially complex multivariate statistical techniques. The techniques required can be found in many statistical software packages, but a lot of movement of data among different programs may be necessary to complete the development of TVs. In addition, some specialized statistical training or experience may be required to carry out the necessary techniques and interpret the results.

The EPT approach is the least desirable in terms of defensibility because it does result in a somewhat circular process. If no other approach is feasible, using EPT as the disturbance gradient could be considered a last resort. Still, there must be confidence in how well EPT values represent the full range of conditions occurring in the region. In addition, there must be a rationale for defining the disturbance gradient using EPT for the specific region of interest. In some types of streams, EPT richness may not be well-represented in general and specifically may not represent stream condition well. For example, in the EMAP-West study, samples from streams in the Plains region of the United States had a median EPT richness of only 5 and a 75th percentile of only 9 taxa (T. Whittier, Oregon State University, personal communication). Although this approach appears to be the simplest and most straightforward one, it has many drawbacks and limitations in practice.

No approach described in this report can be selected and carried out blindly. All require careful evaluation of the data available and the statistical techniques involved. The data set to be used in developing TVs is often the most limiting factor in terms of the choice of approach. Typically, data has already been collected, and the variables in that data set may or may not include those that are necessary to carry out a particular approach. The statistical techniques necessary for a particular methodology can also limit the choices available. Not only must the user have access to and familiarity with the appropriate software package to run analyses, but he/she must also be able to understand and interpret the results obtained. If suitable attention is given to these issues, an approach to developing TVs can be identified that is defensible and appropriate.

4 Literature Cited

Barbour, M.T., J. Gerritsen, B.D. Snyder, and J.B. Stribling. 1999. Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, Second Edition. EPA 841-B-99-002. U.S. Environmental Protection Agency; Office of Water; Washington, D.C.

Blocksom, K.A., J.P Kurtenbach, D.J. Klemm, F.A. Fulk, and S.M. Cormier SM. Development and evaluation of the lake Macroinvertebrate Integrity Index (LMII) for New Jersey lakes and reservoirs. Environmental Monitoring and Assessment 77 (3):311-333.

Bode, R.W., M.A. Novak, and L.E. Abele. 1996. Quality assurance work plan for biological stream monitoring in New York State. NYS Dept. of Environmental Conservation. Albany, NY. 89 p.

Chutter, F.M. 1972. An Empirical Biotic Index of the Quality of Water in South African Streams and Rivers. Water Res. 6:19-30.

Flotemersch, J. E., B.C. Autrey, and S.M. Cormier. 2001. Comparisons of Boating and Wading Methods Used to Assess the Status of Flowing Waters. EPA/600/R-00/108. Office of Research and Development, U.S. Environmental Protection Agency, Cincinnati, OH.

Green, J. 1990. Freshwater Macroinvertebrate Species List Including Tolerance Values and Functional Feeding Group Designations for Use in Rapid Bioassessment Protocols, Report No. 11075.05, Assessment and Watershed Protection Division, U.S. Environmental Protection Agency, Wheeling, West Virginia.

Guisan, A., T.C. Edwards, Jr., and T. Hastie. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modeling 157:89-100.

Hastie, T.J., and R.J. Tibshirani. 1990. Generalized Additive Models. Chapman and Hall, London. 335 pp.

Hawkins, C.P., R.H. Norris, J.N. Hogue, and J.W. Feminella. 2000. Development and evaluation of predictive models for measuring the biological integrity of streams. Ecological Applications 10:1456-1477.

Hawkins, C.P. 2004. Probabilities of capture, RIVPACS-based tolerance values, and diagnostic O/E indices. Proceedings from Western Tolerance Values Workshop, February 3-5, 2004, Corvallis, Oregon.

Hilsenhoff, W.L. 1977. Use of Arthropods to Evaluate Water Quality of Streams. Technical Bulletin No. 100. Dept. of Natural Resources. Madison, Wisconsin.

Hilsenhoff, W.L. 1982. Using a biotic index to evaluate water quality in streams. Department of Natural Resources Tech. Bull. No. 132. Madison, WI. 22pp.

Hilsenhoff, W.L. 1987. An improved biotic index of organic stream pollution. Great Lakes Entomologist 20: 31-39.

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

Hilsenhoff, W.L. 1988b. Seasonal correction factors for the biotic index. The Great Lakes Entomologist 21:9-13.

Johnson, R.K., T. Wiederholm, and D.M. Rosenberg. 1993. Freshwater biomonitoring using individual organisms, populations, and species assemblages of benthic macroinvertebrates. Pages 40-158 in D.M. Rosenberg and V.H. Resh (editors). Freshwater Biomonitoring and Benthic Macroinvertebrates. Chapman and Hall, New York.

Klemm, D.J., K.A. Blocksom, F.A. Fulk, A.T. Herlihy, R.M. Hughes, P.R. Kaufmann, D.V. Peck, J.L. Stoddard, W.T. Thoeny, M.B. Griffith, and W.S. Davis. 2003. Development and evaluation of a macroinvertebrate biotic integrity index (MBII) for regionally assessing mid-Atlantic highlands streams. Environmental Management 31:656-669.

Klemm D.J., K.A. Blocksom, W.T. Thoeny, F.A. Fulk, A.T. Herlihy, P.R. Kaufmann, and S.M. Cormier. 2002. Methods development and use of macroinvertebrates as indicators of ecological conditions for streams in the Mid-Atlantic highlands region. Environmental Monitoring and Assessment 78 (2):169-212.

Kolkwitz, R., and K. Marsson. 1909. Ökologie der tierischen Saprobien. Beiträge zur Lehre von des biologischen Gewasserbeurteilung. Internationale Revue der gesamten Hydrobiologie und Hydrographie, 126-152.

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

Lewis, P.A., D.J. Klemm, and W.T. Thoeny. 2001. Perspectives on use of a multimetric lake bioassessment integrity index using benthic macroinvertebrates. Northeastern Naturalist. 9. 233-246.

Mississippi Department of Environmental Quality (MDEQ). 2003. Development and Application of the Mississippi Benthic Index of Stream Quality (M-BISQ). June

30, 2003. Prepared by Tetra Tech, Inc. (Owings Mills, MD) for the Mississippi Department of Environmental Quality, Jackson, Mississippi.

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

Putman, R.J., and S.D. Wratten. 1984. Principles of Ecology. University of California Press, Berkeley, California. 388 pp.

Ter Braak, C.J.F., and C.W.N. Looman. 1995. Regression. Pages 29-77 in R.H.G. Jongman, C.J.F. ter Braak, and O.F.R. van Tongeren (editors), Data Analysis in Community and Landscape Ecology. Cambridge University Press, Cambridge.

USEPA. 2002. Consolidated Assessment and Listing Methodology Toward a Compendium of Best Practices. 1st Edition. Office of Wetland, Oceans and Watersheds. United States Environmental Protection Agency. Washington D.C.

Yuan, L.L. 2004. Assigned macroinvertebrate tolerance classifications using generalized additive models. Freshwater Biology 49:662-677.

For a Historical Review of Biotic Indices:

Metcalfe, J. L. 1989. Biological water quality assessments of running waters based on macroinvertebrate communities: history and present status in Europe. Environmental Pollution 60:101-139.

Washington, H. G. 1984. Diversity, biotic and similarity indices. A review with special relevance to aquatic ecosystems. Water Research 18:653-694.

United States
Environmental Protection
Agency

Office of Research
and Development (8101R)
Washington, DC 20460

Official Business
Penalty for Private Use
$300

EPA 600/R-06/045
April 2006
www.epa.gov

Please make all necessary changes on the below label, detach or copy, and return to the address in the upper left-hand corner.

If you do not wish to receive these reports CHECK HERE

☐; detach, or copy this cover, and return to the address in the upper left-hand corner.

PRESORTED
STANDARD
POSTAGE & FEES
PAID