Refinement of the Basin-Wide Index of Biotic Integrity for Non-Tidal Streams and Wadeable Rivers in the Chesapeake Bay Watershed

REPORT

May 25, 2017

Zachary M. Smith
Claire Buchanan
Andrea Nagel

Interstate Commission on the Potomac River Basin (ICPRB)
30 West Gude Drive, Suite 450
Rockville, MD 20850
www.potomacriver.org

ICPRB Report 17-2

This report can be downloaded from the Publications tab of the Commission's website, www.potomacriver.org. To receive hard copies of the report, please write

Interstate Commission on the Potomac River Basin
30 West Gude Dr., Suite 450
Rockville, MD 20850

or call 301-984-1908.

Disclaimer

The opinions expressed in this report are those of the authors and should not be construed as representing the opinions or policies of the U. S. Government, the U. S. Environmental Protection Agency, the Potomac basin states of Maryland, Pennsylvania, Virginia, and West Virginia, and the District of Columbia, or the Commissioners to the Interstate Commission on the River Basin.

Suggested citation for this report

Smith, Zachary M., Claire Buchanan, and Andrea Nagel. 2017. Refinement of the Basin-Wide Benthic Index of Biotic Integrity for Non-Tidal Streams and Wadeable Rivers in the Chesapeake Bay Watershed. ICPRB Report 17-2. Interstate Commission on the Potomac River Basin, Rockville, MD.

Executive Summary

The "Chessie BIBI," or Chesapeake Basin-wide Index of Biotic Integrity, is a multi-metric index that measures the biological quality of streams and wadeable rivers on a common scale. It is calculated from macroinvertebrate data collected by federal, state, and local stream monitoring programs in the Chesapeake Bay region. The index was first developed in 2011. This refinement was done for two reasons: recent additions to the stream macroinvertebrate database significantly increased the potential to hone the index's sensitivity, and it is now possible to develop and test genus-level metrics.

The analysis database contained 25,067 sampling events from across the Chesapeake Bay watershed. Sampling sites in 1st to 4th order streams were classified into five disturbance categories based on habitat and water quality information: Reference (best quality), Minimally Degraded, Mixed (indeterminate quality), Moderately Degraded, and Degraded (poorest quality). Biological populations in Reference streams represent the best attainable community structure and function and were used as a benchmark to measure the biological integrity of other streams. Key attributes of the stream macroinvertebrates (taxonomic serial number, functional feeding group, habit, pollution tolerances) were reviewed and updated. Eighty-four metrics were calculated from the raw counts of March to November samples. Metrics were scored with a method that identifies Reference and Degraded sites equally well. Metrics selected for the index were typically the most sensitive to degradation. Eight possible constructs for a multi-metric index were examined.

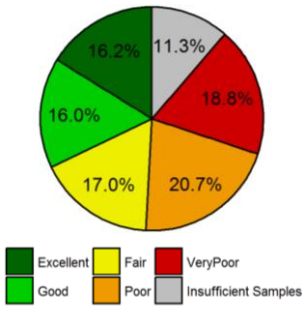
To address different information needs, the Chessie BIBI index was developed for two spatial scales: bioregion and region. The twelve bioregions accommodate natural variation in stream biota caused by hydrology, topography, and climate. The bioregion-specific indices are particularly suited for identifying local reasons of changing stream conditions and for measuring biological responses to restoration efforts. A coarser spatial division into the Inland and Coast regions proved most effective for reporting stream health for the Chesapeake Bay watershed as a whole. The Inland and Coast indices are sensitive to degradation but do not necessarily reflect natural differences between the bioregions.

Metrics keyed to order-, family-, or genus-level attributes were used to build versions of the index for different taxonomic resolutions of the raw counts. Order-level metrics are less sensitive, but they do not require laboratory enumeration and are suited for rapid screening in the field. Family-level metrics performed very well in most cases. They are recommended for use in the bioregion and region indices. Genus-level indices performed marginally better than family-level indices in some but not all bioregions. This is likely because genus-level metrics are affected by seasonal differences that are not accounted for in the indices.

A common scale of five narrative ratings was applied to the index scores of each taxonomic and spatial version of the Chessie BIBI index to compare stream health across jurisdictional boundaries in the Chesapeake watershed. The 50th, 25th, and 10th percentiles of each version's index scores in Reference environmental conditions were used to define Excellent, Good, Fair, and Poor macroinvertebrate status. A fifth rating, Very Poor, was defined by half the value of the 10th percentile. Paired comparisons demonstrate the family-level versions of the bioregion and regional indices produce comparable ratings in all but the Mid-Atlantic Coastal (MAC) bioregion.

The family-level region (Coast, Inland) indices are recommended for assessments of the Chesapeake Bay watershed. The region indices represent large geographic areas in the watershed. They have high CEs, and less complexity, lower metric variability, and lower variability in rating thresholds compared to the bioregion indices.

A simple count of the narrative ratings indicates biological integrity is Very Poor or Poor at 49.5% of sampling sites and Fair, Good, or Excellent at 50.5% of sites in the entire, updated database (1992 – 2015). The counts are roughly comparable to those reported in 2011 for the 2000-2008 period (Very Poor or Poor at 54% of sites and Fair, Good, or Excellent at 46% of sites). Straightforward counts such as these are misleading, however, because some areas—especially urban ones around Washington, D.C.—are more heavily sampled than others. When station ratings are weighted by the proportion of their local (HUC12) watershed area they represent and the weighted ratings are summed, the results indicate stream health is likely Very Poor or Poor in 39.5% of the Chesapeake watershed; Fair, Good, or Excellent in 49.2% of the watershed; and not known in 11.3% of the watershed. Many unsampled HUC12 watersheds are in predominantly agricultural or forested areas and, when sampled, may improve percentages of the Fair, Good and Excellent ratings. Area-weighted ratings provide a better starting point for measuring trends than simple counts of the ratings.

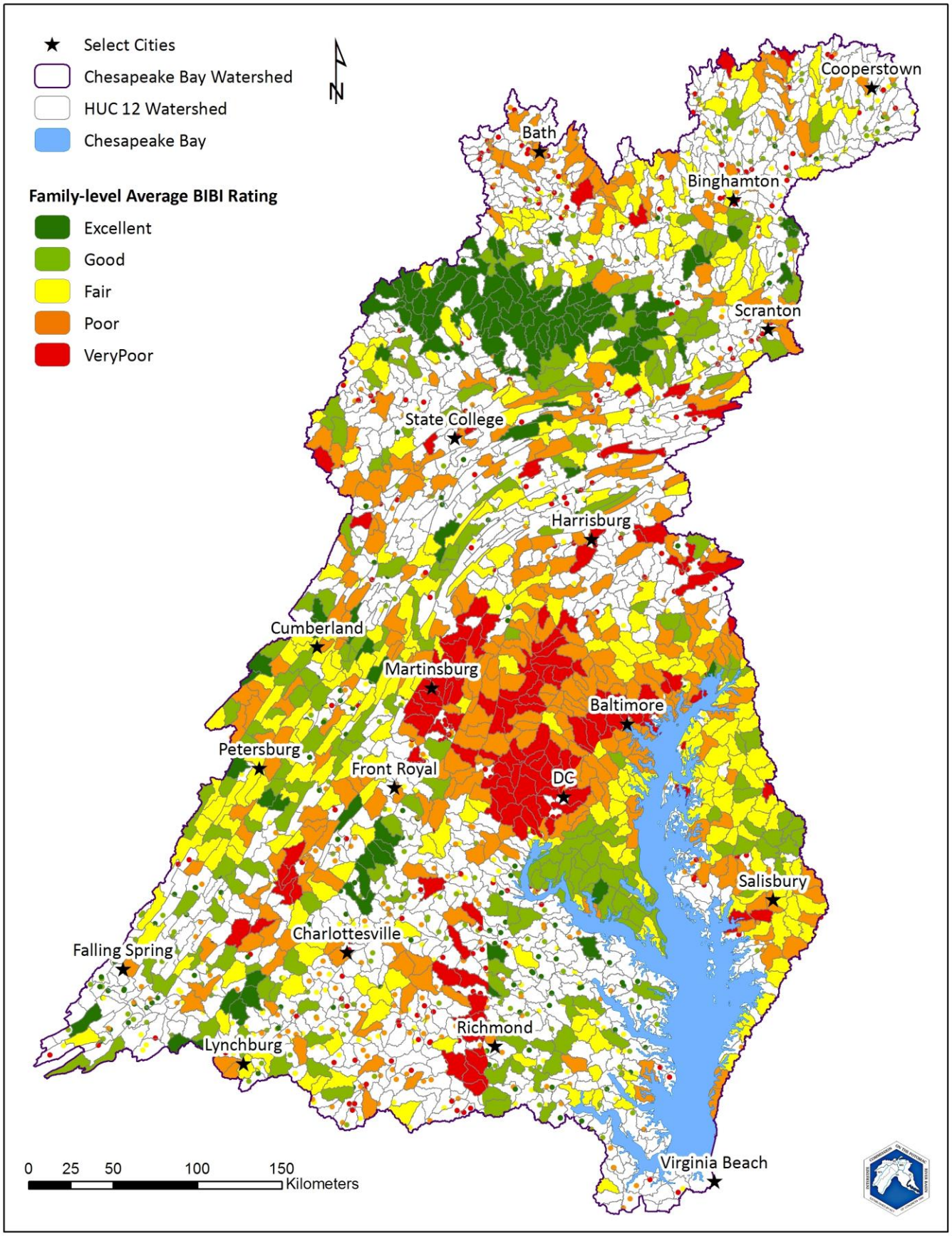


Area-weighted ratings of the family-level version of the Regional Index for Chesapeake watershed (1992 – 2015 data).

Like all indices, the Chessie BIBI index is dependent upon the idiosyncrasies of the data used to build it. The benefit of a large database is the increase in statistical power and the ability to transcend geopolitical borders. We strongly recommend that IBIs be developed cooperatively, across jurisdictional boundaries, to allow for coordinated analysis and evaluation of regions that are environmentally similar (i.e., bioregions). Collaboration will enhance the accuracy and reliability of the macroinvertebrate attribute assignments used to calculate many of the metrics. It will provide a succinct set of results that are more readily interpreted by non-experts—as opposed to differing index values and ratings reported by multiple programs for the same region.

Refinement of the Chessie BIBI was hampered by the fact that only eight habitat and three water quality parameters occur frequently enough in the database to be useful in classifying stream environmental conditions. There was also uncertainty in how various monitoring programs score the habitat metrics described in EPA's Visual-Based Rapid Bioassessment protocols (Barbour et al. 1999). We recommend stronger efforts to ensure that a standard suite of habitat and water quality measurements are made with comparable methods at all stream biological monitoring sites across the Chesapeake watershed. These measurements will benefit stream biological assessments in the long run. They will also improve each jurisdiction's ability to track and report incremental improvements in stream functions ("lift") that have not yet reached the point of benefiting biological populations and stream ecological health.

Facing page: Chessie BIBI (family-level version of the Region Index) ratings for streams and small rivers in the Chesapeake Bay watershed. When sufficient data ($n \geq 3$ catchments) are available, HUC12 watersheds are colored per the rating of their average index score. Otherwise, individual sampling locations are indicated and colored per their ratings.


Select Cities
Chesapeake Bay Watershed
HUC 12 Watershed
Chesapeake Bay
Family-level Average BIBI Rating
Excellent
Good
Fair
Poor
VeryPoor
N
Cooperstown
Bath
Binghamton
Scranton
State College
Harrisburg
Cumberland
Martinsburg
Baltimore
Petersburg
Front Royal
DC
Salisbury
Charlottesville
Falling Spring
Richmond
Lynchburg
Virginia Beach
0 25 50 100 150
Kilometers

Acknowledgements

This project was made possible through partial support provided by U. S. Environmental Protection Agency grants CB-96305701 (Chesapeake Bay Program) and I-98339413 (CWA §106), and by the Interstate Commission on the Potomac River Basin (ICPRB).

The Chessie BIBI index of non-tidal stream health was developed in 2011 in collaboration with biologists from across the Chesapeake Bay watershed. This report presents a refinement of the 2011 index. Technical work was performed by ICPRB staff on an updated database of stream macroinvertebrate data. An adhoc Technical Advisory Group (TAG) was created to guide the process and consisted of benthic macroinvertebrate experts from New York, Pennsylvania, Maryland, Virginia, West Virginia, Delaware, and the District of Columbia as well as federal, academic, and River Basin Commission partners. The authors wish to thank members of the technical advisory group for their guidance and feedback: Alexander J. Smith (NYDEC), Brianna Hutchison (SRBC), Christopher J. Victoria (Ann Arundel Co.), Dan Boward (MDDNR), David Ward (Loudoun Co.), Don Smith (VADEQ), Dustin Shull (PADEP), Elisha S. Rubin (DOEE), Ellen Dickey (DNREC), Ellyn Campbell (SRBC), Ginger Rogers (Versar Inc. for Howard Co.), Greg Pond (USEPA Region 3), Jason Hill (VADEQ), Jeff Bailey (WVDEP), Jennifer St. John (Montgomery Co.), John Wirts (WVDEP), Kelly Maloney (USGS), Mark Secrist (FWS), Michael Whitman (WVDEP), Mike Bilger (Susquehanna University), Mike Kashiwagi (MDDNR), and Richard Mitchell (USEPA). Other members of the Chesapeake Bay Program's Stream Health Workgroup provided input on final presentation of the results.

The Chessie BIBI could not have been produced without the cooperation of staff of the monitoring programs who responded to our data requests. The long hours and sometimes harsh conditions endured by field crews and the diligence and taxonomic expertise of the laboratory staff are the very solid foundation on which the Chessie BIBI index was built in 2011 and then refined in this study. Contributing monitoring programs are listed in Table 1 of the report. Finally, the authors especially thank Greg Pond (USEPA Region 3), Karen Blocksom (USEPA ORD) and A. J. Smith (NYDEC) for their many insightful comments and suggestions during critical stages of the index refinement; Mike Mallonee (ICPRB-CBPO) for his guidance and help in modifying, populating, and updating the relational database that houses the primary data; and Peter Tango (USGS-CBPO), Scott Phillips (USGS), and Jennifer Greiner (USFWS) for their support and encouragement.

Table of Contents

I. Introduction 1

II. Methods 2

A. Data Sources 2

B. Taxonomic Data Preparation 4

i. Taxonomic Classification and Hierarchy 4

ii. Taxonomic Attributes 4

iii. Method Standardization 5

iv. Rarefaction 7

v. Biological metrics 8

C. Environmental Data Preparation 8

i. Habitat 8

ii. Water Quality 9

D. Environmental Condition Classification 9

E. Spatial Classification 11

F. Metric Testing 16

i. Metric Sensitivity 16

ii. Range and Variability 19

iii. Metric Redundancy 19

G. Metric Scoring Approach 20

H. Index Construction 21

I. Index Classification Efficiency 24

J. Taxonomic Tiers 24

K. Delete-d Jackknife Validation 24

L. Delete-d Jackknife Precision 26

M. Narrative Rating Categories 27

N. Area-Weighting of Rating Results 29

III. Results 30

A. Index Construction 31

B. Chesapeake-Wide Index 31

C. Two Region Indices 32

i. Order-Level Indices 33

ii. Family-Level Indices 34

iii. Genus-Level Indices 35
D. Bioregion Indices 38
i. Order-Level Indices 39
ii. Family-Level Indices 44
iii. Genus-Level Indices 50
E. Index Validation and Precision 58
F. Narrative Ratings 58
G. Chesapeake Watershed Stream Health 63
IV. Discussion 66
A. Spatial Scales 66
B. Taxonomic Versions 67
C. Distributions of Index Scores 68
D. Narrative Ratings 69
V. Conclusions and Recommendations 70
A. Sample Collection 70
B. Data Analysis 71
C. Assessments 72
VI. Citations 75

Appendices

A. Taxonomic Classification (27 pgs)
B. Taxonomic Attributes (24 pgs)
C. Taxonomic Standardization (5 pgs)
D. Rarefaction (3 pgs)
E. Biological Metric Descriptions (11 pgs)
F. Abiotic Parameters for Evaluating Stream Environment (5 pgs)
G. Stream Classification (26 pgs)
H. HUC12 Watershed Characteristics in Bioregions (49 pgs)
I. Index Methodologies (19 pgs)
J. Scoring Methodologies (4 pgs)
K. Index Performance, Accuracy, and Precision (12 pgs)
L. Narrative Ratings and Maps of Index Scores (14 pgs)
M. Potential Biases in the Regional Index Ratings (9 pgs)
Appendix Citations (3 pgs)

Refinement of the Basin-Wide Index of Biotic Integrity for Non-Tidal Streams and Wadeable Rivers in the Chesapeake Bay Watershed

I. Introduction

Water quality assessments are often limited by geopolitical borders in the United States even though streams and rivers frequently cross state borders. Protocols performed by multiple independent monitoring programs can lead to inconsistent, fragmented assessments of waterbodies if those protocols are substantially different. In March 2008, Chesapeake Bay Program (CBP) partners combined state agency stream assessments in a map of stream macroinvertebrate impairment (Wolf 2008) and concluded the result could not adequately represent stream condition on a Chesapeake basin-wide scale.

The United States Environmental Protection Agency (USEPA) and Geological Survey (USGS) have used macroinvertebrates to evaluate United States waterbodies on a national scale. The USEPA collected 63 stream macroinvertebrate samples from across the 165,760 km^2 Chesapeake Bay watershed as part of its 2004-2005 Wadeable Stream Assessment (WSA) (USEPA 2006). WSA was replaced in 2008-2009 by the National Rivers and Streams Assessment (NRSA) and that program collected 37 samples within the watershed (USEPA 2016a). The USGS National Water Quality Assessment (NAWQA) program collected 254 stream macroinvertebrate samples in the Chesapeake Bay watershed between 1993 and 2012, primarily from its Potomac and lower Susquehanna study units. These federal programs benefit from a standard protocol and strict QA/QC measures; the results provide a statistical estimate of stream status for large regions of the country. However, these data sets have limited use when evaluating stream status on smaller scales.

All the states and several counties in the Chesapeake watershed routinely monitor stream biota for regulatory purposes. Other groups monitor for research or to measure restoration. These programs collected and enumerated more than 25,000 stream macroinvertebrate samples between 1992 and 2015. Although field methods can differ, there is frequently more similarity between methods than dissimilarity. Creating a unified database of the raw data from the various data sets and evaluating them with a standard protocol vastly improves the statistical power to characterize stream status, identify stressors, and detect responses to restoration efforts. Stream macroinvertebrate samples collected using slightly different sampling and processing protocols require more intensive post-collection QA/QC measures. However, they can yield similar biological assessment results (Ostermiller and Hawkins 2004, Astin 2006, Friberg et al. 2006, Gerth and Herlihy 2006, Herbst and Silldorff 2006, Southerland et al. 2006, Astin 2007, Rehn et al. 2007).

In 2011, the Chesapeake Bay Program (CBP) developed a Basin-wide Index of Biotic Integrity for stream macroinvertebrates, known as the "Chessie BIBI," for non-tidal streams and wadeable rivers in the Chesapeake Bay basin (Buchanan et al. 2011). The Chessie BIBI stems from the work of Astin (2006, 2007), who integrated existing data collected by seven agencies/programs and developed a wadeable stream index for the Potomac River basin.

Foreman et al. (2008) and Buchanan et al. (2011) subsequently expanded the index to the entire Chesapeake Bay basin, an area of approximately 165,760 km^2 that extends across seven states/districts (i.e., VA, WV, MD, DC, DE, PA, and NY) and several geomorphic regions. The index quantifies stream health seamlessly across jurisdictional boundaries in the Chesapeake Bay watershed. It is the stream health indicator named in the Strategy for Protecting and Restoring the Chesapeake Bay watershed (Executive Order 13508, 2010) and is identified in the Management Strategy for the Stream Health Outcome (Chesapeake Bay Program 2015) as the indicator for tracking improvements in stream health and function above an as yet undetermined 2008 baseline.

A refinement of the 2011 Chessie BIBI index was performed at this time for two reasons: recent additions to the stream macroinvertebrate database significantly enhanced the potential to hone the index's sensitivity, and it is now possible to develop and test genus-level metrics. The project evaluated: (1) intra-agency/program data integration, (2) multiple spatial scales, (3) site classification parameters, (4) new and old biological metrics, (5) metric scoring methodologies, (6) eight constructs of a multi-metric index, (7) the applicability of order, family, and genus level biotic indices, and (8) area-weighting of the index ratings to remove spatial biases. R-scripts were written to make the procedure for calculating the index faster, repeatable, and more accessible to future analysts. A Technical Advisory Group (TAG) was established to aid the project and review products. Results of the Chessie BIBI refinement are intended to support the establishment of a 2008 baseline for CBP reporting purposes.

II. Methods

A. Data Sources

In a series of data calls between 2007 and 2015, stream macroinvertebrate data and associated water quality and in-stream variables were obtained from twenty-nine federal, state, county, and non-profit agencies/programs that collect samples within the Chesapeake Bay basin (Table 1). A total of 25,067 samples collected with various methods between 1989 and 2015 have been incorporated into a common database. The elements and relationships of the common database are described in detail in Johnson (2013). Modifications made recently in the course of updating the database are described in Nagel (2016).

Table 1. Twenty-nine federal, state, and non-profit agencies/programs contributed to the Chessie BIBI database. The total count represents the number of unique sampling events contributed by the agency/program. A subset of the total count was used in the analysis.

	AGENCY/PROGRAM CODE	AGENCY/PROGRAM NAME	START DATE	END DATE	TOTAL COUNT
1	AAC_DPW_WERS	AACO-Watershed, Ecosystem, and Restoration Service	3/8/2004	4/14/2008	239
2	BAL_DPW_SMP	City of Baltimore - Stream Monitoring Program	4/3/2002	5/6/2010	277
3	BC_DEP_BCWMP	Baltimore County Watershed Management and Monitoring	4/1/2003	4/29/2008	601

	AGENCY/PROGRAM CODE	AGENCY/PROGRAM NAME	START DATE	END DATE	TOTAL COUNT
4	DC_DDOE_SMP	District of Columbia - Stream Monitoring Program	6/19/2003	5/21/2009	44
5	DNREC_DEBM	Delaware Biological Monitoring Program	10/16/2001	11/9/2011	106
6	FC-DPW_FCWMP	Frederick County Watershed Management Program	4/23/2001	8/21/2014	355
7	FC-SPS_FCSQAP	Fairfax County Stream Quality Assessment Program	7/31/2001	4/10/2008	239
8	HC_DPW_HCBMSA	Howard County Bio-Monitoring and Assessment Program	3/7/2001	5/12/2014	354
9	LC-DBD_LCSAP	Loudoun County Stream Quality Assessment Program	3/27/2009	10/12/2010	201
10	MC-SPS_MCSMP	Montgomery County Dept. of Environmental Protection	9/1/1989	10/21/2015	2,338
11	MDDNR_MBSS	Maryland Biological Stream Survey	5/10/1994	11/18/2010	7,472
12	MDDNR_MDCT	Maryland Core/Trend Monitoring Network	6/12/2000	8/6/2013	145
13	NYDEC_RSMP	New York Routine Statewide Monitoring Program	7/29/2002	9/29/2014	508
14	PADEP_PAOWQA	Pennsylvania other Water Quality Assessments	10/13/2003	2/20/2014	719
15	PADEP_PASWM	Pennsylvania Surface Water Monitoring Program	4/13/2000	8/9/2011	1,569
16	PADEP_PAUSGS	Pennsylvania USGS	3/12/1999	9/27/2012	149
17	PADEP_PAUW	Pennsylvania Unassessed Watersheds	6/6/2002	12/4/2003	43
18	PGC-DER_PGCSS	Prince George's County Programs and Planning Division	3/11/1996	4/7/2008	501
19	SRBC_TMDL	SRBC - Watershed Assessment and Protection - TMDL	9/4/2002	8/8/2013	53
20	SRBC_WA	SRBC - Watershed Assessment Program	7/6/1998	10/23/2013	1,799
21	USEPA_EMAP	EPA - EMAP Wadeable Streams Assessment	4/27/1993	9/13/1996	328
22	USEPA_MAHA	EPA - Mid-Atlantic Highlands Assessment	5/21/1997	9/14/1998	156
23	USEPA_NRSA	National Rivers and Streams Assessment	7/1/2008	9/28/2009	37
24	USEPA_WSA	EPA - Wadeable Stream Assessment Program	7/20/2004	11/10/2004	63

	AGENCY/PROGRAM CODE	AGENCY/PROGRAM NAME	START DATE	END DATE	TOTAL COUNT
25	USFS_SA	National Forest Service Stream Assessment	5/18/2000	5/8/2003	7
26	USGS_NAWQA	National Water Quality Assessment Program	6/2/1993	8/16/2012	254
27	VADEQ_SA	Virginia DEQ Benthic Monitoring Program	5/20/1992	11/28/2014	4,598
28	VCU_INSTAR	INteractive STream Assessment Resource	6/11/1999	11/3/2011	772
29	WVDEP_SA	West Virginia Dept. of Environmental Protection, Div. of Water and Waste Management	8/19/1996	10/1/2014	1,134
				TOTAL	**25,067**

B. **Taxonomic Data Preparation**

i. Taxonomic Classification and Hierarchy

The taxonomic status of the benthic macroinvertebrate taxa identified in the Chessie BIBI database were confirmed with the Integrated Taxonomic Information System (ITIS) database (*Retrieved [06/01/2016], from the Integrated Taxonomic Information System On-Line Database, http://www.itis.gov*). Up to ten taxonomic ranks were assigned to each taxon when available and applicable: phylum, subphylum, class, subclass, order, suborder, family, subfamily, tribe, and genus (Appendix A). ITIS also provides a Taxonomic Serial Number (TSN), a unique positive integer assigned to each taxon. Taxa in the Chessie BIBI database were paired with the appropriate TSN. Taxa that were not found in the ITIS database but deemed valid based on a literature review were assigned a unique negative integer. A negative TSN will never overlap with the officially assigned TSN from ITIS, which will allow for the database to be continually updated without incorrectly assigning the same TSN more than once. When applicable, spelling errors were corrected and invalid taxonomic identifications were updated to reflect current taxonomic nomenclature. The reported taxonomic name is archived as originally stated but the updated taxonomic name was used during analyses. If the taxon was identified to a taxonomic rank not included in the database (e.g., Superfamily or Subgenus), the final ID was rolled up to the nearest taxonomic rank. Additionally, complexes (i.e., an unofficial grouping of two or more closely related taxa) were also rolled up to the nearest taxonomic rank included in the database. Complexes were excluded because they have the potential to incorrectly inflate richness and diversity values. The list of taxa was further reviewed by members of the Technical Advisory Group (TAG).

ii. Taxonomic Attributes

Calculations of many benthic macroinvertebrate metrics rely on assigned taxonomic attributes, or traits. Tolerance values, functional feeding groups (FFG), and habits were assigned from available sources (Barbour et al. 1999, USEPA 2008, Chalfant 2009, Bollman et al. 2010, Buchanan et al. 2011, USEPA 2012, WVDEP 2015, Smith 2016). Inconsistencies and gaps

occur in the assignment of these attributes. In some cases, taxa have not been assigned a taxonomic attribute and assigning new attributes is beyond the scope of this study. In other cases, the sources that provided the attributes had differing assignments for the same taxon. If multiple sources provided tolerance values for the same taxon, the average of the tolerance values, rounded to the nearest integer, was assigned to the taxon. Categorical attributes (i.e., FFGs and habits) required more attention. Each categorical variable was assessed individually. For each attribute source and taxon, a total count of each variable was recorded. The variable with the highest total count was assigned as the final attribute. Another issue with categorical variables was that multiple attributes were often assigned to the same taxon within and between sources. Therefore, multiple taxa were assigned more than one attribute (e.g., collector-gather/predator) because all the variables had the same total count. Some of our attribute sources (USEPA 2008, Bollman et al. 2010, WVDEP 2015) assign multiple attributes to a single taxon to encompass attributes that are present at different life stages or the taxon exhibits a variety of attributes. Taxa with multiple attributes are effectively double counted in analyses; thus, incorrectly inflating the percentage of each metric class and resulting in a total percentage greater than 100% within a metric class (e.g., the sum of all percent FFG metrics). When the taxon with multiple attributes is abundant in the sample it has a substantial influence on two or more metrics within a metric category. To avoid any possible issues associated with multiple attribute assignments, each taxon with more than one attribute was reviewed and best professional judgement was used to select a single attribute to represent the taxon. The final attribute table was reviewed further by members of the Technical Advisory Group (TAG), and is provided in Appendix B. It should be repeatedly reviewed and updated in the future as individual taxa are better understood and characterized.

iii. Method Standardization

Differences in field and laboratory methodology can influence the taxonomic composition of samples and unintentionally bias analysis results. An analysis data set was created from the larger Chessie BIBI database that minimizes or removes the influences of many of these factors. Obvious factors were field method, stream size, and season. Only samples collected with a kick-net or a similar procedure were included in the analysis data set (i.e., D-Frame Net, Rectangular Dip Net, Kick Net, Kick Seine, and Slack Sampler). Hester Dendy Multi-Plate samplers, Surber Sampler, Hand-Picked samples, and unspecified collection methods were excluded from the analyses. Additionally, we limited our analyses to a Strahler stream order less than or equal to 4, which we considered to represent wadeable streams/rivers. Very few samples were collected between December and February (Figure 1). Samples collected during these months were excluded during analyses.

Undocumented differences in the laboratory procedures for enumerating stream macroinvertebrates can create bias. For example, some laboratories fail to explain their taxonomic identification rules beyond "the taxa were identified to the genus level or the lowest possible taxonomic resolution." To reduce variability among agencies/programs, taxa were standardized to Operational Taxonomic Units (OTUs) (USEPA 2016a) deemed appropriate for the Chessie BIBI database (Appendix C). The data were reviewed for taxonomic inconsistencies and taxonomic standards were set to reduce inter-agency/program variability. Taxonomic information is lost when specifying OTUs but this loss was necessary to assess data acquired from multiple sources. To identify taxa inconsistencies in the database, the taxa were aggregated by agency/program and a total count was provided for each phylum, subphylum, and class.

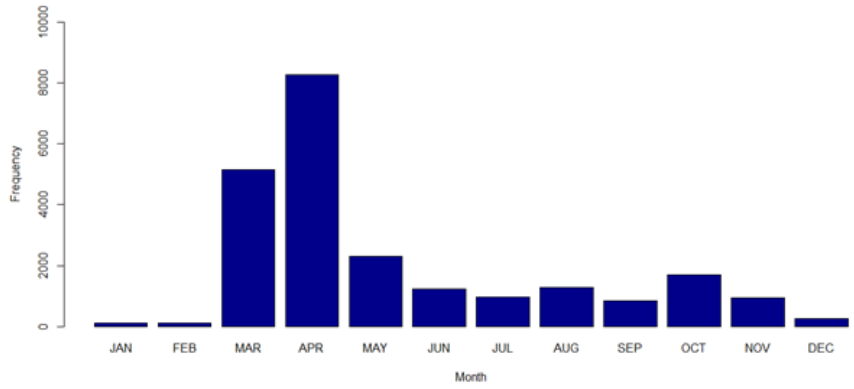


Figure 1. All the kick-net samples in the Chessie BIBI database were aggregated together and the frequency of unique sample events were plotted for each month.

Often it was apparent that some agencies or programs identified a taxon while others excluded the taxon from their assessment. For example, MD_MBSS and NYSDEC did not included mites (Chelicerata) during subsampling procedures, while other large data contributors, such as VADEQ, did include mites. As an additional confirmation, the number of sampling events containing the taxon and the mean relative abundance of the taxon in the samples for which the taxon was present were calculated. If it appeared that 1) at least one agency/program did not include a taxon, 2) the number of samples that the taxon was observed in was low (i.e., Phylum ≤ 5%, Subphylum ≤ 5%, and Class ≤ 1%), and 3) the mean relative abundance of the taxon was low (i.e., Phylum ≤ 3%, Subphylum ≤ 3%, and Class ≤ 1%), then the taxon was excluded entirely from the analysis. Although there is a loss of information and a minor loss in sample integrity, eliminating the taxon was a necessary action to reduce variability between agencies/programs.

We required counts of more than 70 individuals per sample in order to avoid skewing the percentage metrics in our analysis data set. If only two individuals were observed in a sample, each would receive a metric weight of 50% whereas when 100 individuals are observed each receives a metric weight of only 1%. Seventy was considered the greatest acceptable deviation from our lowest agency/program standard count (n = 100).

For the analysis data set, any taxon not classified within the phyla Annelida, Arthropoda, Mollusca, or Platyhelminthes was excluded. At the subphylum-level, taxa were excluded if they were not classified as Clitellata, Crustacea, Hexapoda, or Rhabditophora; if no subphylum-level existed within the ITIS database but the taxon could be classified within the four specified phyla, the taxon was not excluded from the analysis. At the class-level, taxa that were classified as Branchiopoda, Maxillopoda, and Ostracoda were excluded. Additionally, taxa within the families Gerridae, Hebridae, Velliidae, Hydrometridae, and Salidae were excluded from the analysis because they are classified as skimmer taxa. Skimmer taxa are considered semi-aquatic because they live on the surface of the water. They are not directly associated with the benthic macroinvertebrate communities, and therefore, should not be included in the development of a benthic macroinvertebrate index of biotic integrity. Finally, taxa of the order Hymenoptera were excluded because aquatic Hymenoptera are often small, parasitic organisms that may easily go unnoticed during processing. Carter and Resh's (2013) review of state agency benthic macroinvertebrate indices indicated that a similar list of taxa are excluded by one or more agencies in the United States.

Once the agencies/programs were standardized to exclude the same taxa, the taxonomic resolution of the organisms was assessed by agency/program. Generally, the lowest common denominator among agencies/programs was used to standardize the taxa. To observe different taxonomic resolutions, the taxa were first aggregated at a higher level taxonomic rank (e.g., Class). The total count at the higher level rank and subsequent lower level ranks were compared. If there was a large decline in the number of taxa identified at a higher level rank relative to the lower level rank, all taxa were rolled-up to the higher level rank during analysis. Bivalvia, Gastopoda, Oligochaeta, and Trepaxonemata taxa were rolled-up to the class-level, despite some agencies/programs identifying these organisms to the species-level. A relatively large number of Gastropoda individuals (n = 2,396) were not identified beyond the class-level. A conservative decision was made to aggregate all Gastropoda to the class-level. However, the 2,396 individuals not identified at the order-level only represented a 4% reduction from the individuals identified at the class-level. In future endeavors, it may be advantageous to aggregate Gastropoda at the order- or family-level, which may further improve index sensitivity to the defined disturbance gradient. Additionally, Collembola, Lepidoptera, Neuroptera, and Neoophora taxa were all rolled-up to the order-level. Again, this standardization process results in a loss of information but reduces the variability observed between the samples reported by each agency/program. It was difficult to assess the influence of agency/program beyond this point because sampling period, drainage, and ecoregion could confound observed differences. We concluded that the standardization process reduced the influence of agency/program on benthic macroinvertebrate composition, and subsequent divisions based on environmental factors addressed any remaining discrepancies.

iv. Rarefaction

Most agencies/programs have a standard subsampling procedure for randomly "picking" organisms from their stream samples. Standard counts are as low as 100 and some are greater than 500. Richness and diversity metrics are positively correlated with standard count because of the increased probability of finding rare taxa as the standard count increases (Gotelli and Colwell 2011). Such a relationship was observed in the Chessie BIBI database for family-level richness plotted against sample count (Figure 2).

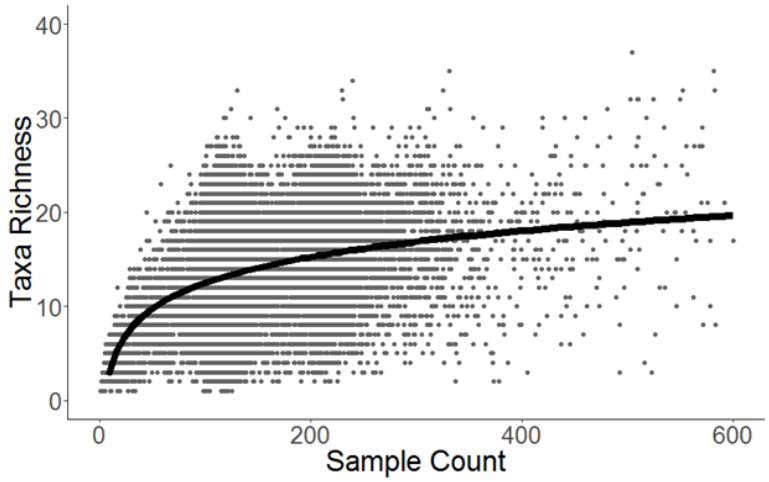


Figure 2. Sample with counts less than or equal to 600 (n = 22,240) were plotted against family richness. A base 10 logarithmic curve was generated with the available data.

To reduce the bias associated with sample count, all richness and diversity measures were calculated with each of the assemblages rarefied to a count of 100. A standard count of 100 was selected because it was the lowest common denominator among all the agencies and programs represented in the Chessie BIBI database. Rarefaction refers to a sample of the original assemblage without replacement until a

standard count is reached. A hypergeometric distribution is formed when sampling without replacement (Bunge and Fitzpatrick 1993). We propose that the rarefied count of each taxon can be estimated, just as a rarefied richness can be estimated. We developed a modified rarefaction method, probabilistic rarefaction with R-programming (R Core Team 2016) using a combination of the rarefied richness and the rarefied counts (Appendix D). Probabilistic rarefaction reduces the variability associated with estimating taxonomic composition at a lower standard count, whereas, rarefaction is more susceptible to higher variability. We used probabilistic rarefaction for calculations of richness and diversity measures in this study.

v. Biological metrics

Eighty-four biological metrics were identified in the literature (GADNR 2007, Pond et al. 2011, Carter and Resh 2013, Smith 2016). Additionally, the percentage of individuals in each Phylum, Subphylum, Class, Subclass, Order, Suborder, Family, Tribe, and Genus were systematically calculated when applicable. These additional composition metrics typically allowed for the assessment of 100-200 metrics.

During metric calculations, the taxonomic data were aggregated at a specific taxonomic level (i.e., Order, Family, or Genus). For Composition, Tolerance, Functional Feeding Group (FFG), and Habit metrics, the specified taxonomic level or the next lowest taxonomic level was used. However, richness/diversity metrics were only calculated using taxa identified to the specified taxonomic level. To prevent richness or diversity inflation, any taxon identified to a coarser taxonomic level was aggregated into an "unidentified" group.

The Hilsenhoff Biotic Index (HBI) and Modified Average Score Per Taxon (ASPT_MOD) calculations required each taxon in a sample to have an assigned TV. If the taxon did not have an assigned TV, it was not included in the calculation of these metrics. An issue arises when including taxa without TVs because these taxa are effectively assigned a TV of zero during metric calculations and would not accurately represent the sample. Therefore, the total number of individuals with an assigned TV was used as the denominator during TV related percentage metrics; taxa without TV were excluded from the calculations.

The metrics were calculated with custom R-functions (R Core Team 2016) and R-functions from the *Vegan* package (Oksanen et al. 2016). A complete list of the biological metrics, and their codes and descriptions, is given in Appendix E.

C. **Environmental Data Preparation**

Environmental data collected with the macroinvertebrate samples were used to create a standardized gradient of environmental conditions from Reference (best quality available) to Degraded (poorest quality). Agencies and monitoring programs have different protocols for collecting *in situ* environmental data, so the variables needed to be standardized prior to index development.

i. Habitat

Twenty-four habitat parameters are reported in the stream macroinvertebrate database (Appendix F). The EPA visual-based Rapid Bioassessment Protocol (Barbour et al. 1999) sought to standardize habitat measures for low and high gradient streams, however many monitoring programs modified these measures to suite their regulatory needs. Thus, only nine habitat parameters were measured consistently and frequently (i.e., more than 75% of sampling

events) and none of these parameters were collected at all sampling locations. One parameter, the velocity/depth ratio, was excluded because it contained odd values outside of the 0-20 standard scale used to score other habitat parameters. Future analyses may benefit from excluding odd velocity/depth ratio measurements and including velocity/depth ratio as a habitat parameter. The eight remaining habitat parameters were bank stability, bank vegetation, channel alteration, embeddedness, epifaunal substrate, flow, riffle/run/pool ratio, and sedimentation. These eight habitat parameters were used in conjunction with water quality parameters to classify environmental condition.

ii. Water Quality

Seventy-eight water quality parameters are reported in the stream macroinvertebrate database (Appendix F). Only four were collected frequently (i.e., more than 75% of the sampling events): temperature, specific conductivity, pH, and dissolved oxygen. Temperature was not included in the site classification process because the diel and monthly range can vary drastically. The remaining three water quality parameters were used to classify environmental condition.

D. **Environmental Condition Classification**

Karr's (1981) original fish IBI did not include the use of reference sites. After the introduction of the IBI concept, it quickly became apparent that assemblages collected at undisturbed sites could be used as a baseline to rate subsequent samples (Fausch et al. 1984, Karr 1991, Gibson et al. 1996). Reference sites typically represent the best obtainable or least-disturbed condition. Metrics indicative of degradation are discovered by performing pairwise comparisons of reference and test/degraded site metric distributions. Test sites represent all sites that were not considered reference, while degraded sites represent poor environmental conditions.

For this study, the condition of a sampling site was classified based on the three water quality parameters and scores of the eight habitat parameters above. Each water quality parameter received a score of 0 – 3 based on values used for state water quality assessments in EPA Region 3 or reported in the literature (e.g., Pond et al. 2011) (Table 2). Zero was assigned to the range of water quality values considered to be naturally occurring and to have minimal influence on stream macroinvertebrate survival. Reference thresholds for conductivity were selected after discussions with Technical Advisory Committee members and a review of the draft field-based methods for developing aquatic life use criteria for conductivity (USEPA 2016b). Reference thresholds for pH and dissolved oxygen were based on stream water quality standards in the Chesapeake basin states. Higher scores represented water quality conditions considered to be associated with anthropogenic stress and increasingly limiting to macroinvertebrate survival. Sites were classified as Reference if 75% or more of the available habitat scores were ≥ 16 and none were less than 12, and the sum of the three water quality scores was zero. Degraded sites had half or more of all available habitat scores ≤ 6 and the sum of the three water quality scores > 1. The Reference and Degraded environmental condition classifications (Table 3) were the only classes used to test metric sensitivity and index classification efficiency during the development of the indices. Three intermediate classes were also defined, i.e., Minimally Degraded, Mixed (includes all sites with insufficient data to classify condition), and Moderately Degraded. The Minimally Degraded and Moderately Degraded classes were used as a visual

validation that the index was appropriately detecting ecological response on a declining gradient from the Reference to the Degraded environmental condition.

Sampling events were excluded if the number of individuals counted was less than or equal to seventy were evaluated for the rating system. Low sample counts may be indicative of a degraded condition, and thus, it may be appropriate to categorize these samples as "Very Poor." However, when the low counts were associated with sample environmental condition classes there was no definitive pattern (Table 4). Although there were more Moderately Degraded and Degraded samples with low counts than Reference and Minor Degradation, most samples were classified as Mixed.

Table 2. Criteria used to assign water quality degradation scores to each sampling event. A total sum of degradation scores equal to zero was necessary to meet Reference conditions. Sampling events with summed degradation scores greater than zero were classified as various levels of Degraded.

	Score	Specific Conductivity (µS/cm)	pH (SU)	Dissolved Oxygen (mg/L)
Reference	**0**	$x \leq 300$	$6.0 \leq x \leq 8.5$	
	1	$300 < x < 750$	$5.0 \leq x < 6.0$ or $8.5 < x \leq 9.0$	
	2	$750 \leq x < 1000$	$4.0 \leq x < 5.0$ or $9.0 < x \leq 9.5$	$x \leq 5.0$
	3	$x \geq 1000$	$x < 4.0$ or $x > 9.5$	

Table 3. Site condition classifications. If more than five habitat scores were missing, the site could not be appropriately classified and were placed in the Mixed class.

Site Class	Habitat Requirements	Water Quality Requirements
Reference	• ≥ 75% of available habitat scores are ≥ 16 • No habitat scores < 12	• The sum of the assigned water quality scores equals 0
Minimally Degraded	• 66% - < 75% of available habitat scores are ≥ 16	• The sum of the assigned water quality scores equals 0
Mixed	• Does not meet the requirements of the other site classes	• Does not meet the requirements of the other site classes
Moderately Degraded	• ≥ 50% of available habitat scores ≤ 12 (excluded Degraded)	• The sum of the assigned water quality scores is ≤ 1
Degraded	• ≥ 50% of available habitat scores ≤ 6	• The sum of assigned water quality scores is > 1

Table 4. The number of sampling events with less than or equal to seventy individuals identified, aggregated by environmental condition class.

Reference	Minimally Degraded	Mixed	Moderately Degraded	Degraded	Total Count
38	19	729	212	129	1,127

E. **Spatial Classification**

Classifying least-disturbed streams into spatial units with similar features reduces the underlying "noise" in the data analysis and can reveal key relationships between biota and natural factors. Geology, topography, soils, vegetation, slope, and other natural factors affect the structure and function of stream macroinvertebrate assemblages (Kennen 1999, Feminella 2000, Hawkins et al. 2000). For example, taxa in stream assemblages on steep hillsides, with frequent riffles and falls, tend to be more adapted to high flow velocities than those in the flatter valleys or coastal plains. Taxa in karst regions can be more heavily influenced by cooler groundwater. Macroinvertebrates are more likely to disperse along connected stream corridors than across the mountain ridges or other barriers separating major drainage basins (Bilton et al. 2001, Petersen et al. 2004).

The natural landscape of the Chesapeake Bay drainage basin has been classified by hydrologic unit, physiography, and ecoregion. The hydrologic classification system was created by the United States Geological Survey (Seaber et al. 1987). It catalogs surface waters in a hierarchical system, dividing large hydrologic regions into successively smaller units. Hydrologic Unit Codes, or HUCs, indicate the level of classification. Geology and distinct landforms on the earth's surface are the basis for physiographic classifications (Fenneman 1917). The Appalachian Highlands is the largest physiographic region along the east coast of North America, stretching 2,400 km (1,500 mi) from Newfoundland to central Alabama. Four provinces of the Appalachian Highlands contain most of the Chesapeake drainage area: Appalachian Plateau, Valley and Ridge, Blue Ridge, and Piedmont. The other major physiographic region in the Chesapeake drainage is the Atlantic Plain, which lies between the Piedmont province and Atlantic Ocean. Ecoregion, the third classification system, builds on physiographic provinces and considers non-geological factors such as climate, soils, elevation, and vegetation (Omernik 1987, Woods et al. 1999). Ecoregions subdivide physiographic provinces into relatively homogeneous landscapes that support distinct ecosystems.

Indices for three spatial scales were explored: 1) Chesapeake-wide, 2) region, and 3) bioregion. The Chesapeake-wide index used a single suite of macroinvertebrate metrics for the entire basin. The metrics generalize the response of benthic macroinvertebrates to one degradation gradient for the entire basin. The Chesapeake Bay basin is 167,000 km^2, and this spatial resolution may be considered too coarse for index development. However, the National Rivers and Streams Assessment (NRSA) developed indices for geographic areas much larger than the Chesapeake Bay basin, such as the Southern Appalachians, Atlantic Coastal Plains, and Temperate Plains (USEPA 2016a). The feasibility of a single, basin index was explored for reporting purposes.

For the regional spatial scale, the basin was divided into two regions—Coast and Inland. Level III ecoregions 63 (Mid-Atlantic Coast) and 65 (Southeastern Plains) (Woods et al. 1999)

were used to define the Coast region of the basin. The remaining ecoregions located in the Piedmont and Appalachian Highland provinces were aggregated to represent the Inland region. Hydrogeomorphologic differences between these two regions are well known in the literature (Maxted et al. 2000, Klemm et al. 2003, USEPA 2016a). Benthic macroinvertebrate assemblages in these regions are significantly dissimilar (Dail et al. 2013).

For the third spatial scale, the basin was divided into twelve bioregions. These were areas with distinct differences in their natural, undisturbed stream macroinvertebrate assemblages. Bioregion classifications identified in Buchanan et al. (2011) were confirmed or adjusted. Table 5 lists the bioregions in the Chesapeake Bay drainage that were identified; Figure 3 shows their locations. The bioregion scale was the highest spatial resolution used in this report. Indices developed for individual bioregions provide assessments for relatively small geographic areas and identify biological responses specific to disturbances in that area. Appendix G presents in detail how hydrologic unit, physiography, and ecoregion classification approaches were applied to arrive at twelve bioregions.

While bioregion classifications are intended to reflect differences in natural features, they also capture differences in anthropogenic features that can influence stream macroinvertebrate assemblages. We recognized that Reference conditions in one bioregion can differ from Reference conditions in another bioregion despite both bioregions meeting the eight habitat and three water quality criteria. The anthropogenic influences are not necessarily evident in the five environmental condition categories (above). They may include factors such as road density, upstream impervious cover, agricultural contamination of the hydrologically connected zone, groundwater withdrawals for agricultural irrigation that affect baseflow, and nitrogen deposition. The levels of some important natural and anthropogenic features in HUC12 watersheds are shown by bioregion in Appendix H.

Table 5. Hydrologic and physiographic features used to delineate the twelve bioregions of the Chesapeake Bay basin.

Bioregion Code	Bioregion Name	Area (km^2)	Subregion (HUC4)	Additional Distinctions	EPA Level III Ecoregion	EPA Level IV Ecoregion	State(s)
NAPU	Northern Appalachian Plateau and Uplands	24,690			60 Northern Appalachian Plateau and Uplands 83 Eastern Great Lakes and Hudson Lowlands	60a Glaciated Low Plateau 60b Northeastern Uplands 60d Finger Lakes Uplands and Gorges 60e Glaciated Allegheny Hills 83f Mohawk Valley	NY, PA
NCA	North Central Appalachians	10,964			62 North Central Appalachians	62a Pocono High Plateau 62b Low Poconos 62c Glaciated Allegheny High Plateau 62d Unglaciated Allegheny High Plateau	NY, PA
CA	Central Appalachians	5,986			69 Central Appalachians	69a Forested Hills and Mountains 69b Uplands and Valleys of Mixed Land Use	MD, PA, WV
MAC	Middle Atlantic Coastal Plain	14,345			63 Middle Atlantic Coastal Plain	63b Chesapeake-Pamlico Lowlands and Tidal Marshes 63c Swamps and Peatlands 63d Virginian Barrier Islands and Coastal Marshes 63e Mid-Atlantic Flatwoods 63f Delmarva Uplands	DE, MD, VA
SEP	Southeastern Plains	16,464			65 Southeastern Plains	65n Chesapeake Rolling Coastal Plain 65m Rolling Coastal Plain	DC, MD, VA
BLUE	Blue Ridge	5,175			66 Blue Ridge	66a Northern Igneous Ridges 66b Northern Sedimentary and Metasedimentary Ridges	MD, PA, VA, WV
NRV	Northern Ridge and Valley	21,471	Susquehanna		67 Ridge and Valley	67a Northern Limestone/Dolomite Valleys 67b Northern Shale Valleys 67c Northern Sandstone Ridges 67d Northern Dissected Ridges and Knobs 67e Anthracite Subregion	PA
SRV		20,052				67a Northern Limestone/Dolomite Valleys	

Bioregion Code	Bioregion Name	Area (km^2)	Subregion (HUC4)	Additional Distinctions	EPA Level III Ecoregion	EPA Level IV Ecoregion	State(s)
	Southern Ridge and Valley		Potomac and Lower Chesapeake-James		67 Ridge and Valley	67b Northern Shale Valleys 67c Northern Sandstone Ridges 67d Northern Dissected Ridges and Knobs 67f Southern Limestone/Dolomite Valleys & Low Rolling Hills 67g Southern Shale Valleys 67h Southern Sandstone Ridges 67i Southern Dissected Ridges and Knobs	MD, PA, VA, WV
UNP	Upper-Northern Piedmont	12,294	Susquehanna and Upper Chesapeake	Great Valley	64 Northern Piedmont 67 Ridge and Valley	64a Triassic Lowlands 64b Trap Rock and Conglomerate Uplands 64c Piedmont Uplands 64d Piedmont Limestone/Dolomite Lowlands 67a Northern Limestone/Dolomite Valleys 67b Northern Shale Valleys	DE, MD, PA
SGV	Southern Great Valley	8,910	Potomac and Lower Chesapeake-James	Great Valley	67 Ridge and Valley	67a Northern Limestone/Dolomite Valleys 67b Northern Shale Valleys	MD, PA, VA, WV
LNP	Lower-Northern Piedmont	10,989	Potomac and Lower Chesapeake-James		64 Northern Piedmont	64a Triassic Lowlands 64b Trap Rock and Conglomerate Uplands 64c Piedmont Uplands 64d Piedmont Limestone/Dolomite Lowlands	DC, MD, VA
PIED	Piedmont	15,660			45 Piedmont 58 Northeastern Highlands	45c Carolina Slate Belt 45e Northern Inner Piedmont 45f Northern Outer Piedmont 45g Triassic Basins 58 Reading Prong	DC, MD, VA

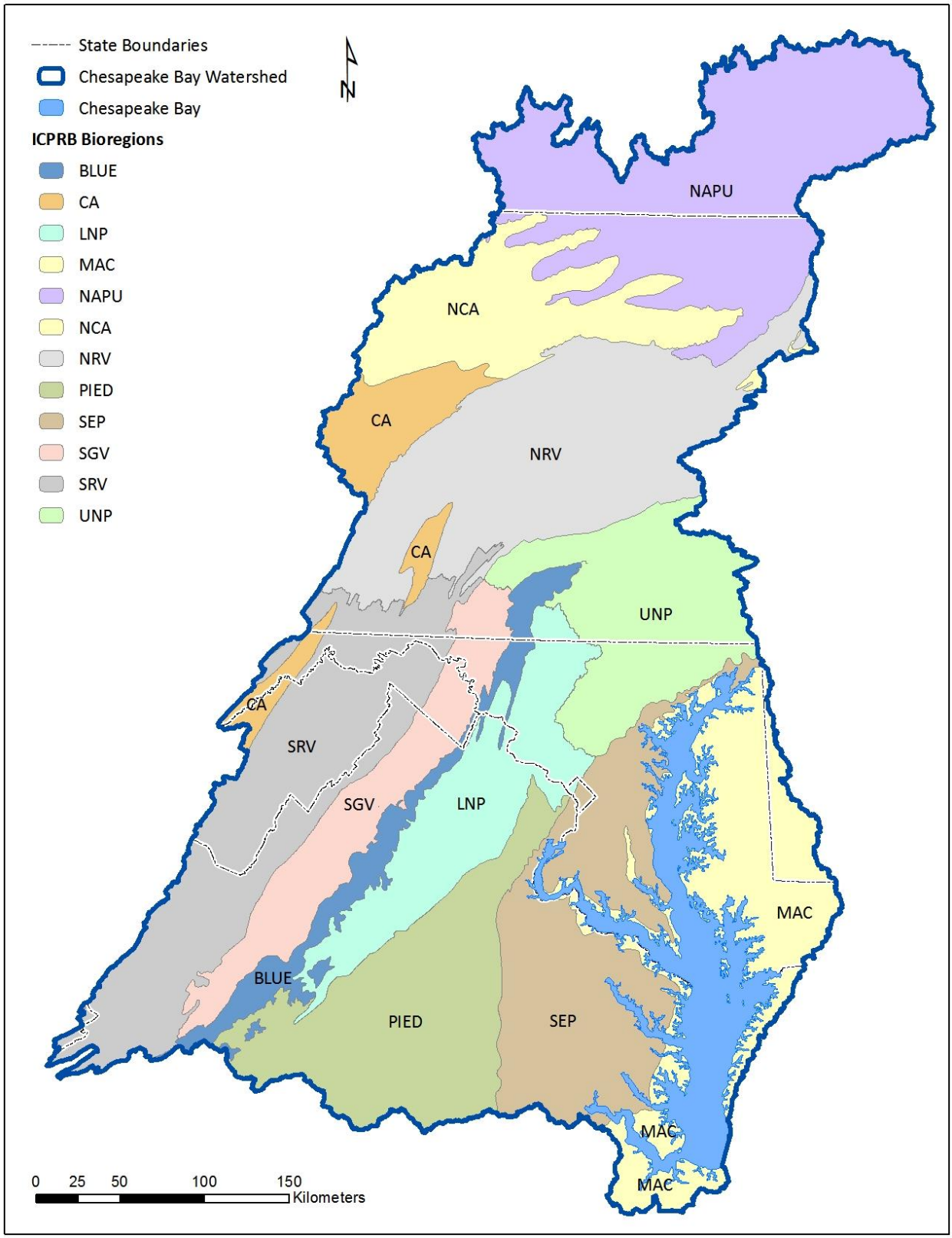


Figure 3. The Chesapeake Bay basin was divided into twelve bioregions for the Chessie BIBI refinement.

F. **Metric Testing**

Biological metrics included in the indices were selected based on their responses to an environmental gradient and their ecological relevance. Evaluations of metric sensitivity, range, variability, and redundancy were conducted for each metric. Metrics which had high range, low variability, were not redundant, and consistently distinguished between environmental condition classes (i.e., sensitive metrics) were considered for the final index. R-statistical language (R Core Team 2016) was used to create functions that automated many of the processes involved in metric selection and index development. Decisions generally made during development were built into the functions or were made into variables easily manipulated with each iteration of the function. Programming the index development process provided rapid, repeatable, and precise results.

i. Metric Sensitivity

Metric sensitivity is the measure of a metric's responsiveness to environmental degradation (Barbour et al. 1999). A metric's Discrimination Efficiency (DE) is often used to quantify metric sensitivity. We developed a new method for measuring metric sensitivity referred to as Balanced Discrimination Efficiency (BDE). BDE is essentially the same as the sensitivity measure used in the 2011 Chessie BIBI report where it was called DE (Buchanan et al. 2011).

DE and BDE are modifications of the Classification Efficiency (CE) equation. CE is a measure used to assess the ability of a multi-metric index to discriminate between Reference and Degraded sites (Equation 1).

Equation 1

$$CE = \left(\frac{\%Ref + \%Deg}{2}\right)$$

Where:

$$\%Ref = \left(\frac{Ref_{correct}}{n_{Ref}}\right) \times 100$$

$$\%Deg = \left(\frac{Deg_{correct}}{n_{Deg}}\right) \times 100$$

$Ref_{correct}$ = the number of Reference samples correctly identified by a threshold.
n_{Ref} = the total number of Reference samples.
$Deg_{correct}$ = the number of Degraded samples correctly identified by a threshold.
n_{Deg} = the total number of Degraded samples.

A threshold is selected to create a binary measure of the index performance. For metrics that decrease with degradation, values greater than or equal to the threshold value are considered to represent a Reference condition and values less than the threshold represent a Degraded condition. The percentage of Reference samples (%Ref) and the percentage of Degraded samples (%Deg) correctly identified by the threshold are calculated, and the mean of %Ref and %Deg provides a measure of the index's ability to correctly classify environmental condition.

The DE measure uses specific percentiles of an individual metric's Reference distribution to establish thresholds for the metric. For metrics that decrease with disturbance, DE uses the 25th percentile of the Reference distribution as a threshold for distinguishing Reference and Degraded samples (Gerritsen et al. 2000). For metrics that increase with disturbance, DE uses the 75th percentile of the Reference distribution. The percentage of Degraded samples correctly identified by the threshold is then calculated using Equation 2, which is equivalent to the *%Deg* formula from Equation 1.

Equation 2

$$DE = \frac{a}{b} \times 100$$

Where:
a = the number of Degraded samples correctly identified by the Reference threshold.
b = the total number of Degraded samples.

During the DE calculation, the percentage of reference sites correctly identified is a static 75% based on the 25th or 75th Reference percentile. If these thresholds were applied to the CE equation (Equation 1), %Ref would always be represented as 75%. Because %Ref is a constant, %Deg is the dynamic factor influencing CE. Therefore, DE simplifies the CE equation to focus on the dynamic factor (i.e., %Deg). The DE methodology provides a simplistic evaluation of metric sensitivity but is prone to classification bias (i.e., DE favors the correct classification of Degraded samples).

The sensitivity measure performed during this assessment is an iterative process, with the objective of finding metric thresholds where %Ref and %Deg are roughly equal. Each Reference percentile was systematically checked as a possible threshold. For each threshold, the percentage of samples correctly identified as Reference and Degraded was measured (Equation 3).

Equation 3

$$B_i = \frac{\%Ref + \%Deg}{2} - |\%Ref - \%\,Deg|$$

B_i was the discrimination efficiency using the i^{th} percentile of the metric's Reference distribution as the threshold. The absolute value of the difference between %Ref and %Deg was used as a balancing factor. Subtracting the balancing factor from the average reduced the probability of selecting a threshold that was biased towards correctly identifying one of the two environmental conditions.

The threshold which produced the maximum B_i was approximately the point that bisected the Reference and Degraded distributions. We refer to this threshold as the metric's Best Separation Point (BSP). The BSP was used as the threshold to calculate the Balanced Discrimination Efficiency (BDE) for each metric (Equation 4). The metrics with the greatest BDE's were considered as candidates for the final index. The BDE equation is effectively the same as Equation 1 for CE.

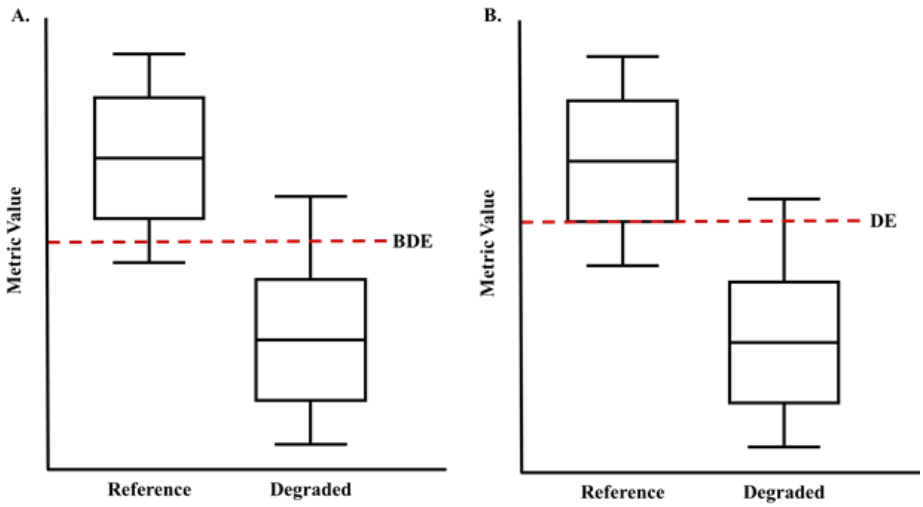


Figure 4. Balanced Discrimination Efficiency (BDE) generally measures metric sensitivity at a different threshold than Discrimination Efficiency (DE). The figures depict metrics that decrease with disturbance. BDE is based on the Best Separation Point (BSP), the point at which the percentage of Reference and Degraded samples correctly identified are approximately equal (Figure 4A). DE is measured based on a standard threshold defined by the Reference distributions 25th percentile (Figure 4B).

Equation 4

$$BDE = \left(\frac{\%REF_{BSP} + \%DEG_{BSP}}{2}\right)$$

Where:

$$\%Ref_{BSP} = \left(\frac{Ref_{correct}}{n_{Ref}}\right) \times 100$$

$$\%Deg_{BSP} = \left(\frac{Deg_{correct}}{n_{Deg}}\right) \times 100$$

$Ref_{correct}$ = the number of Reference samples correctly identified by the BSP threshold.
n_{Ref} = the total number of Reference samples.
$Deg_{correct}$ = the number of Degraded samples correctly identified by the BSP threshold.
n_{Deg} = the total number of Degraded samples.
BSP = indicates that the Best Separation Point was used as the threshold for discerning Reference and Degraded samples.

%Ref and %Deg were dynamic factors in the BDE equation (Figure 4 A), providing a more specific assessment of a metrics ability to discriminate than the standard DE method (Figure 4 B). Additionally, the BSP is used in the scoring procedure (See *II. G. Metric Scoring Approach*) providing continuity between metric sensitivity and metric scoring.

ii. Range and Variability

The selected metrics should respond to environmental degradation and not to variability in the data (Barbour et al. 1999). Setting standards for metric range and variability can protect against overfitting the index. Only Reference samples were used to assess metric range and variability. Blocksom and Johnson (2009) calculated range as the difference between the maximum metric value and the minimum metric value. To avoid the influence of outliers we calculated range as the difference between the Reference 95th percentile and the Reference 5th percentile. Table 6 summarizes range requirements specified for the metrics assessed in the analysis. Selecting metrics with low range restricts the Reference criteria beyond expected natural variability and in effect creates a high probability for false-negatives.

Table 6. The metrics selected for the final indices were required to meet specified Reference distribution range requirements.

Metrics	**Range Requirement**
Simpson, Pielou, and Hurlbert's PIE	≥ 0.1
Shannon, Menhinick, and Margalef	≥ 1.0
HBI and ASPT	≥ 2.0
Richness metrics and variations of Beck's Index	≥ 3.0
Percent metrics	≥ 10.0

Measuring variability acts as a counter measure to range. Preferably metrics with high range and low variability are selected for further analysis. Variability was measured as the range of the Reference interquartiles relative to the range between 0 and the reference 25th percentile (Blocksom and Johnson 2009). Metrics were selected if the ratio of the interquartile range relative to the range between 0 to the Reference 25th percentile was less than 3. Blocksom and Johnson (2009) recommended a ratio of less than 1 but this standard too frequently eliminated sensitive metrics.

iii. Metric Redundancy

Spearman correlation was used to assess metric redundancy. Including two highly correlated metrics in the final index is analogous to doubling the weight of a single metric in the index. The final biological index is composed of multiple metrics that are not strongly correlated and each metric evaluates the response of different aspects of the biological assemblage to disturbance. A correlation coefficient of 0.85 ($r \geq 0.85$ or $r \leq -0.85$) was selected for this study. A coefficient of 0.85 is a relatively high correlation coefficient but it has been used in other indices (Gerritsen et al. 2000, Butcher et al. 2003) and indicates ~72% of paired metric values have a positive or negative relationship. Redundant metrics ($r \geq 0.85$ or $r \leq -0.85$) were compared to determine which metric showed greater separation between the Reference and Degraded distributions. The metric with the lower *p*-value was retained. The metrics remaining after the metric redundancy assessment were considered for the final index.

G. **Metric Scoring Approach**

Metrics are often scored using a continuous range between two thresholds that represent "floor" and "ceiling" values (Minns et al. 1994, Hughes et al. 1998, Blocksom 2003, Pond et al. 2011). Buchanan et al. (2011) developed a scoring approach that relied on finding the Best Separation Point (BSP) between the Reference and Degraded distributions. They used the BSP and the median of the Reference distribution as the "floor" and "ceiling," respectively, for the scoring gradient. Metric values between the BSP and Reference median thresholds were scored on a continuous gradient ranging from 0 - 100; values outside the range were scored 0 or 100, depending on the direction of change with disturbance. The range between the BSP and Reference median was often small, with few metric values falling on the gradient, and many of metrics scored in a binary (i.e., 0 or 100) rather than a continuous (i.e., 0 - 100) manner.

The Buchanan et al. (2011) scoring approach was modified in this report to expand the range of values that could be scored on the continuous gradient. The BSP was established as the midpoint (i.e., 50) of the continuous gradient and the Reference median (X_M) was established as the ceiling (i.e., 100). Metrics that decrease with disturbance received a score of 100 if the metric value was greater than or equal to the X_M threshold (Figure 5A). Metrics that increase with disturbance received a score of 100 if the metric value was less than or equal to the X_M threshold (Figure 5B). To identify the floor of the continuous gradient (X_T) for metrics that decrease with degradation, the difference between X_M and the BSP (*a*) was subtracted from the BSP (Figure 5A). Metric values less than the value of X_T received a score of zero. For metrics that increase with degradation, the floor ($X_{T'}$) was established by adding *a* to the BSP (Figure 5B). Metric values greater than the value of $X_{T'}$ received a score of zero. On rare occasions, the calculated values of the thresholds X_T and X_T' for percentage metrics were less than 0% or greater than 100%, respectively. Since percentage metrics of a sample cannot fall below 0% or exceed 100%, the metric values of these thresholds were adjusted to 0 or 100, respectively.

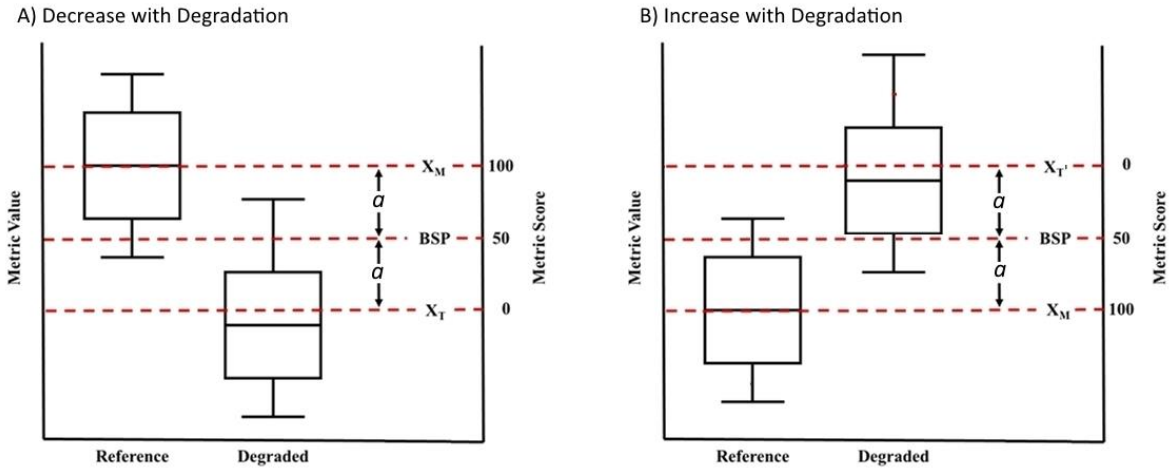


Figure 5. Metric scoring procedure. For metric values that decrease with degradation (Figure 5A), the Reference median (X_M) is the ceiling and X_T is the floor of the 0 – 100 point gradient. Values greater than X_M receive a score of 100; values less than X_T receive a score of 0; values in-between X_M and X_T are scored proportionally. For metric values that increase with degradation (Figure 5B), the Reference median (X_M) is the ceiling and X_T' is the floor of the 0 – 100 point gradient. Values less than X_M receive a score of 100; values greater than X_T receive a score of 0; values in-between X_M and X_T' are scored proportionally. BSP, Best Separation Point between Reference and Degraded metric values.

Metric values falling between the floor and ceiling thresholds were scored proportionally to the range of values between the thresholds. Equation 5 and Equation 6 were used for metrics that decrease with disturbance and metrics that increase with disturbance, respectively:

Equation 5

$$Score = \frac{X - X_T}{X_M - X_T} \times 100$$

Where:
X = the metric value.
X_T = the lower threshold (i.e. floor).
X_M = the upper threshold (i.e. ceiling).

Equation 6

$$Score = \frac{X_{T'} - X}{X_{T'} - X_M} \times 100$$

Where:
X = the metric value.
$X_{T'}$ = the upper threshold (i.e. ceiling).
X_M = the lower threshold (i.e. floor).

H. **Index Construction**

Eight ways of constructing a multi-metric index from family-level metrics were examined (Appendix I). The purpose of the exercise was to gain insights into the consequences of choosing a specific index structure or combination of metrics. In one extreme, only the single most sensitive metric from each of five metric categories (richness/diversity, tolerance, functional feeding group, habit, and composition) was incorporated into the index. In another extreme, over 200 metrics were included. Zero inflation protection was applied in some methods and not in others. Zeros can be abundant in biological data and large proportion of zeros has the potential to mask relationships between environmental and biological data (McCune et al. 2002). For three methods, several metrics in each of the five metric categories were pre-selected or selected as most sensitive by the R-program and their scores averaged; subsequently, the five category averages were averaged to obtain a final index score. All eight methods were tested in the twelve new bioregions. We concluded that the development strategy closest to strategies described in the literature was the most practical (Method A in Appendix I).

Order, family, and genus-level indices were developed for each of the three spatial scales: Chesapeake-wide, region, and bioregion. For the family- and genus-level indices, metrics remaining after the range, variability, and redundancy checks were divided into the five metric categories and the four most sensitive metrics of the five categories were selected to be the foundation of the index using the metric testing and scoring approaches described above. Therefore, one category was not represented initially. Additional sensitive metrics were then added to the index only if they improved the index CE. Metrics omitted during the selection of the top four metrics were readmitted if the metric improved the index CE.

Functional feeding group (FFG), habit, and some tolerance metrics are inappropriate at the order-level. Thus, only richness/diversity, composition, and a subset of tolerance metrics were considered for the order-level version of the indices. When applicable, one richness/diversity metric, one composition metric, and one tolerance metric were required for order-level indices. After conducting range, variability, and redundancy checks, it was possible to have a very small set of potential metrics that represented a single metric type, and therefore, it was not possible to represent the three metric types. Additional sensitive metrics were then included if they improved index CE.

Within the Chessie BIBI database there are sampling stations that were revisited multiple times during a year or revisited during subsequent years. The scores of multiple sampling events for a station were not averaged during index development. Approximately 2,154 stations (10.1%) in the analysis dataset have between 2 and 35 sampling events each. The index scores for the sampling events at one station have the potential to vary considerably, and narrative ratings (described below) at some stations occasionally ranged across all five rating categories, from Excellent to Very Poor. At some stations, variation was associated with changes in stream water quality and habitat conditions; in others, it appeared to be natural inter-annual variability. It was assumed that each sampling event represents the biological response to the immediate environment and not the previous year's biological status.

However, treating these repeat-visit sampling events as independent samples could result in pseudoreplication. Sampling events collected from the same sampling stations at different times have a higher probability of not being statistically independent. For example, two sampling events collected from the same sampling station two years apart have a higher probability of representing a more similar macroinvertebrate community than two sampling events collected at different sampling stations. This pseudoreplication could create a bias towards the community found at repeatedly sampled stations. To minimize the potential influence of pseudoreplication, each sampling station was represented by a single sampling event. If multiple sampling events were recorded for a single sampling station, then one sampling event was selected at random. This process should have eliminated pseudoreplicates at a given station, but pseudoreplication may still be an issue in this dataset. In some instances, multiple agencies collected samples from the same sampling station but assigned different station names. Currently, sampling stations with two or more station names are difficult to find in the database. Additionally, some of the data represents intensively sampled streams, where the sampling stations are different but the distance between these stations is small. In these cases, the probability of these sampling events representing statistically independent samples is low. Although this situation appears to represent pseudoreplication it may be more appropriate to refer to this a spatial autocorrelation. It was assumed that the occurrence of these pseudoreplicates and spatially autocorrelated sampling events was infrequent but GIS assessment in future refinements of the Chessie BIBI may be able to identify and manage these potential issues. As the sample size increases, the influence of relatively small set of pseudoreplicates or spatially autocorrelated sampling events should diminish because each sampling event has less weight and the number of independent samples represents a majority of the data. We contend that the large sample size of the Chessie BIBI dataset and the process described above for managing sampling station pseudoreplicates minimizes any potential bias associated with pseudoreplication.

Metrics were scored using the Balanced Discrimination Efficiency (BDE) approach described in *II. F. G. Metric Scoring Approach*. This method is compared to three others in Appendix J. Each index produces different scores but their ability to correctly classify Reference and Degraded sites (classification efficiency) is essentially the same. The Balanced Discrimination Efficiency method was best at spreading index scores across the entire 0 – 100 continuous scoring scale.

Sampling events and sampling stations were not evenly distributed throughout the Chesapeake Bay basin. Maryland had broad coverage and a high density of sampling events, in part, due to their Stream Waders volunteer program. The high density of sampling events in certain areas, such as Maryland, may create indices biased towards the conditions of that area. To reduce the potential bias when creating the regional and basin-wide indices, fifty Reference and fifty Degraded sampling events were randomly selected from each bioregion. If the bioregion had less than fifty Reference or less than fifty Degraded sampling events, then all the sampling events where retained during the assessment. Reducing the Reference and Degraded sample sizes to a maximum of fifty reduced the potential for spatial bias, e.g., a basin-wide or regional index developed with one bioregion containing two or three times the number of Reference sampling events relative to the other bioregions. After the bioregions were subset to a maximum of fifty Reference and fifty Degraded sampling events, they were aggregated to the appropriate basin-wide or region spatial resolutions.

In repeated runs of the R-program scripts to select and test metrics for the bioregion and region indices, we noticed that random choices made early in the data preparation's probabilistic rarefaction step affected the metric selections and scoring thresholds. The probabilistic rarefaction process reduces the variability inherent in the general rarefaction process; however, in each run it randomly selects from equally rare taxa to make up a sample count of about one-hundred. Slight differences in the results influence metric redundancy, range, variability, and sensitivity, and ultimately affect which metrics are selected (see Appendix D). The richness/diversity metrics appear to be the only metrics with the potential to change scoring thresholds with each run due to the probabilistic rarefaction process. Additionally, with each new run of the R-program scripts the random selection of a single sampling event to represent each station and/or the random sub-setting of Reference and Degraded, described in the previous paragraphs, added additional variability to the metrics selected for the final index and the associated scoring thresholds.

To reduce the potential error created by running the program only once, an iterative development process was adopted. R-program scripts were developed to automate the entire process associated with index development: the random selection of a single sampling event to represent each sampling station, if applicable a maximum of fifty Reference and fifty Degraded samples selected at random from each bioregion, metric calculations, metric range test, metric variability test, metric sensitivity test, metric redundancy test, metric selection (based on metric type, range, variability, sensitivity, and redundancy), metric scoring, and averaging the scores into the final index score. We then identified the metrics that occurred most often in fifty runs and incorporated them into the final indices. Metrics occurring in 20% or more of the fifty runs were included in the final indices. If fewer than five metrics occurred in 20% or more of the runs, the metrics were ranked in descending order and the frequency of the fifth ranked metric was used as the new frequency threshold to select metrics. The selected metrics were subjected to a final metric redundancy assessment using all the available data. Any metrics identified as

redundant were removed following the procedure outlined in *II. F. iii. Metric Redundancy*. The means of the metric scoring thresholds calculated in the fifty runs were used as the metric scoring thresholds in the final indices. This iterative process provides more robust indices that are less susceptible to overfitting the indices and more sensitive to the underlying biological patterns associated with the defined disturbance gradient.

To calculate the index score of a sampling event, the metrics corresponding to the selected spatial (basin-wide, region, bioregion) and taxonomic (order, family, genus) index are scored and averaged.

I. **Index Classification Efficiency**

The ability of the index to correctly classify Reference and Degraded sites is tested in a manner similarly to how individual metrics are tested. The BSP for the index was determined from distributions of the Reference and Degraded index scores using Equation 3. CE was calculated using the BSP as the threshold separating Reference and Degraded index values (Equation 1). Since the percentage of correctly identified Reference samples is approximately equal to the percentage of correctly identified Degraded samples at the index's BSP, the BSP provides a more accurate representation of CE, as opposed to an arbitrarily selected Reference percentile value or index value (i.e., 50 on a scale of 0 – 100).

J. **Taxonomic Tiers**

Benthic macroinvertebrate indices of biotic integrity are typically developed for a single taxonomic resolution (e.g., family-level). However, this limits the accessibility and/or applicability of these indices to some monitoring programs that operate within the Chesapeake Bay basin. Order-, family-, and genus-level indices were created for the basin-wide index, the two region indices (Inland and Coast) and each of the twelve bioregions. The order-level provides a coarse assessment but can be easily used by volunteer groups or programs with limited funding and/or little experience identifying macroinvertebrates. Metrics that required assigned taxonomic attributes (e.g., FFG, habit, and tolerance value metrics) were excluded from the order-level analysis. The family-level indices will be applicable to monitoring programs with moderate amounts of funding and experience identifying macroinvertebrates. The genus-level index is appropriate for monitoring programs with staff certified in taxonomic identification. For all indices, we required a minimum of 90% of taxa to be identified to the corresponding taxonomic resolution. We did this because samples which include taxa identified to a resolution lower than specified index resolution are susceptible to under-representation in richness and diversity metrics and overly coarse taxonomic attribute assignments.

K. **Delete-d Jackknife Validation**

Sensitive metrics are assumed to reflect ecological response to an environmental gradient. However, each multi-metric index is susceptible to overfitting of the data (Barbour et al. 1996). In this case, overfitting refers to the selection of metric(s) that appear to reflect an ecological response to the disturbance gradient but in fact reflect random variability or nuances of the data set used to construct the index. Validation procedures verify that the index measures an ecological response to a defined gradient, and thus, protects against overfitting.

In general, validation requires the data set to be divided into a training set and a validation set prior to index development (Southerland et al. 2005, Pond et al. 2011). The

training set is used to develop the index, while the validation set is used to verify that the index classifies data appropriately. When sample size is small, it may not be possible to set aside an independent dataset for validation purposes (Hawkins 2004). In such instances, Cross Validation (CV) can be used to create and validate an index with the same dataset. The Delete-d Jackknife CV procedure was used to validate each bioregion index. Buchanan et al. (2011) referred to this method as a jackknife with replacement but this is more frequently referred to as a Delete-d Jackknife. This is an iterative process creating a unique training dataset and validation dataset with each iteration. For each iteration, *d* samples are removed from the dataset to form a validation dataset; the remaining samples constitute the training dataset. A true Delete-d Jackknife removes *d* samples and re-computes the final value (e.g., mean, median, or CE) for each possible data combination. This quickly becomes computationally impossible for the average desktop computer. For example, with a sample size of 100 and *d* equal to 25 there are greater than 2.4×10^{23} possible combinations. Therefore, five-hundred unique Delete-d Jackknife combinations were used in this study as an estimate of the results of all the possible combinations.

A portion of the samples (*d*) in the Reference and Degraded populations were randomly removed. Shao and Wu (1989) recommend that *d* should be greater than the square root of *n* but less than *n* ($\sqrt{n} < d <$ n). Buchanan et al. (2011) removed 10% of the reference population during CV but 10% of any bioregion with one-hundred or fewer reference samples would produce a *d* value lower than the recommended range. Therefore, we set *d* equal to 25% of the Reference population. Additionally, 25% of the Degraded population was also removed during the CV procedure because the Degraded distribution in combination with the Reference distribution influences metric scoring thresholds. The removal of 25% of Reference and 25% of Degraded samples placed *d* well within the range recommended by Shao and Wu (1989) for all indices.

All the available and applicable data within each spatial resolution was used to develop the appropriate Chesapeake-wide, region, or bioregion specific indices. A Delete-d Jackknife CV was used to verify that the index was not overfit to the data. Five-hundred CV iterations were conducted. With each iteration of the CV process the index was reconstructed with a unique training set and CE was checked using the corresponding, independent validation set. The CV tests utilized only the metrics selected when using all the data to build the index. The goal of this process was to test the validity of the original index. Therefore, allowing the program to deviate from the metrics originally selected using all the data would not address the accuracy of the original index. CE of the validation set calculated with each iteration was used to calculate the expected CE and RMSE. Mean simulated CE was the average CE of all iterations ($\hat{\theta}_{(.)}$). RMSE provides a measure of standard deviation associated with the expected CE (Equation 7).

Equation 7

$$RMSE = \sqrt{\frac{\sum_i (\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2}{n}}$$

Where:
$\hat{\theta}_{(i)}$ = the estimated CE for each iteration.

$\hat{\theta}_{(.)}$ = the mean simulated CE from *i* iterations.
n = the number iterations.

The CV method described above is an iterative modification of the validation process typically found in the literature. Instead of parsing the data into training and validation sets prior to development, the index was developed using all the data and post-development the data was iteratively parsed into training and validation data sets. Although both methodologies verify that the index reflects ecological responses to an environmental gradient, the CV method should provide a more robust assessment because the validation process was repeated five-hundred times.

L. **Delete-d Jackknife Precision**

After the final index had been established, a threshold was calculated to find the Best Separation Point (BSP) between the Reference and Degraded distributions using the BDE equation (Equation 3). A Delete-d Jackknife was used to measure variation of the BSP and the associated CE. The indices were constructed using all the available data and will be referred to as the original indices. To assess precision, the metrics selected for the original index were used to iteratively create new indices based on subsets of the available data. Twenty-five percent of the Reference population and 25% of the Degraded population were randomly removed to create each unique subset. During each iteration, the metrics were scored and used to create a new index. The process was repeated five-hundred times. RMSE (Shao 1989) was calculated for the BSP and CE of the five-hundred iterations. The RMSE indicated the variability associated with the measures of interest (Equation 8). Shao (1989) provided the Mean Square Error (MSE) formula for a delete-d jackknife and the square root of this formula was used to calculate RMSE.

Equation 8

$$RMSE = \sqrt{\frac{n-d}{d(N)}\sum_{i}(\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2}$$

Where:
n = the number of reference samples.
d = the number of reference samples removed for each iteration.
N = the number of iterations.
$\hat{\theta}_{(i)}$ = the estimated threshold or CE for each iteration.
$\hat{\theta}_{(.)}$ = the mean estimated threshold or CE from *i* iterations.

Delete-d Jackknife was used to test the precision of the BSP and CE. This was not a CV procedure because a validation dataset (i.e., an independent dataset) was not utilized during the assessment. Five-hundred subsets of the data (i.e., training dataset) were used to provide an estimate of precision.

The scoring thresholds are determined by the data used to construct the index. Therefore, different datasets from the same geographic area would be expected to generate different scoring thresholds. The variability associated with the scoring thresholds attests to the robustness of the metrics. Low variability suggests that the scoring thresholds are repeatable and most likely

indicative of stream condition; however, high variability suggests that the scoring thresholds reflect random noise in the data and may not be robust measures of stream condition. Estimating precision of the indices BSPs and CEs provided a measure for which we could judge index performance.

M. **Narrative Rating Categories**

The numeric thresholds of the rating scale used in the 2011 Chessie BIBI (Buchanan et al. 2011) could not be applied to the family-level index scores for the twelve new bioregions because refinements in the metric scoring procedure to improve resolution caused an overall increase in the Reference index scores. Preliminary results in this study indicate that using the same threshold values for rating the index scores of the twelve new bioregions could create bioregion biases. Although each bioregion index is scored on a standard 0 - 100 continuous scale, a score from one index may not be directly comparable to a score in another index (Pond et al. 2011). The influence of natural and anthropogenic factors become more pronounced at the smaller bioregion spatial scales (see above), and using the same thresholds will penalize bioregions that score low for causes that were not accounted for when Reference condition criteria are applied (e.g., Figures L-1 and L-2). For each bioregion, index scores were rated according to their individual, bioregion-specific Reference distributions. For the same reasons, separate narrative rating scales based on percentiles of the Reference and Degraded index scores were developed for the Coast and Inland region indices.

The order-, family- and genus-level indices were rated on a 5-category scale based on the Reference distribution without outliers. Outliers were Reference index scores outside of 1.5 times the Reference interquartile range. After removing the outliers, the thresholds were derived from the 50th, 25th, and 10th percentiles of Reference and half of the value of the 10th percentile (Table 5). Scores equal to the rating thresholds were always categorized as better of the two ratings. For example, a sampling event with a score equal to the rating threshold between Poor and Fair would receive a Fair rating.

Table 5. Index scores used as thresholds to separate the narrative rating categories for the order-, family- and genus-level index scores. Thresholds were defined by the 50th, 25th, and 10th Reference percentiles and half of the value of the 10th percentile.

Narrative Rating Threshold:		Very Poor | Poor	Poor | Fair	Fair | Good	Good | Excellent
Watershed-wide Index					
Chesapeake	Order	19.60	39.20	58.20	73.90
Chesapeake	Family	22.60	45.20	64.40	79.90
Chesapeake	Genus	19.90	39.80	63.00	80.70
Region Indices					
Coast	Order	13.63	27.25	42.44	67.70
Coast	Family	18.39	36.78	52.85	78.43
Coast	Genus	22.05	44.09	55.30	65.92
Inland	Order	21.73	43.46	57.55	68.64
Inland	Family	23.15	46.29	62.28	77.78
Inland	Genus	20.27	40.54	56.62	71.61

Narrative Rating Threshold:		Very Poor | Poor	Poor | Fair	Fair | Good	Good | Excellent
Bioregion Indices					
BLUE	Order	18.90	37.70	58.20	75.00
BLUE	Family	30.90	61.70	82.20	92.10
BLUE	Genus	23.90	47.80	57.10	72.10
CA	Order	26.10	52.20	56.90	67.90
CA	Family	27.90	55.80	69.50	78.60
CA	Genus	30.20	60.50	65.70	77.80
LNP	Order	31.40	62.80	69.60	81.90
LNP	Family	35.20	70.30	80.20	91.30
LNP	Genus	28.80	57.60	68.80	80.70
MAC	Order	29.20	58.50	62.60	74.60
MAC	Family	22.70	45.40	63.20	76.80
MAC	Genus	21.60	43.30	64.50	78.40
NAPU	Order	10.70	21.40	33.30	55.40
NAPU	Family	17.90	35.90	47.40	61.90
NAPU	Genus	19.00	37.90	54.00	68.50
NCA	Order	18.30	36.70	50.30	65.50
NCA	Family	15.70	31.50	56.30	78.70
NCA	Genus	31.30	62.60	75.00	85.90
NRV	Order	19.50	39.10	54.20	69.70
NRV	Family	19.10	38.20	50.80	75.00
NRV	Genus	24.50	48.90	66.70	79.90
PIED	Order	10.00	20.10	42.50	70.60
PIED	Family	30.30	60.60	73.20	81.80
PIED	Genus	24.00	48.00	59.80	69.10
SEP	Order	23.80	47.50	54.60	67.30
SEP	Family	16.30	32.60	56.70	83.90
SEP	Genus	23.20	46.30	59.00	74.80
SGV	Order	21.90	43.70	54.60	68.70
SGV	Family	29.90	59.90	66.00	76.50
SGV	Genus	24.00	47.90	56.60	67.50
SRV	Order	19.50	39.10	53.50	65.60
SRV	Family	21.60	43.20	58.00	71.90
SRV	Genus	18.80	37.50	53.90	68.60
UNP	Order	17.20	34.50	58.80	73.70
UNP	Family	31.30	62.60	69.90	80.50
UNP	Genus	30.70	61.40	73.10	81.70

N. **Area-Weighting of Rating Results**

Several data processing steps were done to prepare area-weighted estimates of the percentages represented by each rating. All the data were used to map and compare ratings in this report, but most of the 21,552 samples in the final analysis database were collected in the 12-year period between 2000 and 2011 (Figure 6). This period dominates the results. The index scores of stations with multiple samples were averaged and the average rated to avoid giving any location disproportionate importance. The 21,552 sampling events in the analysis database condensed to 12,922 stations represented by a single sampling event and 2,154 stations represented by multiple (2-35) sampling events for a total of 15,146 stations. A check of unique locations identified by their latitude and longitude found 184 locations that shared 2 - 5 stations with different names. These cases are difficult to detect so we let them stand as unique stations.

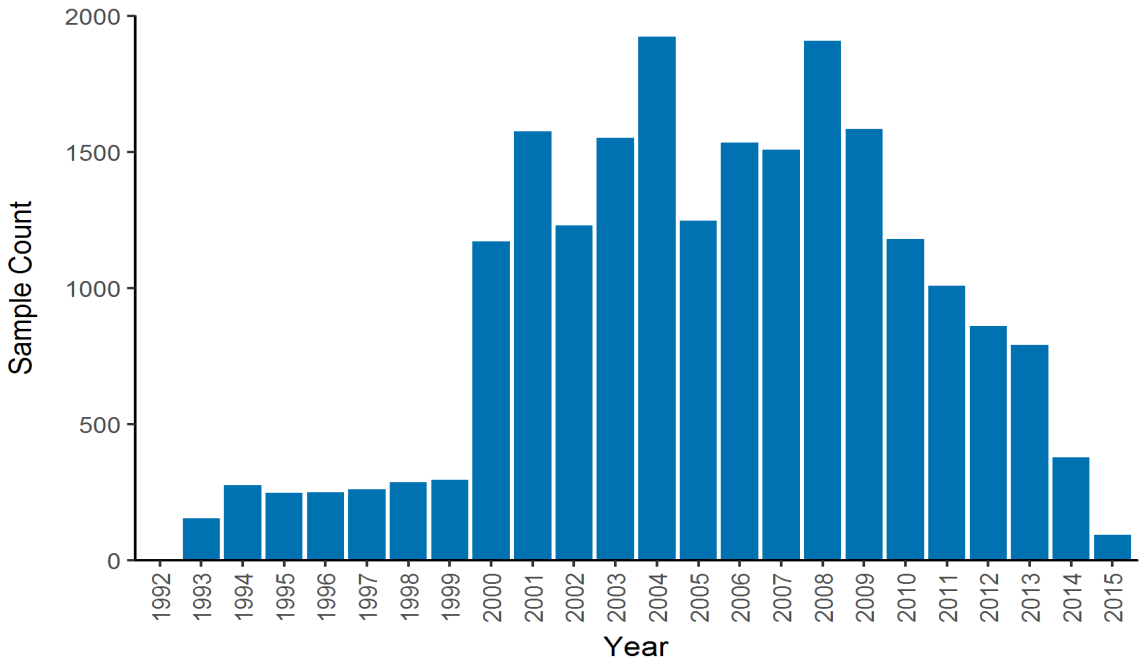


Figure 6. Sample count by year in the analysis database after the data preparation steps described in Methods were applied.

Randomly or systematically located stations are best suited for estimating the percentages of streams that can be statistically expected in each of the five rating categories. Monitoring programs usually indicate in their data sets or documentation how they selected their monitoring locations, and we incorporated this information into the stream database. In the analysis database, 3,689 (24.4%) of the 15,146 stations are currently listed as targeted and not random or systematic. This seems high given the number of monitoring programs that use random-stratified sampling designs or revisit stations that were first selected randomly. We believe there are inaccuracies in how some stations are classified in our database, and the true number of targeted stations is lower. For example, 2,981 (80.8%) of the 3,689 stations have only one sample and come from monitoring programs that typically do not target sampling locations. Thirty more stations are first listed as targeted and in subsequent years listed as random/systematic. These inconsistencies suggest that only 678 (4.5%) of the stations in the

analysis database may be targeted sites. We decided to include all stations in the analysis database at this time, pending further investigation.

Furthermore, combining data from multiple random sampling designs is not equivalent to a single, basin-wide random sampling design. Using a basin-wide random sampling design, each sampling location would have an equal probability of being selected. The random sampling design applied by each agency in the Chessie BIBI database were limited by the agencies study area (e.g., PADEP was limited to PA) and the number of samples collected by the agency. Different areas in the Chesapeake Bay watershed have different random sample densities, and therefore, each sampling location did not have an equal probability of being selected.

To avoid giving heavily sampled areas an unfair weight, the rating for each station's index score or average score was area-weighted. We used HUC12s as the basis for area-weighting because they are relatively small (10.7 – 197.1 km^2) and homogeneous (67.8% fall entirely within the same bioregion). When a HUC12 overlapped two or more bioregions, the areas of each bioregion within the HUC12 were used to area-weight their respective scores (Table H-3). Sampling stations were aggregated first by high resolution catchments ("Retrieved [03/15/2017], from the ecosheds product downloads, http://ecosheds.org/assets/nhdhrd/v2/" n.d.). This reduced the potential for spatial autocorrelation caused by stations located close together with different names. The mean value of all station scores within each catchment was calculated and assigned a rating. Each catchment rating was then multiplied by the appropriate factor for its HUC12-bioregion unit:

Equation 9

$$\text{Factor} = \frac{\text{the area of HUC12} - \text{Bioregion unit}}{\text{the number of catchments in HUC12} - \text{Bioregion unit}}$$

All the weighted ratings for each rating category (Excellent, Good, Fair, Poor, and Very Poor) in a bioregion are summed and the sum divided by bioregion total area to obtain %Excellent, %Good, etc.

For Coast and Inland, the two regional indices, station ratings are similarly weighted by how much of their respective HUC12-region unit they represent. Most HUC12s (1,922) fall entirely within one region; only 59 overlap the Coast-Inland boundary. The weighted ratings for each rating category are summed and the sum divided by the region's total area to estimate the percent of streams in that category in the region. The weighted ratings can be rolled up to basin.

III. Results

Of the 25,067 sampling events included in the stream macroinvertebrate database, 21,314 remained after applying the data preparation steps. The majority of sampling events in all bioregions represent intermediate environmental conditions and were not used to develop the indices (i.e., Minimally Degraded, Mixed, and Moderately Degraded). The Mixed category contains samples with insufficient water quality and stream habitat data. For this study, 1,866 Reference and 1,323 Degraded samples were identified and used to develop the order- and family-level versions of the Chesapeake-wide index, the Coast and Inland indices, and the twelve bioregion indices. Even fewer samples—1,587 Reference and 1,228 Degraded samples—were

used to develop the genus-level version of the indices because we required a minimum of 90% of taxa reported in a sample to be identified to the genus-level during laboratory enumerations.

A. **Index Construction**

Results of the index construction comparisons (Appendix I) suggest that index development strategy has a minor influence on index CE. However, it has a major influence on the distributions of Reference and Degraded index scores and consequently on the thresholds for scoring metric values. None of the strategies consistently outperformed the others in terms of sensitivity or variability. This appears to be due in large part to a counter-balancing effect in the metric scoring process. For example, metrics in a Reference-quality sample that score low are typically countered by a larger number of metrics that score high, resulting in an overall high index score that often classifies the sample correctly as Reference-like. The key is to have sufficient numbers of sensitive metrics so the counter-balancing effect can occur. Our results suggest a minimum of five metrics is sufficient to achieve this counter-balancing effect.

Zero inflation influenced the distributions of Reference and Degraded index scores. Zero inflation occurs when abundance values for a given taxon are dominated by zeros, i.e. the taxon is rare (McCune et al. 2002). This effectively masks any underlying differences in abundance between Reference and Degraded conditions when the taxon occurs. Metric range, variability, and sensitivity criteria tended to protect against the zero inflation issue.

For the Chessie BIBI refinement, we used the index construction strategy that was most similar to strategies described in the literature (Method A in Appendix I). Multi-metric stream macroinvertebrate indices in the literature often include at least one metric in each of the five metric categories, which is thought to provide a holistic evaluation of the macroinvertebrate community. For the family- and genus-level indices in this study, we first ensured that four of the five metric categories were represented by their most sensitive metric. Additional metrics were then added without regard to category if they improved the index's overall CE. For order-level indices, the most sensitive metric in each of the richness/diversity, composition, and tolerance categories were ensured, and additional metrics were included if they improved CE. Functional feeding group (FFG), habit, and some tolerance metrics are inappropriate to use at the order-level.

Composition, richness/diversity, and tolerance metrics generally had the highest discrimination efficiencies, reflecting strong responses to degradation. Habit and FFG metrics often had the lowest discrimination efficiencies, but in some instances, they could improve index CE by surprising amounts when included in the indices. Family- and genus-level versions of the region and bioregion indices generally contained more than ten metrics. Order-level indices generally contained fewer metrics, ranging from three metrics (PIED) to eight (SRV).

B. **Chesapeake-Wide Index**

Following the iterative process outlined in *II. H. Index Construction*, Chesapeake-wide indices were developed for the three taxonomic resolutions. Each taxonomic version of the index had relatively high classification efficiencies (CEs): 81.5% for order-level, 84.9% for family-level, and 85.0% for genus-level (Table K-2). Five, six, and six metrics were included in the order-, family-, and genus-level versions, respectively. The most frequently selected metrics (Table 6) were used to build the index. Despite efforts to minimize spatial bias, reference sampling events, on the Inland side of the watershed, had an uneven influence in shaping the index because the

number of Reference and Degraded sampling events in the Inland region were 4.4 times more numerous than in the Coast region (

Table *7*). We concluded this spatial resolution was biased and misrepresented the Coast region; therefore, this index was not evaluated further.

Table 6. Metrics included in the order-, family-, and genus-level versions of the single Chesapeake-wide index.

Name	Metric category	Order-level	Family-level	Genus-level
PCT_COTE	Composition			X
PCT_DIPTERA	Composition	X		
PCT_EPHEMEROPTERA	Composition	X	X	
PCT_EPHEMEROPTERA_NO_BAETID	Composition			X
PCT_EPT	Composition	X	X	
PCT_HEXAPODA	Composition	X		
PCT_PISCIFORMA	Composition			X
PCT_SCRAPE	FFG		X	
PCT_CLING	Habit		X	X
RICH_CLING	Habit		X	
PCT_EPT_RICH	Richness/Diversity		X	
PIELOU	Richness/Diversity	X		
RICH_EPHEM_EPEORUS	Richness/Diversity			X
PCT_INTOL_0_4	Tolerance			X

Table 7. Macroinvertebrate sample numbers in the Coast and Inland environmental condition categories (and percent of the Region's total sample number).

Region	Reference	Minimally Degraded	Mixed	Moderately Degraded	Degraded	Total Count
COAST	63 (1.2%)	92 (1.8%)	3,719 (71.5%)	798 (15.3%)	533 (10.2%)	5,205
INLAND	1,803 (11.2%)	512 (3.2%)	8,429 (52.3%)	4,575 (28.4%)	790 (4.9%)	16,109

C. **Two Region Indices**

Reference and Degraded samples are unevenly distributed across the Inland and Coast region. The Inland region has ten bioregions and over half of its Reference samples are in the LNP, NCA, and SRV bioregions. The Coast has two bioregions and about two-thirds of Reference samples are in SEP. The uneven distributions of Reference sampling events give the few well sampled areas more influence in shaping each region's index. This inherent bias again was minimized by randomly selecting 50 Reference and 50 Degraded sampling events from each bioregion in a region before implementing the 50 index development iterations to select metrics and scoring thresholds for the region.

i. Order-Level Indices

Both order-level regional indices consisted of three composition metrics, one richness/diversity metric, and one tolerance metric (Table 8). Overall, the mean metric BDE was 64.8% and ranged from 55.6% to 78.0%. The Coast BSP was 53.0; the Inland BSP, 54.0 (Table K-2). The Inland index CE was 77.6%; the Coast index, 69.5%. Figure 7 shows declining trends in index scores as degradation increases.

Table 8. Metrics included in the Coast and Inland order-level indices.

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
COAST	PCT_ARTHROPODA	Composition	Increase	92.26	93.30	52.38
COAST	PCT_DIPTERA	Composition	Increase	37.63	50.37	65.08
COAST	PCT_HEXAPODA	Composition	Decrease	78.04	75.63	55.56
COAST	PIELOU	Richness/Diversity	Decrease	0.70	0.65	68.25
COAST	PCT_DOM5	Tolerance	Increase	94.30	97.82	73.02
INLAND	PCT_PTERYGOTA	Composition	Decrease	99.09	98.12	58.01
INLAND	PCT_EPHEMEROPTERA	Composition	Decrease	31.67	6.55	77.98
INLAND	PCT_EPT	Composition	Decrease	67.71	46.07	73.93
INLAND	PIELOU	Richness/Diversity	Decrease	0.76	0.71	61.90
INLAND	PCT_DOM4	Tolerance	Increase	94.75	96.69	61.79
					Mean	64.79

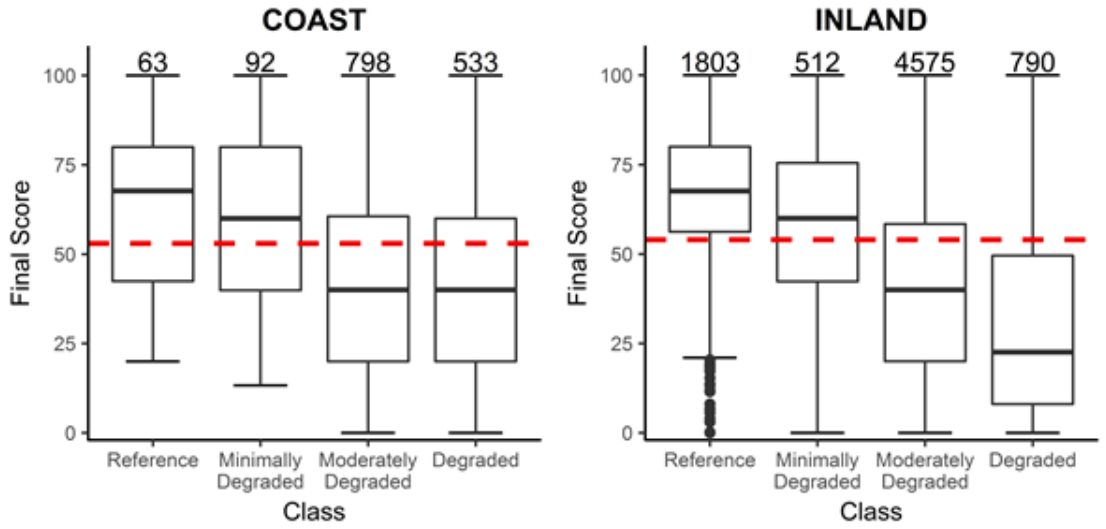


Figure 7. Distributions of index scores for the Coast and Inland order-level indices. The whisker lengths are designated by the interquartile range multiplied by 1.5. The horizontal red line represents the Best Separation Point (BSP) between the Reference and Degraded distributions. BSP was used as a threshold for evaluating Classification Efficiency (CE).

ii. Family-Level Indices

Both regional indices included at least one metric from the five metric classes (Table 9). Ten metrics were included in the Coast index and nine metrics were included in the Inland index. The mean metric BDE for both regions was 73.0%. The Inland metric BDEs ranged from 59.5% - 84.0%, and the Coast metric BDEs ranged from 63.5% - 81.0%. The Inland index CE was 82.4%; the Coast, 72.3% (Table K-2). The Inland and the Coast BSPs were 53.0, which did not differ greatly from the expected value of 50.0. Figure 8 shows declining trends in index scores as degradation increases.

Table 9. Metrics included in the Coast and Inland family-level indices and the associated scoring thresholds.

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
COAST	GOLD	Composition	Decrease	52.59	32.15	66.67
COAST	MARGALEFS	Richness/Diversity	Decrease	2.54	1.84	73.02
COAST	PCT_EPT_RICH	Richness/Diversity	Decrease	33.16	14.81	71.43
COAST	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	2.00	0.00	80.95
COAST	PCT_COLLECT	FFG	Increase	80.57	91.70	66.67
COAST	PCT_PREDATOR	FFG	Decrease	5.19	3.65	71.43
COAST	RICH_FILTER	FFG	Decrease	2.00	0.51	80.95
COAST	RICH_CLIMB	Habit	Decrease	2.00	0.30	63.49
COAST	ASPT_MOD	Tolerance	Decrease	4.29	2.72	74.60
COAST	HBI	Tolerance	Increase	5.28	6.13	71.43
INLAND	PCT_EPHEMEROPTERA_NO_BAETID	Composition	Decrease	24.23	1.76	78.87
INLAND	PCT_EPT_RICH	Richness/Diversity	Decrease	62.68	49.12	68.50
INLAND	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	4.00	2.08	59.46
INLAND	RICH_PLECOPTERA	Richness/Diversity	Decrease	3.00	0.93	72.38
INLAND	PCT_SCRAPE	FFG	Decrease	10.37	0.80	77.76
INLAND	PCT_BURROW	Habit	Increase	17.47	33.61	75.10
INLAND	RICH_CLING	Habit	Decrease	8.41	5.36	84.03
INLAND	ASPT_MOD	Tolerance	Decrease	6.57	4.68	71.21
INLAND	PCT_INTOL_0_4	Tolerance	Decrease	61.03	34.10	79.03
					Mean	73.00

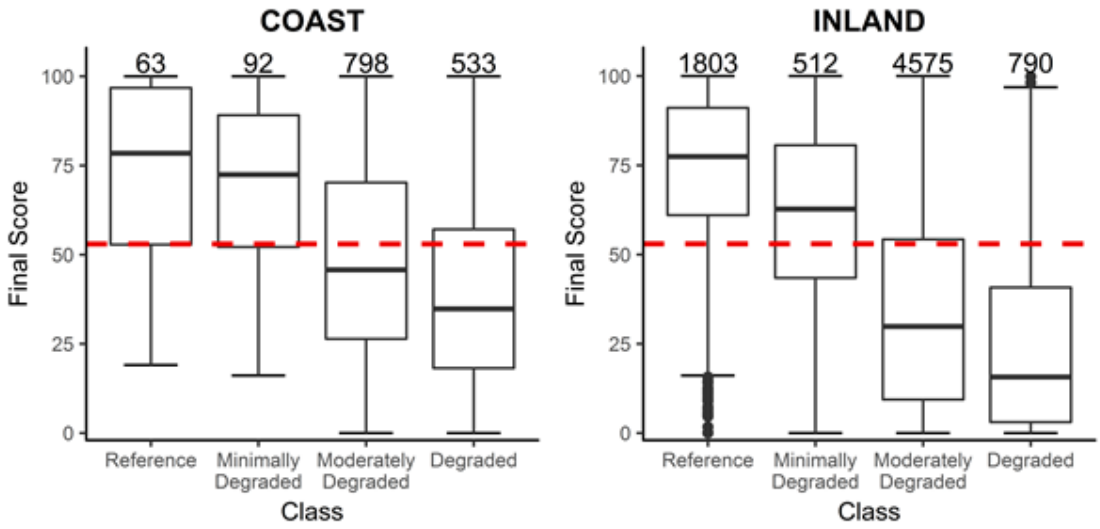


Figure 8. Distributions of index scores for the Coast and Inland family-level indices. The whisker lengths are designated by the interquartile range multiplied by 1.5. The horizontal red line represents the Best Separation Point (BSP) between the Reference and Degraded distributions. BSP was used as a threshold for evaluating Classification Efficiency (CE).

iii. Genus-Level Indices

The Reference and Degraded sample sizes remained about the same in the Coast region after removing samples that did not meet the minimum taxonomic resolution requirements for a genus-level index. They were somewhat lower in the Inland region but still sufficient for index development (Table 10). For both indices, at least one metric was selected to represent each of the five metric classes (

Table 11). The metric BDEs in the Coast region ranged from 51.6% to 79.0% while BDEs in the Inland region ranged from 63.9% to 79.7%. Overall, the mean BDE for both regions was 67.3%. The Inland index CE was 82.6%; the Coast index, 74.6% (Table K-2). The BSP values for both indices did not differ much from the expected BSP value of 50.0 (Coast BSP = 55.0, Inland BSP = 50.0). Figure 9 depicts a declining trend in the index scores with increasing degradation.

Table 10. Sample numbers in the Coast and Inland region site classes that met the genus-level index requirements.

Region	Reference	Minimally Degraded	Mixed	Moderately Degraded	Degraded	Total Count
COAST	62 (1.3%)	89 (1.9%)	3301 (69.1%)	791 (16.6%)	533 (11.2%)	4,776
INLAND	1,525 (11%)	449 (3.2%)	6,922 (50.1%)	4,236 (30.6%)	695 (5%)	13,827

Table 11. Metrics included in the Coast and Inland genus-level indices.

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
COAST	PCT_CAECIDOTEA	Composition	Increase	0.00	0.85	79.03
COAST	PCT_EPT_NO_HYDRO	Composition	Increase	83.37	93.55	58.06
COAST	GOLD	Composition	Decrease	51.52	33.37	66.13
COAST	PCT_DIPTERA	Composition	Increase	38.63	50.97	66.13
COAST	HURLBERTS_PIE	Richness/Diversity	Increase	0.66	0.81	66.13
COAST	PCT_COLLECT	FFG	Decrease	75.57	70.12	51.61
COAST	PCT_GATHER	FFG	Increase	49.12	56.83	59.68
COAST	PCT_PREDATOR	FFG	Increase	6.62	8.57	54.84
COAST	PCT_BURROW	Habit	Decrease	30.96	17.92	51.61
COAST	RICH_BURROW	Habit	Increase	2.00	3.55	70.97
COAST	HBI	Tolerance	Increase	5.33	6.08	69.35
COAST	PCT_DOM1	Tolerance	Decrease	52.98	34.61	67.74
COAST	PCT_TOLERANT_5_10	Tolerance	Increase	84.02	92.84	66.13
COAST	PCT_TOLERANT_7_10	Tolerance	Increase	14.32	21.15	64.52
COAST	RICH_TOL	Tolerance	Increase	2.93	4.66	77.42
INLAND	PCT_COTE	Composition	Decrease	60.28	34.42	76.13
INLAND	PCT_EPHEMEROPTERA	Composition	Decrease	31.47	3.02	79.67
INLAND	PCT_EPT_RICH_NO_TOL	Richness/Diversity	Decrease	50.00	24.78	64.72
INLAND	PIELOU	Richness/Diversity	Decrease	0.82	0.75	63.93
INLAND	PCT_SCRAPE	FFG	Decrease	16.83	6.64	70.75
INLAND	PCT_CLING	Habit	Decrease	60.07	36.10	74.89
INLAND	PCT_SWIM	Habit	Decrease	8.55	0.83	72.39
INLAND	RICH_CLING	Habit	Decrease	11.08	6.26	64.33
INLAND	PCT_INTOL_0_3	Tolerance	Decrease	53.15	11.18	78.36
INLAND	PCT_MOD_TOL_4_6	Tolerance	Increase	43.71	63.17	68.39
					Mean	67.32

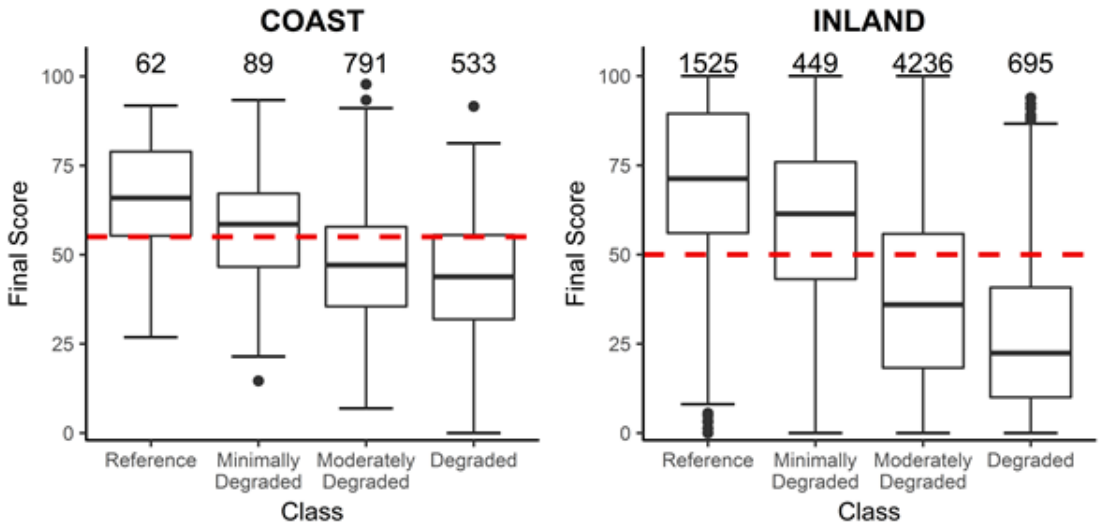


Figure 9. Distributions of index scores for the Coast and Inland genus-level indices. The whisker lengths are designated by the interquartile range multiplied by 1.5. The horizontal red line represents the Best Separation Point (BSP) between the Reference and Degraded distributions. BSP was used as a threshold for evaluating Classification Efficiency (CE).

D. **Bioregion Indices**

The CA, MAC, and SEP bioregions had fewer than 50 Reference sampling events; the BLUE, CA, and PIED bioregions had fewer than 50 Degraded sampling events (Table 12). Fewer than approximately 50 samples from either Reference or Degraded conditions begins to increase variability, and thus uncertainty, in an index's ability to correctly classify a sample (Buchanan et al. 2011). The CA bioregion was the poorest represented bioregion (order and family n = 367; genus n = 297), LNP was the best represented bioregion at the order- and family-levels (n = 3,625), and SEP was the best represented bioregion at the genus-level (n = 3,261).

Table 12. Order- and family-level macroinvertebrate sample numbers in each bioregion's environmental condition class (and percent of the bioregion's total sample number).

Bioregion	Reference	Minimally Degraded	Mixed	Moderately Degraded	Degraded	Total Count
BLUE	133 (29.9%)	27 (6.1%)	235 (52.8%)	43 (9.7%)	7 (1.6%)	445
CA	35 (9.5%)	15 (4.1%)	176 (48%)	98 (26.7%)	43 (11.7%)	367
LNP	347 (9.6%)	77 (2.1%)	1,698 (46.8%)	1341 (37%)	162 (4.5%)	3,625
MAC	17 (1.1%)	30 (1.9%)	1113 (69.7%)	180 (11.3%)	257 (16.1%)	1,597
NAPU	89 (6.7%)	53 (4%)	735 (55.6%)	374 (28.3%)	70 (5.3%)	1,321

Bioregion	Reference	Minimally Degraded	Mixed	Moderately Degraded	Degraded	Total Count
NCA	318 (36.3%)	28 (3.2%)	383 (43.7%)	90 (10.3%)	58 (6.6%)	877
NRV	149 (13.6%)	31 (2.8%)	611 (55.6%)	242 (22%)	65 (5.9%)	1,098
PIED	133 (10.7%)	57 (4.6%)	665 (53.5%)	359 (28.9%)	29 (2.3%)	1,243
SEP	46 (1.3%)	62 (1.7%)	2,606 (72.2%)	618 (17.1%)	276 (7.6%)	3,608
SGV	127 (7.9%)	35 (2.2%)	497 (30.8%)	835 (51.7%)	122 (7.5%)	1,616
SRV	396 (18.6%)	128 (6%)	1,111 (52.2%)	426 (20%)	68 (3.2%)	2,129
UNP	76 (2.2%)	61 (1.8%)	2,318 (68.4%)	767 (22.6%)	166 (4.9%)	3,388

i. Order-Level Indices

An attempt was made to include at least one richness/diversity metric, one tolerance metric, and one composition metric in each order-level index. On average, six metrics were selected for each order-level index (Table 13). Redundancy, range, variability, and sensitivity assessments limited the representation of these metric classes in six of the twelve bioregions (i.e., BLUE, LNP, MAC, PIED, SEP, and UNP). The greatest number of metrics was selected for the SRV bioregion (n = 8). The performance of the indices varied greatly. The average index CE was 74.2% but CE values ranged from 57.8%, or little better than a coin toss, in NAPU to 87.6% in LNP (Table 14). The mean index BSP ($\bar{x} = 54.3$) did not vary drastically from the expected value of 50.0 (Table 14). Index scores, in general, depicted a weak descending trend from Reference to Degraded conditions (Figure 10 and Figure 11).

Table 13. Metrics included in the bioregion-specific, order-level indices.

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
BLUE	GOLD	Composition	Decrease	83.44	63.12	82.71
BLUE	PCT_COTE	Composition	Decrease	63.63	42.94	74.44
BLUE	PCT_EPHEMEROPTERA	Composition	Decrease	39.77	4.38	74.44
BLUE	PCT_DOM2	Tolerance	Decrease	73.94	69.60	62.41
CA	PCT_COTE	Composition	Decrease	48.12	46.24	54.29
CA	PCT_DIPTERA	Composition	Decrease	20.06	20.89	51.43
CA	PCT_EPHEMEROPTERA	Composition	Decrease	19.42	4.00	68.57
CA	PCT_PLECOPTERA	Composition	Decrease	26.55	0.00	71.43
CA	RICH	Richness/Diversity	Decrease	6.00	4.42	65.71
CA	PCT_DOM3	Tolerance	Increase	84.33	94.10	71.43

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
LNP	PCT_ARTHROPODA	Composition	Decrease	99.17	95.63	75.50
LNP	PCT_COTE	Composition	Decrease	60.65	9.35	89.91
LNP	PCT_EPHEMEROPTERA	Composition	Decrease	28.68	0.00	90.20
LNP	PCT_EPT	Composition	Decrease	66.92	0.00	92.51
LNP	PCT_DOM1	Tolerance	Increase	48.22	55.53	66.57
LNP	PCT_DOM5	Tolerance	Increase	97.78	98.18	62.54
MAC	PCT_PTERYGOTA	Composition	Increase	68.94	72.86	52.94
MAC	GOLD	Composition	Decrease	58.54	34.80	64.71
MAC	PCT_ARTHROPODA	Composition	Increase	73.97	95.76	70.59
MAC	PCT_ODONATA	Composition	Decrease	1.82	0.42	76.47
MAC	MARGALEFS	Richness/Diversity	Decrease	1.58	1.19	64.71
MAC	PIELOU	Richness/Diversity	Decrease	0.71	0.62	64.71
NAPU	PCT_COTE	Composition	Decrease	62.93	58.29	55.06
NAPU	PCT_EPHEMEROPTERA	Composition	Decrease	31.39	18.88	57.30
NAPU	PCT_EPT	Composition	Decrease	62.23	54.27	56.18
NAPU	PIELOU	Richness/Diversity	Increase	0.77	0.79	53.93
NAPU	PCT_DOM3	Tolerance	Decrease	88.71	86.92	55.06
NAPU	PCT_DOM4	Tolerance	Increase	96.11	96.63	51.69
NCA	PCT_COTE	Composition	Decrease	64.00	47.02	69.18
NCA	PCT_EPHEMEROPTERA	Composition	Decrease	40.72	9.95	77.04
NCA	PCT_EPT	Composition	Decrease	75.02	74.39	50.63
NCA	PCT_PLECOPTERA	Composition	Increase	15.14	18.03	60.06
NCA	RICH	Richness/Diversity	Decrease	6.00	5.66	53.46
NCA	PCT_DOM1	Tolerance	Increase	45.91	54.68	62.26
NCA	PCT_DOM4	Tolerance	Increase	95.79	99.02	71.70
NRV	PCT_PTERYGOTA	Composition	Decrease	98.91	98.08	61.07
NRV	PCT_DIPTERA	Composition	Increase	18.23	29.78	59.06
NRV	PCT_EPHEMEROPTERA	Composition	Decrease	33.53	4.36	69.80
NRV	PCT_POTEC	Composition	Decrease	75.16	57.38	66.44
NRV	RICH	Richness/Diversity	Decrease	6.00	5.26	72.48
NRV	PCT_DOM1	Tolerance	Increase	46.00	50.08	56.38
NRV	PCT_DOM4	Tolerance	Increase	93.20	98.12	65.77
PIED	PCT_ARTHROPODA	Composition	Decrease	98.07	95.26	62.41
PIED	PCT_EPHEMEROPTERA	Composition	Decrease	33.31	16.05	75.94
PIED	PCT_EPT	Composition	Decrease	65.88	36.65	75.19
PIED	RICH	Richness/Diversity	Increase	7.50	7.89	45.11
SEP	GOLD	Composition	Decrease	50.56	30.40	69.57
SEP	PCT_ARTHROPODA	Composition	Decrease	95.27	91.43	54.35
SEP	PCT_COTE	Composition	Decrease	26.19	2.11	80.43
SEP	PCT_HEXAPODA	Composition	Decrease	83.99	78.40	56.52
SEP	PIELOU	Richness/Diversity	Decrease	0.70	0.64	73.91

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
SEP	RICH	Richness/Diversity	Decrease	8.01	6.38	71.74
SGV	PCT_PTERYGOTA	Composition	Decrease	96.87	88.95	68.50
SGV	GOLD	Composition	Decrease	80.19	65.96	61.42
SGV	PCT_EPHEMEROPTERA	Composition	Decrease	30.33	8.91	76.38
SGV	PCT_EPT	Composition	Decrease	63.28	31.77	82.68
SGV	PIELOU	Richness/Diversity	Decrease	0.75	0.72	64.57
SGV	PCT_DOM3	Tolerance	Increase	85.46	90.04	66.93
SGV	PCT_DOM5	Tolerance	Increase	96.97	98.26	56.69
SRV	PCT_ARTHROPODA	Composition	Decrease	99.42	99.05	45.71
SRV	PCT_COTE	Composition	Decrease	57.58	44.28	68.43
SRV	PCT_DIPTERA	Composition	Increase	17.49	30.31	70.20
SRV	PCT_EPHEMEROPTERA	Composition	Decrease	30.48	13.52	69.44
SRV	PCT_EPT	Composition	Decrease	69.27	56.82	60.35
SRV	PIELOU	Richness/Diversity	Decrease	0.78	0.68	69.70
SRV	PCT_DOM1	Tolerance	Increase	42.82	51.54	63.89
SRV	PCT_DOM4	Tolerance	Increase	95.40	96.35	60.35
UNP	PCT_COTE	Composition	Decrease	57.46	17.52	77.63
UNP	PCT_HEXAPODA	Composition	Decrease	99.50	96.69	81.58
UNP	PCT_DOM1	Tolerance	Increase	50.00	62.01	60.53
UNP	PCT_DOM4	Tolerance	Increase	95.62	96.49	56.58
					Mean	66.19

Table 14. The Best Separation Point (BSP) for bioregion index scores The BSP is used as the threshold for calculating Classification Efficiency (CE).

	Order-Level Index		**Family-Level Index**		**Genus-Level Index**	
Bioregion	Index BSP	Index CE	Index BSP	Index CE	Index BSP	Index CE
BLUE	47.0	85.3	60.0	85.7	50.0	87.0
CA	55.0	76.9	54.0	82.1	53.0	84.4
LNP	63.0	87.6	61.0	89.6	57.0	87.9
MAC	59.0	74.8	57.0	76.8	53.0	82.6
NAPU	50.0	57.8	48.0	70.4	55.0	72.7
NCA	55.0	68.8	47.0	82.4	49.0	88.6
NRV	54.0	73.1	50.0	77.8	46.0	89.4
PIED	49.0	69.8	60.0	83.9	50.0	84.4
SEP	56.0	73.2	55.0	75.4	58.0	77.1
SGV	54.0	74.7	61.0	85.1	54.0	80.7
SRV	56.0	70.5	53.0	79.7	52.0	75.5
UNP	53.0	77.7	57.0	90.0	59.0	89.3
Mean	54.3	74.2	55.3	81.6	53.0	83.3

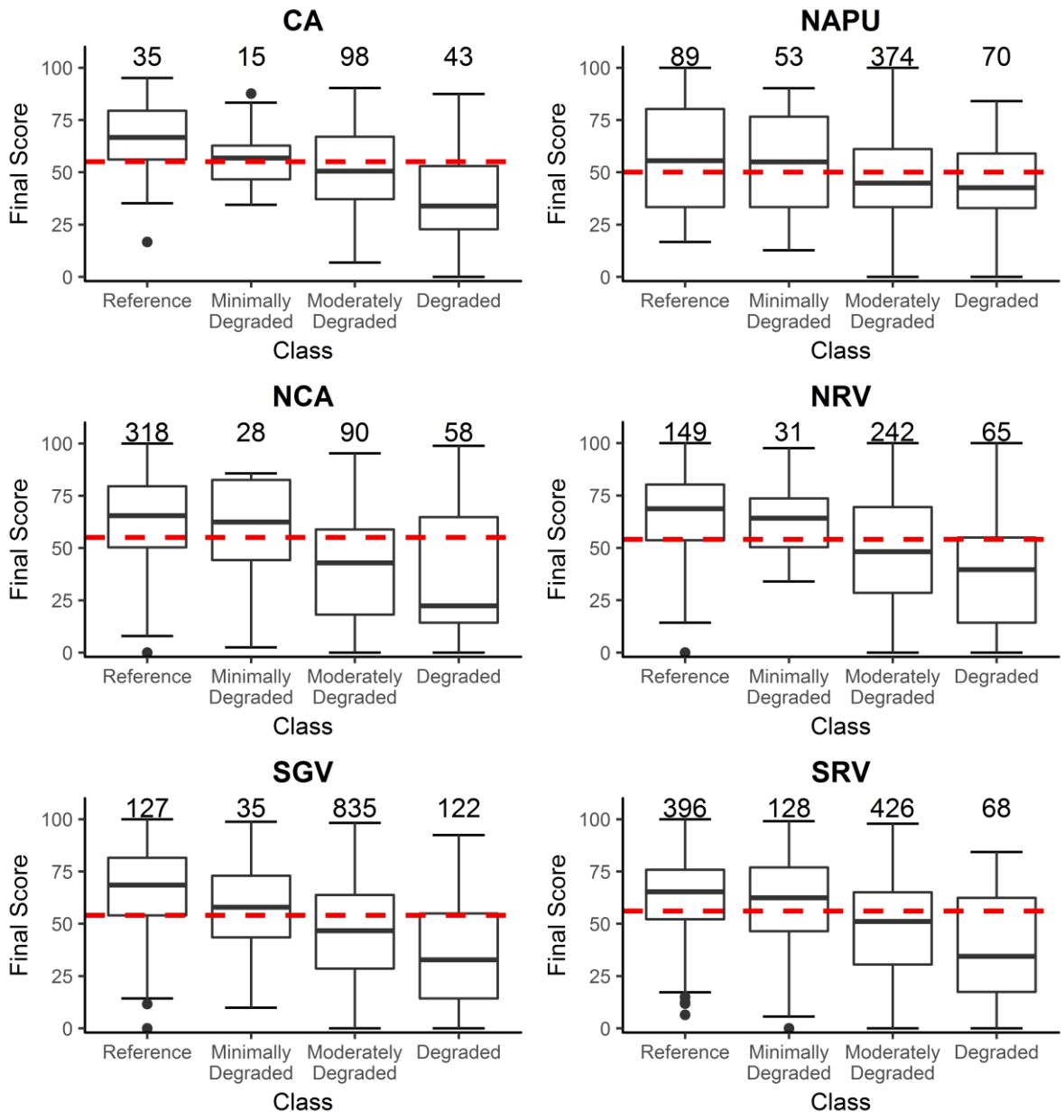


Figure 10. Distributions of index scores for order-level indices in six bioregions: Central Appalachians (CA), Northern Appalachian Plateau and Uplands (NAPU), Northern Central Appalachians (NCA), Northern Ridge and Valley (NRV), Southern Great Valley (SGV), and Southern Ridge and Valley (SRV). The whisker lengths are designated by the Interquartile Range multiplied by 1.5. The horizontal red line represents the Best Separation Point (BSP) between the Reference and Degraded distributions. BSP was used as a threshold for evaluating Classification Efficiency (CE).

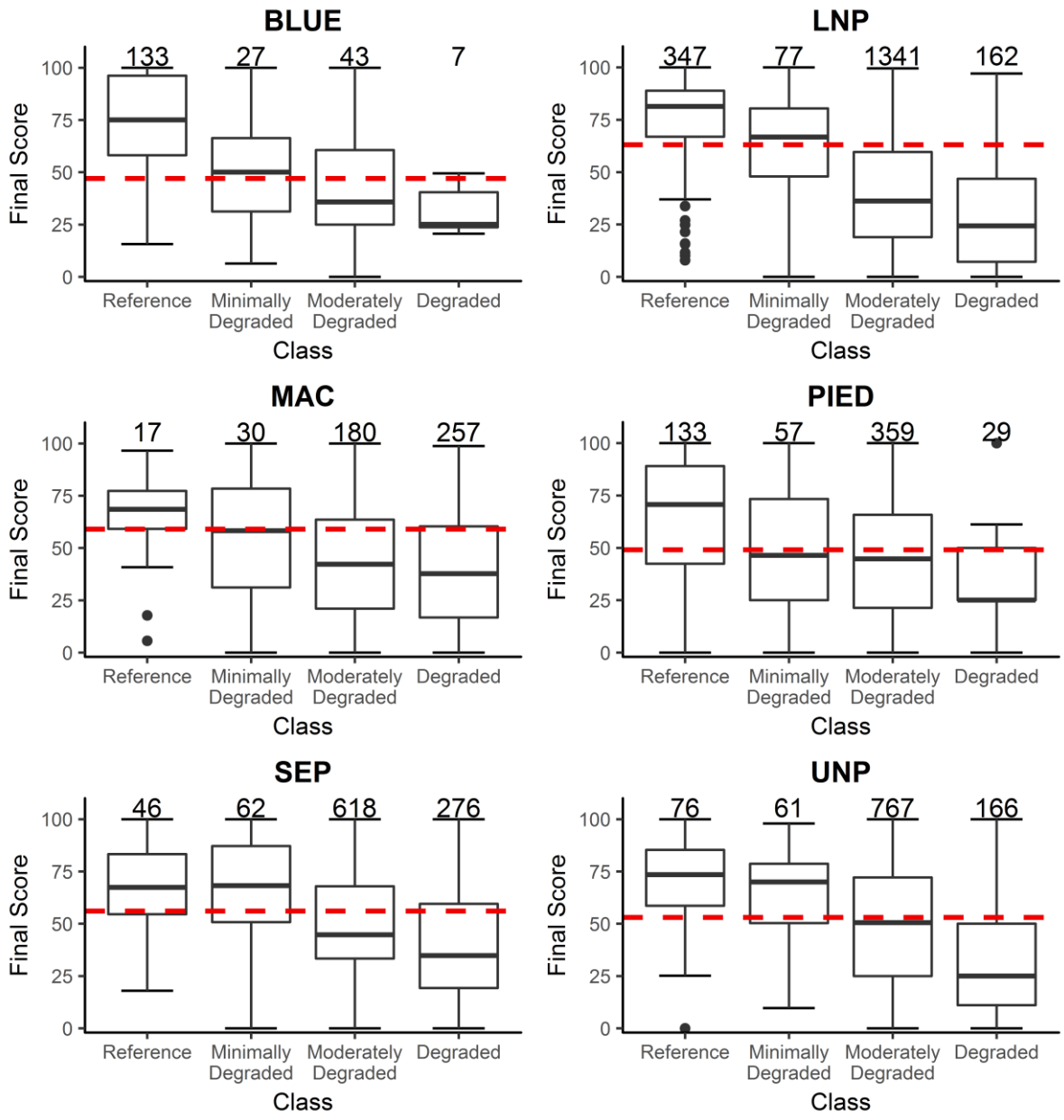


Figure 11. Distributions of index scores for order-level indices in six bioregions: Blue Ridge Mountains (BLUE), Lower Northern Piedmont (LNP), Middle-Atlantic Coast (MAC), Piedmont (PIED), Southeastern Plains (SEP), and Upper Northern Piedmont (UNP). The whisker lengths are designated by the Interquartile Range multiplied by 1.5. The horizontal red line represents the Best Separation Point (BSP) between the Reference and Degraded distributions. BSP was used as a threshold for evaluating Classification Efficiency (CE).

ii. Family-Level Indices

The average number of metrics selected for each bioregion was twelve; the greatest number of metrics (n = 21) was selected in the SGV bioregion (Table 15). Although we required at least four of the five metric classes to be represented in each index, redundancy, range, variability, and sensitivity tests prevented the selection of four metric classes in the PIED bioregion. The PIED index was composed of three richness/diversity metrics, two habit metrics, and three tolerance metrics. The mean BDE of all the metrics was 72.43%. CEs of the indices averaged 81.6%, and ranged from 70.4% in the NAPU to 90.0% in the UNP (Table 14). The BSP's averaged 55.3, and ranged from 47 (NCA) to 61 (LNP and SGV). In general, a descending gradient as degradation increases was observed for each bioregion index (Figure 12 and Figure 13).

Table 15. Metrics included in the family-level bioregion indices.

Bio-region	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
BLUE	PCT_SYSTELLOGNATHA	Composition	Decrease	6.06	0.00	86.47
BLUE	PCT_EPT_RICH_NO_TOL	Richness/Diversity	Decrease	52.78	34.74	82.71
BLUE	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	4.00	0.33	85.71
BLUE	PCT_SCRAPE	FFG	Decrease	15.23	0.00	90.23
BLUE	PCT_BURROW	Habit	Increase	12.65	40.71	87.97
CA	PCT_EUHOLOGNATHA	Composition	Decrease	17.98	0.23	71.43
CA	PCT_HEPTAGENIIDAE	Composition	Decrease	6.69	0.00	77.14
CA	PCT_NON_HYDROP_ TRICHOPTERA	Composition	Decrease	40.40	25.02	71.43
CA	PCT_PISCIFORMA	Composition	Decrease	12.33	5.49	74.29
CA	PCT_SYSTELLOGNATHA	Composition	Decrease	5.44	0.08	71.43
CA	PCT_HYDRO_EPT	Composition	Increase	12.50	21.73	68.57
CA	PCT_EPT_RICH	Richness/Diversity	Decrease	66.82	56.82	65.71
CA	RICH_PLECOPTERA	Richness/Diversity	Decrease	4.00	0.47	74.29
CA	RICH_COLLECT	FFG	Decrease	7.00	5.13	62.86
CA	RICH_FILTER	FFG	Decrease	2.91	1.38	51.43
CA	RICH_PREDATOR	FFG	Decrease	3.00	1.44	80.00
CA	RICH_BURROW	Habit	Decrease	3.00	1.40	82.86
CA	PCT_MOD_TOL_4_6	Tolerance	Increase	43.04	57.26	71.43
LNP	PCT_PTERYGOTA	Composition	Decrease	98.36	92.09	74.64
LNP	PCT_EPT_RICH	Richness/Diversity	Decrease	52.32	38.07	70.89
LNP	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	3.00	1.70	81.56
LNP	PCT_COLLECT	FFG	Increase	72.82	92.49	88.47
LNP	PCT_PREDATOR	FFG	Decrease	7.78	1.43	87.90
LNP	PCT_CLING	Habit	Decrease	62.77	4.54	91.35
LNP	RICH_CLING	Habit	Decrease	7.00	3.69	88.18
LNP	ASPT_MOD	Tolerance	Decrease	6.62	2.70	93.66
LNP	PCT_MOD_TOL_4_6	Tolerance	Increase	46.54	77.78	88.47
MAC	GOLD	Composition	Decrease	58.96	33.91	64.71

Bio-region	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
MAC	PCT_ARTHROPODA	Composition	Increase	74.01	95.38	70.59
MAC	PCT_CHIRONOMIDAE	Composition	Increase	23.01	42.68	64.71
MAC	PCT_ODONATA	Composition	Decrease	1.82	0.43	76.47
MAC	PCT_COLLECT	FFG	Increase	73.25	96.19	70.59
MAC	PCT_PREDATOR	FFG	Decrease	5.16	4.01	76.47
MAC	PCT_SCRAPE	FFG	Decrease	12.07	0.00	76.47
MAC	RICH_CLIMB	Habit	Decrease	2.45	0.00	64.71
MAC	ASPT_MOD	Tolerance	Decrease	3.27	2.90	70.59
MAC	PCT_MOD_TOL_4_6	Tolerance	Increase	52.49	75.48	64.71
MAC	PCT_URBAN_INTOL	Tolerance	Increase	65.28	71.78	76.47
MAC	RICH_TOL	Tolerance	Decrease	4.00	2.16	52.94
NAPU	PCT_EPT_NO_HYDRO	Composition	Decrease	87.96	70.51	67.42
NAPU	PCT_NON_HYDROP_ TRICHOPTERA	Composition	Decrease	45.08	20.60	68.54
NAPU	PCT_EPHEMEROPTERA	Composition	Decrease	31.47	19.62	57.30
NAPU	PCT_EPT	Composition	Decrease	62.26	54.08	56.18
NAPU	MARGALEFS	Richness/Diversity	Decrease	2.83	2.74	53.93
NAPU	PCT_EPT_RICH_NO_TOL	Richness/Diversity	Decrease	45.32	32.85	68.54
NAPU	PIELOU	Richness/Diversity	Increase	0.79	0.80	51.69
NAPU	RICH_PLECOPTERA	Richness/Diversity	Decrease	2.13	0.70	67.42
NAPU	PCT_COLLECT	FFG	Increase	71.84	82.80	59.55
NAPU	PCT_FILTER	FFG	Increase	15.51	22.62	61.80
NAPU	PCT_PREDATOR	FFG	Increase	5.36	6.24	56.18
NAPU	PCT_SCRAPE	FFG	Decrease	9.36	6.89	58.43
NAPU	RICH_GATHER	FFG	Increase	5.00	6.01	53.93
NAPU	RICH_SHRED	FFG	Decrease	2.00	0.84	61.80
NAPU	PCT_CLING	Habit	Decrease	59.19	50.98	55.06
NAPU	RICH_BURROW	Habit	Decrease	2.95	1.93	53.93
NAPU	RICH_CLING	Habit	Decrease	8.00	6.77	64.04
NAPU	RICH_SPRAWL	Habit	Increase	1.00	2.30	56.18
NAPU	PCT_INTOL_0_3	Tolerance	Decrease	44.70	24.22	61.80
NAPU	RICH_MODTOL	Tolerance	Increase	6.00	7.03	55.06
NCA	PCT_HEPTAGENIIDAE	Composition	Decrease	13.72	0.00	76.73
NCA	PCT_COTE	Composition	Decrease	63.95	47.71	68.55
NCA	PCT_EPHEMEROPTERA	Composition	Decrease	40.73	9.02	77.36
NCA	PCT_LIMESTONE	Composition	Decrease	8.93	0.00	79.25
NCA	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	4.00	3.83	70.13
NCA	RICH_EPT	Richness/Diversity	Decrease	12.00	8.02	68.55
NCA	RICH_PLECOPTERA	Richness/Diversity	Decrease	4.04	3.45	68.24
NCA	RICH_CLING	Habit	Decrease	10.00	7.75	78.62
NCA	ASPT_MOD	Tolerance	Decrease	7.54	6.67	65.41

Bio-region	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
NRV	PCT_EPHEMEROPTERA_NO_BAETID	Composition	Decrease	23.69	6.01	68.46
NRV	PCT_EPT_RICH	Richness/Diversity	Decrease	64.70	47.53	73.15
NRV	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	4.00	1.80	85.23
NRV	RICH_TRICHOPTERA	Richness/Diversity	Decrease	3.00	2.53	65.10
NRV	RICH_COLLECT	FFG	Decrease	8.00	7.14	62.42
NRV	RICH_PREDATOR	FFG	Decrease	3.00	0.32	79.19
NRV	RICH_SHRED	FFG	Decrease	2.00	1.27	69.80
NRV	ASPT_MOD	Tolerance	Decrease	6.27	4.59	72.48
PIED	PCT_EPT_RICH	Richness/Diversity	Decrease	53.58	38.51	68.42
PIED	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	3.60	0.86	84.21
PIED	RICH_TRICHOPTERA	Richness/Diversity	Decrease	2.03	0.45	78.95
PIED	PCT_CLING	Habit	Decrease	62.62	32.33	81.95
PIED	RICH_CLING	Habit	Decrease	8.00	2.96	91.73
PIED	PCT_INTOL_0_4	Tolerance	Decrease	58.15	27.97	81.20
PIED	PCT_TOLERANT_7_10	Tolerance	Increase	1.60	4.84	75.94
PIED	PCT_URBAN_INTOL	Tolerance	Decrease	95.41	86.93	66.17
SEP	GOLD	Composition	Decrease	50.30	32.16	69.57
SEP	PCT_COTE	Composition	Decrease	26.37	1.62	80.43
SEP	PCT_EPT_RICH_NO_TOL	Richness/Diversity	Decrease	24.73	16.49	63.04
SEP	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	2.00	0.00	86.96
SEP	RICH_TRICHOPTERA	Richness/Diversity	Decrease	2.00	0.41	63.04
SEP	RICH_COLLECT	FFG	Decrease	8.00	4.62	78.26
SEP	RICH_FILTER	FFG	Decrease	2.31	0.04	84.78
SEP	PCT_CLING	Habit	Decrease	43.05	0.00	78.26
SEP	RICH_CLING	Habit	Decrease	4.93	1.03	82.61
SEP	ASPT_MOD	Tolerance	Decrease	4.65	3.18	63.04
SEP	HBI	Tolerance	Increase	5.12	5.86	76.09
SGV	PCT_EPT_NO_HYDRO	Composition	Decrease	83.21	73.53	65.35
SGV	PCT_HEPTAGENIIDAE	Composition	Decrease	7.89	0.00	81.89
SGV	PCT_NON_HYDROP_TRICHOPTERA	Composition	Decrease	30.54	10.57	66.93
SGV	PCT_PISCIFORMA	Composition	Decrease	17.07	0.10	79.53
SGV	PCT_PTERYGOTA	Composition	Decrease	96.88	90.20	64.57
SGV	GOLD	Composition	Decrease	80.89	64.95	61.42
SGV	PCT_EPHEMEROPTERA_NO_BAETID	Composition	Decrease	23.94	5.23	75.59
SGV	PCT_EPT_RICH_NO_TOL	Richness/Diversity	Decrease	32.97	19.39	91.34
SGV	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	3.94	0.61	84.25
SGV	RICH_TRICHOPTERA	Richness/Diversity	Decrease	2.54	0.08	89.76
SGV	PCT_COLLECT	FFG	Increase	71.17	94.57	81.10
SGV	PCT_FILTER	FFG	Decrease	22.79	8.60	62.20

Bio-region	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
SGV	PCT_PREDATOR	FFG	Decrease	5.28	0.02	74.02
SGV	PCT_SCRAPE	FFG	Decrease	15.43	0.32	77.17
SGV	RICH_COLLECT	FFG	Decrease	8.00	6.32	66.93
SGV	RICH_FILTER	FFG	Decrease	3.00	1.36	66.14
SGV	PCT_CLING	Habit	Decrease	62.23	39.70	71.65
SGV	PCT_SPRAWL	Habit	Increase	1.79	4.54	59.84
SGV	ASPT_MOD	Tolerance	Decrease	6.44	3.94	81.89
SGV	PCT_MOD_TOL_4_6	Tolerance	Increase	52.49	76.71	71.65
SGV	PCT_TOLERANT_7_10	Tolerance	Increase	1.73	4.30	75.59
SRV	PCT_COTE	Composition	Decrease	57.55	43.78	68.69
SRV	PCT_EPHEMEROPTERA_NO_BAETID	Composition	Decrease	21.00	2.80	75.25
SRV	PCT_EPT_RICH_NO_TOL	Richness/Diversity	Decrease	47.16	34.94	66.16
SRV	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	4.00	3.66	66.16
SRV	RICH_PLECOPTERA	Richness/Diversity	Decrease	3.00	0.65	73.23
SRV	RICH_TRICHOPTERA	Richness/Diversity	Decrease	3.00	2.79	61.62
SRV	PCT_PREDATOR	FFG	Decrease	7.12	2.24	67.17
SRV	PCT_SCRAPE	FFG	Decrease	10.88	0.47	78.03
SRV	RICH_COLLECT	FFG	Decrease	7.97	6.02	79.29
SRV	RICH_PREDATOR	FFG	Decrease	3.00	2.72	60.10
SRV	PCT_BURROW	Habit	Increase	17.13	31.28	70.96
SRV	PCT_CLING	Habit	Decrease	58.45	40.16	71.46
SRV	RICH_CLING	Habit	Decrease	8.43	6.16	74.49
SRV	ASPT_MOD	Tolerance	Decrease	6.44	4.95	67.42
SRV	RICH_INTOL	Tolerance	Decrease	8.00	4.15	73.23
UNP	PCT_PISCIFORMA	Composition	Decrease	9.43	0.00	80.26
UNP	PCT_RETREAT_CADDISFLY	Composition	Decrease	10.97	2.89	72.37
UNP	PCT_HEXAPODA	Composition	Decrease	99.50	96.59	81.58
UNP	PCT_EPT_RICH_NO_TOL	Richness/Diversity	Decrease	37.09	22.10	93.42
UNP	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	3.00	1.44	73.68
UNP	RICH_PLECOPTERA	Richness/Diversity	Decrease	2.23	0.00	89.47
UNP	RICH_TRICHOPTERA	Richness/Diversity	Decrease	3.00	1.01	56.58
UNP	PCT_COLLECT	FFG	Increase	80.32	90.97	77.63
UNP	PCT_FILTER	FFG	Decrease	18.14	1.91	73.68
UNP	PCT_SCRAPE	FFG	Decrease	5.72	0.22	77.63
UNP	HBI	Tolerance	Increase	4.01	5.74	84.21
UNP	PCT_URBAN_INTOL	Tolerance	Decrease	98.46	91.71	72.37
					Mean	72.43

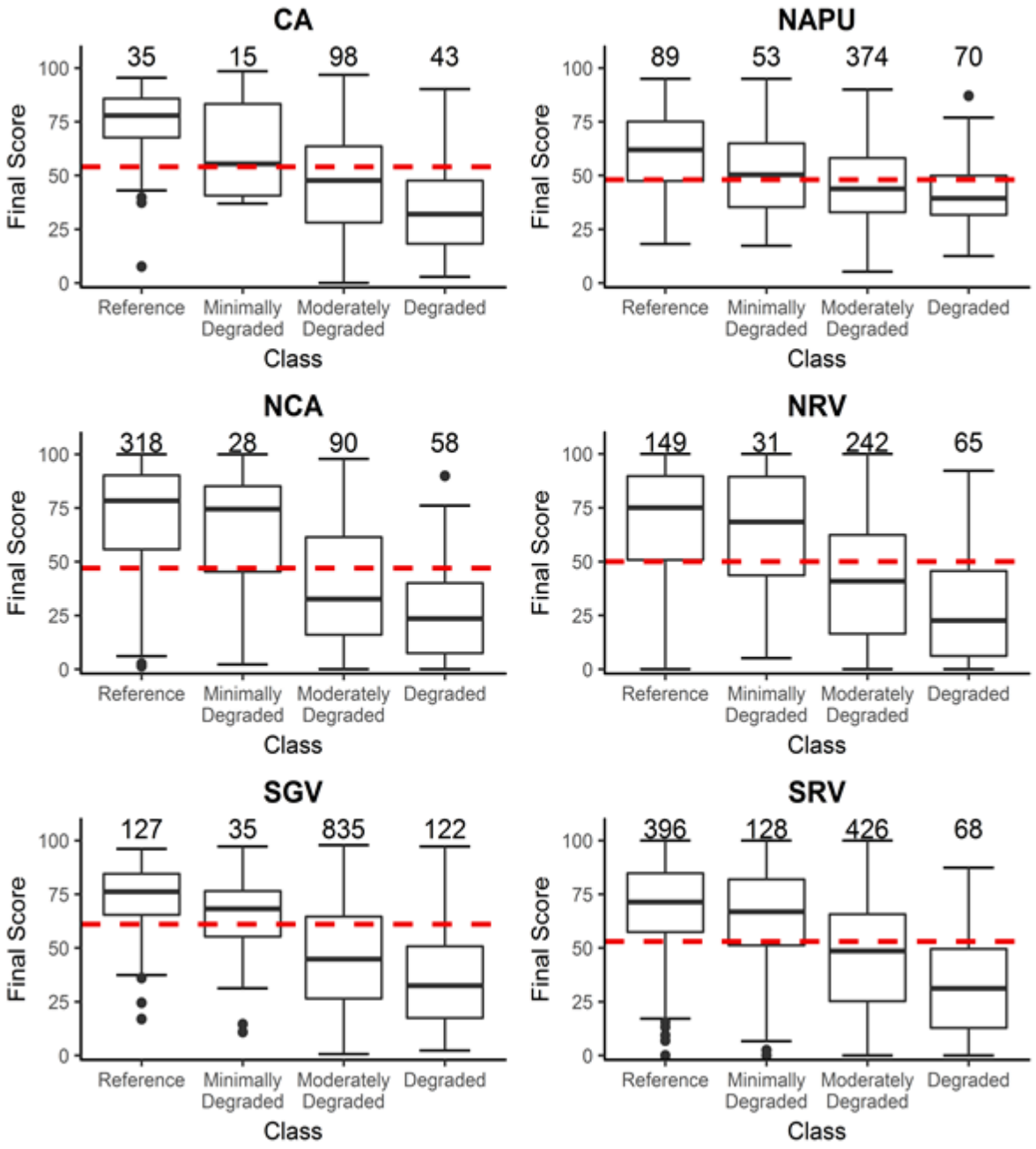


Figure 12. Distributions of index scores for family-level indices in six bioregions: Central Appalachians (CA), Northern Appalachian Plateau and Uplands (NAPU), Northern Central Appalachians (NCA), Northern Ridge and Valley (NRV), Southern Great Valley (SGV), and Southern Ridge and Valley (SRV). The whisker lengths are designated by the Interquartile Range multiplied by 1.5. The horizontal red line represents the Best Separation Point (BSP) between the Reference and Degraded distributions. BSP was used as a threshold for evaluating Classification Efficiency (CE).

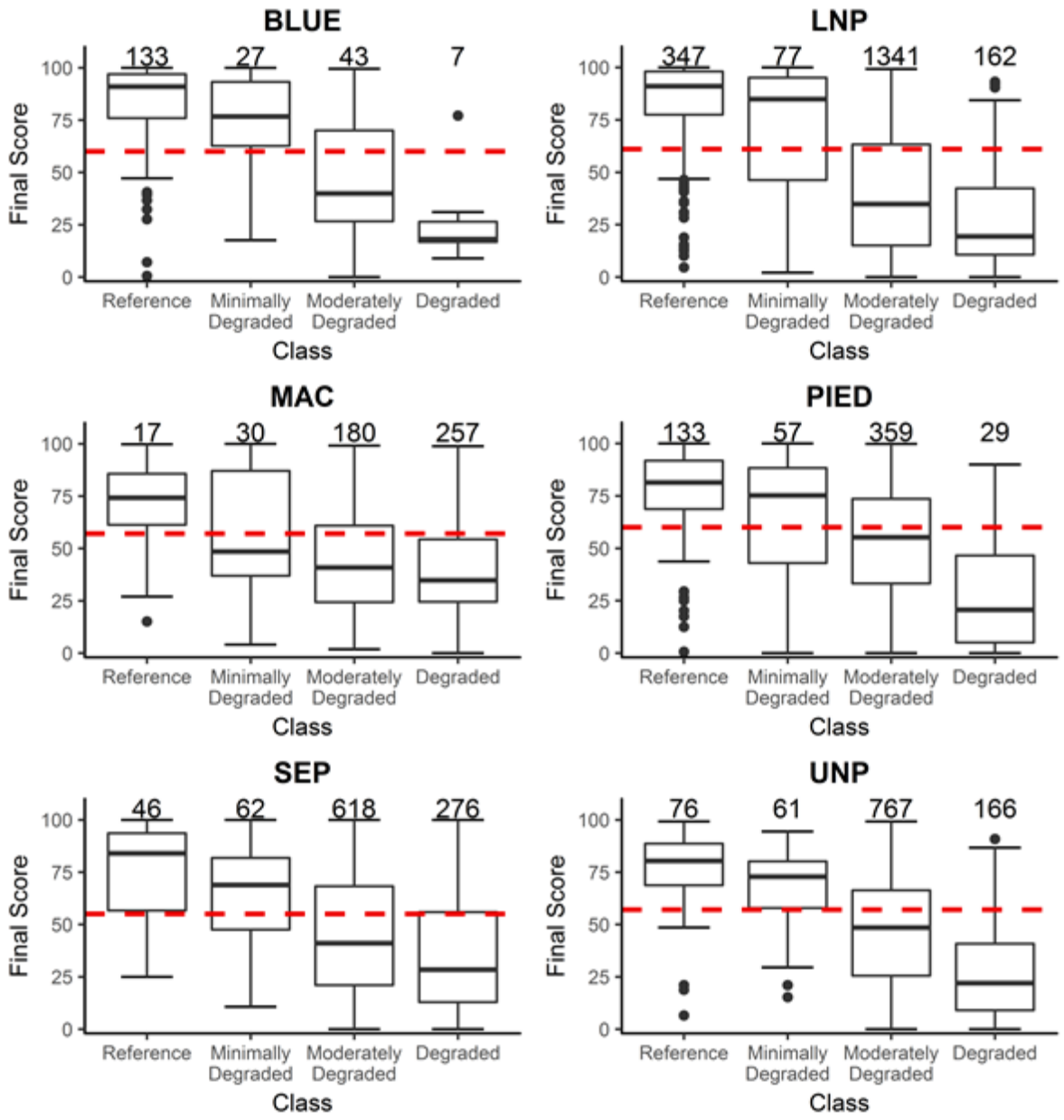


Figure 13. Distributions of index scores for family-level indices in six bioregions: Blue Ridge Mountains (BLUE), Lower Northern Piedmont (LNP), Middle-Atlantic Coast (MAC), Piedmont (PIED), Southeastern Plains (SEP), and Upper Northern Piedmont (UNP). The whisker lengths are designated by the Interquartile Range multiplied by 1.5. The horizontal red line represents the Best Separation Point (BSP) between the Reference and Degraded distributions. BSP was used as a threshold for evaluating Classification Efficiency (CE).

iii. Genus-Level Indices

Sampling events used in this analysis were required to have 90% or more taxa identified to the genus-level. A total of 2,711 sampling events, which were applicable for the order and family-level assessments, had to be omitted because too few taxa were identified to genus (Table 16). Despite the reduction of Reference and Degraded samples, many of the bioregions retained a substantial sample size for index development. However, CA, MAC, and SEP bioregions had fewer than 50 Reference samples, and the BLUE, CA, NAPU, NCA, and PIED had fewer than 50 Degraded samples.

Table 16. Sample numbers in the bioregion condition categories that met the genus-level index requirements.

Bioregion	Reference	Minimally Degraded	Mixed	Moderately Degraded	Degraded	Total Count
BLUE	94 (29.7%)	22 (6.9%)	156 (49.2%)	38 (12%)	7 (2.2%)	317
CA	33 (11.1%)	13 (4.4%)	137 (46.1%)	89 (30%)	25 (8.4%)	297
LNP	207 (6.4%)	57 (1.8%)	1,485 (46.1%)	1,313 (40.7%)	162 (5%)	3,224
MAC	17 (1.1%)	30 (2%)	1,031 (68.1%)	180 (11.9%)	257 (17%)	1,515
NAPU	81 (7.9%)	48 (4.7%)	589 (57.4%)	269 (26.2%)	40 (3.9%)	1,027
NCA	318 (38.3%)	27 (3.3%)	367 (44.2%)	80 (9.6%)	38 (4.6%)	830
NRV	140 (14.7%)	30 (3.2%)	515 (54.2%)	215 (22.6%)	51 (5.4%)	951
PIED	107 (9.4%)	46 (4%)	626 (54.8%)	337 (29.5%)	26 (2.3%)	1,142
SEP	45 (1.4%)	59 (1.8%)	2,270 (69.6%)	611 (18.7%)	276 (8.5%)	3,261
SGV	114 (7.9%)	27 (1.9%)	415 (28.9%)	760 (53%)	119 (8.3%)	1,435
SRV	355 (19.5%)	119 (6.5%)	900 (49.4%)	386 (21.2%)	61 (3.3%)	1,821
UNP	76 (2.7%)	60 (2.2%)	1,732 (62.2%)	749 (26.9%)	166 (6%)	2,783

At least four out of the five metric types were typically represented in each index (Table 17). The fewest metrics (n = 8) were selected for SRV and UNP. The greatest number of metrics (n = 22) were selected for BLUE index. On average, approximately fourteen metrics were selected for each bioregion. The mean BDE of all the metrics selected within the basin was 72.2%. The mean CE was 83.3% but ranged from 72.7% (NAPU) to 89.4% (NRV) (Table 14). The mean BSP ($\bar{x} = 53.0$) did not differ from the expected BSP of 50.0. The lowest BSP was observed in the NRV (x = 46.0), while the greatest BSP was found in the UNP (x = 59.0). Figure 14 and Figure 15 depict the final index scores of the sampling event categories.

Table 17. Metrics included in the bioregion-specific, genus-level indices.

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
BLUE	PCT_HEPTAGENIIDAE	Composition	Decrease	10.89	0.00	87.23
BLUE	PCT_NON_HYDROP_ TRICHOPTERA	Composition	Increase	44.39	68.71	56.38
BLUE	PCT_PERLIDAE	Composition	Decrease	1.60	0.46	67.02
BLUE	PCT_PISCIFORMA	Composition	Decrease	18.54	4.44	78.72
BLUE	PCT_SYSTELLOGNATHA	Composition	Decrease	6.09	0.00	87.23
BLUE	GOLD	Composition	Decrease	82.58	62.48	79.79
BLUE	PCT_COLEOPTERA	Composition	Increase	4.04	5.80	64.89
BLUE	PCT_COTE	Composition	Decrease	64.84	42.52	75.53
BLUE	PCT_EPHEMEROPTERA_ NO_BAETID	Composition	Decrease	36.04	0.70	74.47
BLUE	PCT_PLECOPTERA	Composition	Increase	11.71	23.49	54.26
BLUE	PIELOU	Richness/Diversity	Decrease	0.82	0.73	67.02
BLUE	RICH_PLECOPTERA	Richness/Diversity	Decrease	3.72	1.19	73.40
BLUE	PCT_COLLECT	FFG	Decrease	55.70	44.56	57.45
BLUE	PCT_PREDATOR	FFG	Decrease	8.45	7.11	64.89
BLUE	PCT_SCRAPE	FFG	Increase	21.32	23.56	54.26
BLUE	RICH_FILTER	FFG	Decrease	3.00	1.58	70.21
BLUE	RICH_PREDATOR	FFG	Decrease	3.91	0.86	77.66
BLUE	PCT_BURROW	Habit	Increase	8.24	15.50	69.15
BLUE	PCT_CLING	Habit	Decrease	68.75	52.01	72.34
BLUE	PCT_SWIM	Habit	Decrease	7.80	1.64	71.28
BLUE	RICH_SWIM	Habit	Decrease	1.93	0.00	81.91
BLUE	RICH_MODTOL	Tolerance	Decrease	5.41	3.72	60.64
CA	PCT_SYSTELLOGNATHA	Composition	Decrease	5.30	0.00	81.82
CA	RICH_EPHEM_EPEORUS	Composition	Decrease	3.59	0.00	78.79
CA	PCT_COTE	Composition	Decrease	47.29	22.58	69.70
CA	PCT_EPHEMEROPTERA	Composition	Decrease	19.34	0.00	81.82
CA	PCT_EPT_RICH	Richness/Diversity	Decrease	66.83	42.53	84.85
CA	PIELOU	Richness/Diversity	Decrease	0.82	0.74	69.70
CA	RICH_PLECOPTERA	Richness/Diversity	Decrease	4.00	0.44	75.76
CA	RICH_TRICHOPTERA	Richness/Diversity	Decrease	3.98	1.13	78.79

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
CA	PCT_COLLECT	FFG	Increase	49.08	59.80	63.64
CA	PCT_FILTER	FFG	Decrease	17.24	8.35	63.64
CA	PCT_PREDATOR	FFG	Decrease	11.23	5.50	69.70
CA	PCT_SCRAPE	FFG	Decrease	11.32	0.00	81.82
CA	RICH_PREDATOR	FFG	Decrease	4.65	1.64	75.76
CA	RICH_SPRAWL	Habit	Increase	2.85	4.39	63.64
CA	PCT_MOD_TOL_4_6	Tolerance	Increase	38.56	52.32	69.70
CA	PCT_URBAN_INTOL	Tolerance	Decrease	98.72	96.60	69.70
LNP	PCT_CHIRONOMINI	Composition	Increase	0.00	0.57	89.86
LNP	PCT_EPT_NO_HYDRO	Composition	Decrease	90.82	83.38	50.24
LNP	PCT_PTERYGOTA	Composition	Decrease	97.56	91.66	66.67
LNP	PCT_COTE	Composition	Decrease	56.16	9.00	87.44
LNP	PCT_EPHEMEROPTERA_NO_BAETID	Composition	Decrease	22.62	0.00	88.89
LNP	PCT_COLLECT	FFG	Increase	66.22	81.55	76.81
LNP	PCT_FILTER	FFG	Decrease	15.37	6.02	74.88
LNP	PCT_PREDATOR	FFG	Decrease	8.89	4.88	77.78
LNP	RICH_BURROW	Habit	Increase	1.00	2.24	71.50
LNP	HBI	Tolerance	Increase	4.04	5.86	90.82
LNP	PCT_MOD_TOL_4_6	Tolerance	Increase	46.58	72.02	81.64
LNP	RICH_TOL	Tolerance	Increase	1.00	2.24	59.90
MAC	PCT_CAECIDOTEA	Composition	Increase	0.00	0.92	88.24
MAC	PCT_CHIRONOMINAE	Composition	Increase	0.00	1.30	76.47
MAC	PCT_ORTHOCLADIINAE	Composition	Increase	0.00	2.48	82.35
MAC	PCT_ARTHROPODA	Composition	Increase	74.11	94.94	70.59
MAC	PCT_ODONATA	Composition	Decrease	1.82	0.33	76.47
MAC	HURLBERTS_PIE	Richness/Diversity	Increase	0.56	0.75	76.47
MAC	PCT_EPT_RICH	Richness/Diversity	Increase	0.00	5.82	76.47
MAC	PCT_SCRAPE	FFG	Decrease	12.06	3.43	70.59
MAC	RICH_GATHER	FFG	Increase	1.00	7.13	76.47
MAC	PCT_BURROW	Habit	Decrease	52.33	20.85	70.59
MAC	PCT_URBAN_INTOL	Tolerance	Increase	65.29	71.06	76.47
MAC	RICH_TOL	Tolerance	Increase	3.00	4.78	76.47
NAPU	PCT_EPT_NO_HYDRO	Composition	Decrease	87.99	71.13	65.43
NAPU	PCT_HYDRO_TRICHOPTERA	Composition	Increase	58.02	75.51	66.67
NAPU	PCT_EPHEMEROPTERA	Composition	Decrease	32.35	10.57	69.14
NAPU	PCT_EPT_RICH	Richness/Diversity	Decrease	70.91	37.28	71.60
NAPU	PCT_EPT_RICH_NO_TOL	Richness/Diversity	Decrease	48.92	29.89	74.07
NAPU	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	6.00	3.56	79.01
NAPU	PCT_COLLECT	FFG	Increase	54.65	66.81	62.96
NAPU	PCT_FILTER	FFG	Increase	20.88	26.58	61.73

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
NAPU	PCT_SCRAPE	FFG	Decrease	20.38	12.16	71.60
NAPU	RICH_SHRED	FFG	Decrease	2.00	0.85	54.32
NAPU	RICH_CLING	Habit	Decrease	11.67	8.32	76.54
NAPU	PCT_INTOL_0_4	Tolerance	Decrease	60.97	44.68	74.07
NAPU	RICH_INTOL	Tolerance	Decrease	9.00	3.14	65.43
NCA	PCT_PISCIFORMA	Composition	Decrease	19.63	0.00	85.22
NCA	PCT_SYSTELLOGNATHA	Composition	Decrease	6.70	1.33	75.16
NCA	RICH_EPHEM_EPEORUS	Composition	Decrease	6.00	1.61	85.22
NCA	PCT_COTE	Composition	Decrease	64.02	6.78	87.74
NCA	PCT_LIMESTONE	Composition	Decrease	8.96	0.00	82.39
NCA	PCT_PLECOPTERA	Composition	Increase	15.18	31.05	77.67
NCA	RICH_EPT_NO_TOL	Richness/Diversity	Decrease	8.91	6.37	76.10
NCA	PCT_SCRAPE	FFG	Decrease	25.20	0.00	85.22
NCA	RICH_GATHER	FFG	Decrease	4.00	1.87	87.11
NCA	RICH_SCRAPE	FFG	Decrease	5.00	0.93	88.36
NCA	PCT_CLING	Habit	Decrease	63.16	36.22	78.93
NCA	PCT_SWIM	Habit	Decrease	10.11	0.00	84.91
NCA	RICH_SWIM	Habit	Decrease	2.00	1.60	74.21
NRV	PCT_NON_HYDROP_ TRICHOPTERA	Composition	Decrease	41.58	10.75	72.86
NRV	PCT_PISCIFORMA	Composition	Decrease	13.88	0.00	74.29
NRV	PCT_EPHEMEROPTERA_ NO_BAETID	Composition	Decrease	23.77	0.51	75.00
NRV	PCT_EPT_RICH	Richness/Diversity	Decrease	69.10	37.52	85.00
NRV	RICH_TRICHOPTERA	Richness/Diversity	Decrease	4.00	1.18	82.14
NRV	RICH_COLLECT	FFG	Decrease	8.00	4.87	71.43
NRV	RICH_FILTER	FFG	Decrease	3.61	1.77	77.14
NRV	RICH_PREDATOR	FFG	Decrease	3.86	0.00	79.29
NRV	RICH_SWIM	Habit	Decrease	2.00	1.26	72.86
NRV	PCT_TOLERANT_5_10	Tolerance	Increase	33.46	77.14	74.29
NRV	RICH_INTOL	Tolerance	Decrease	11.00	0.00	90.00
PIED	PCT_NON_HYDROP_ TRICHOPTERA	Composition	Decrease	53.18	30.96	45.79
PIED	GOLD	Composition	Decrease	80.07	47.02	82.24
PIED	PCT_EPHEMEROPTERA	Composition	Decrease	33.40	14.01	75.70
PIED	PIELOU	Richness/Diversity	Decrease	0.84	0.74	59.81
PIED	PCT_COLLECT	FFG	Increase	62.29	77.61	73.83
PIED	PCT_FILTER	FFG	Increase	16.14	18.84	56.07
PIED	PCT_PREDATOR	FFG	Increase	9.58	10.70	51.40
PIED	PCT_SCRAPE	FFG	Decrease	19.85	4.42	76.64
PIED	PCT_SWIM	Habit	Decrease	8.30	3.41	66.36
PIED	RICH_BURROW	Habit	Increase	1.00	2.57	63.55

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
PIED	HBI	Tolerance	Increase	3.97	5.55	85.98
PIED	PCT_DOM1	Tolerance	Increase	46.00	67.85	41.12
PIED	PCT_MOD_TOL_4_6	Tolerance	Increase	55.26	63.70	68.22
PIED	PCT_TOLERANT_7_10	Tolerance	Increase	1.76	7.15	77.57
PIED	PCT_URBAN_INTOL	Tolerance	Decrease	94.47	84.42	68.22
PIED	RICH_TOL	Tolerance	Increase	1.06	2.69	46.73
SEP	PCT_COTE	Composition	Decrease	25.64	2.15	80.00
SEP	PCT_OLIGO_CHIRO	Composition	Increase	28.68	53.44	75.56
SEP	PCT_COLLECT	FFG	Decrease	77.12	67.84	53.33
SEP	PCT_PREDATOR	FFG	Increase	6.43	9.02	55.56
SEP	RICH_PREDATOR	FFG	Increase	2.04	4.25	80.00
SEP	PCT_CLING	Habit	Decrease	47.19	0.00	71.11
SEP	RICH_BURROW	Habit	Increase	2.00	3.58	71.11
SEP	RICH_CLIMB	Habit	Increase	1.00	2.67	82.22
SEP	HBI	Tolerance	Increase	5.23	5.82	73.33
SEP	PCT_MOD_TOL_4_6	Tolerance	Decrease	74.30	59.75	51.11
SEP	PCT_TOLERANT_7_10	Tolerance	Increase	12.01	19.79	71.11
SEP	RICH_TOL	Tolerance	Increase	2.35	5.05	82.22
SGV	PCT_EPT_ CHEUMATOPSYCHE	Composition	Decrease	59.39	20.35	84.21
SGV	PCT_EPT_NO_HYDRO	Composition	Decrease	82.45	73.71	64.91
SGV	PCT_HEPTAGENIIDAE	Composition	Decrease	8.66	0.00	82.46
SGV	PCT_HYDRO_ TRICHOPTERA	Composition	Decrease	69.94	64.01	45.61
SGV	PCT_MALACOSTRACA	Composition	Increase	0.00	0.44	68.42
SGV	PCT_NON_HYDROP_ TRICHOPTERA	Composition	Decrease	28.28	12.28	64.04
SGV	PCT_PHILOPOTAMIDAE	Composition	Decrease	1.93	0.00	78.07
SGV	PCT_PTERYGOTA	Composition	Decrease	96.05	88.28	68.42
SGV	PCT_RETREAT_ CADDISFLY	Composition	Decrease	17.33	2.64	63.16
SGV	GOLD	Composition	Decrease	79.99	65.76	61.40
SGV	PCT_COTE	Composition	Decrease	68.28	42.94	64.04
SGV	PIELOU	Richness/Diversity	Decrease	0.80	0.72	51.75
SGV	PCT_COLLECT	FFG	Increase	68.10	81.55	71.05
SGV	PCT_PREDATOR	FFG	Decrease	5.33	1.77	66.67
SGV	PCT_SCRAPE	FFG	Decrease	18.63	7.80	66.67
SGV	PCT_BURROW	Habit	Increase	11.23	15.12	47.37
SGV	PCT_MOD_TOL_4_6	Tolerance	Increase	53.27	68.03	66.67
SGV	PCT_TOLERANT_7_10	Tolerance	Increase	1.79	4.98	78.07
SGV	PCT_URBAN_INTOL	Tolerance	Decrease	92.82	90.44	63.16
SRV	PCT_COTE	Composition	Decrease	57.61	43.77	69.01

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
SRV	PCT_EPHEMEROPTERA	Composition	Decrease	30.27	12.89	69.58
SRV	PCT_EPT_RICH_NO_TOL	Richness/Diversity	Decrease	42.34	28.94	58.31
SRV	PIELOU	Richness/Diversity	Decrease	0.83	0.74	60.28
SRV	PCT_SCRAPE	FFG	Decrease	13.61	6.87	70.42
SRV	PCT_CLING	Habit	Decrease	54.40	38.98	70.99
SRV	PCT_SWIM	Habit	Decrease	10.41	5.28	60.85
SRV	PCT_INTOL_0_3	Tolerance	Decrease	49.74	24.77	70.42
UNP	PCT_PISCIFORMA	Composition	Decrease	9.35	0.00	80.26
UNP	PCT_RETREAT_ CADDISFLY	Composition	Decrease	10.98	2.76	73.68
UNP	PCT_EPHEMEROPTERA	Composition	Decrease	24.29	0.00	78.95
UNP	PCT_HEXAPODA	Composition	Decrease	99.50	96.66	81.58
UNP	RICH_EPT_NO_TOL	Richness/Diversity	Decrease	5.00	1.40	90.79
UNP	PCT_CLING	Habit	Decrease	60.04	24.16	76.32
UNP	RICH_SWIM	Habit	Decrease	1.14	0.00	89.47
UNP	PCT_INTOL_0_4	Tolerance	Decrease	58.72	13.43	81.58
					Mean	72.21

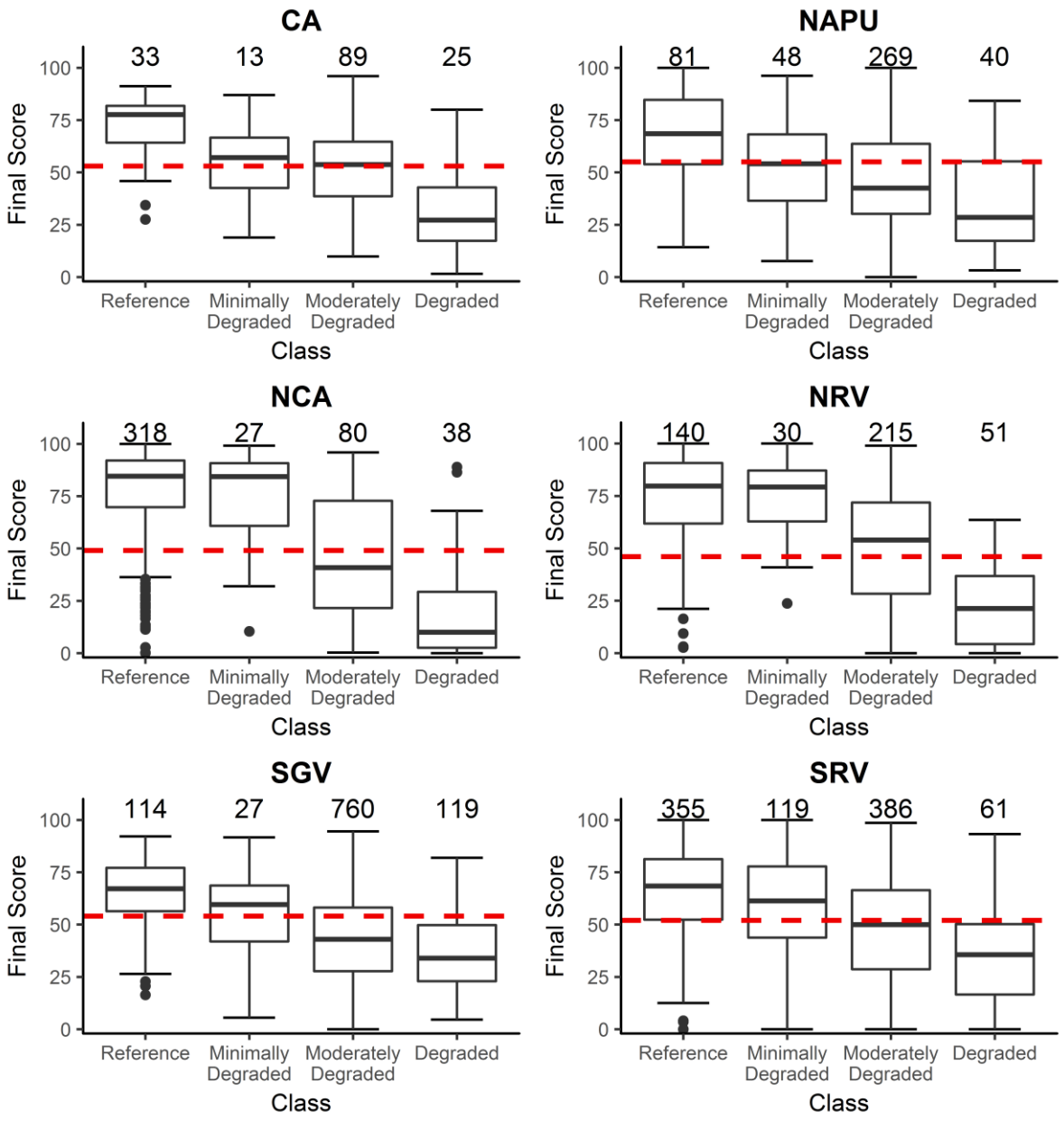


Figure 14. Distributions of index scores for genus-level indices in six bioregions: Central Appalachians (CA), Northern Appalachian Plateau and Uplands (NAPU), Northern Central Appalachians (NCA), Northern Ridge and Valley (NRV), Southern Great Valley (SGV), and Southern Ridge and Valley (SRV). The whisker lengths are designated by the Interquartile Range multiplied by 1.5. The horizontal red line represents the Best Separation Point (BSP) between the Reference and Degraded distributions. BSP was used as a threshold for evaluating Classification Efficiency (CE).

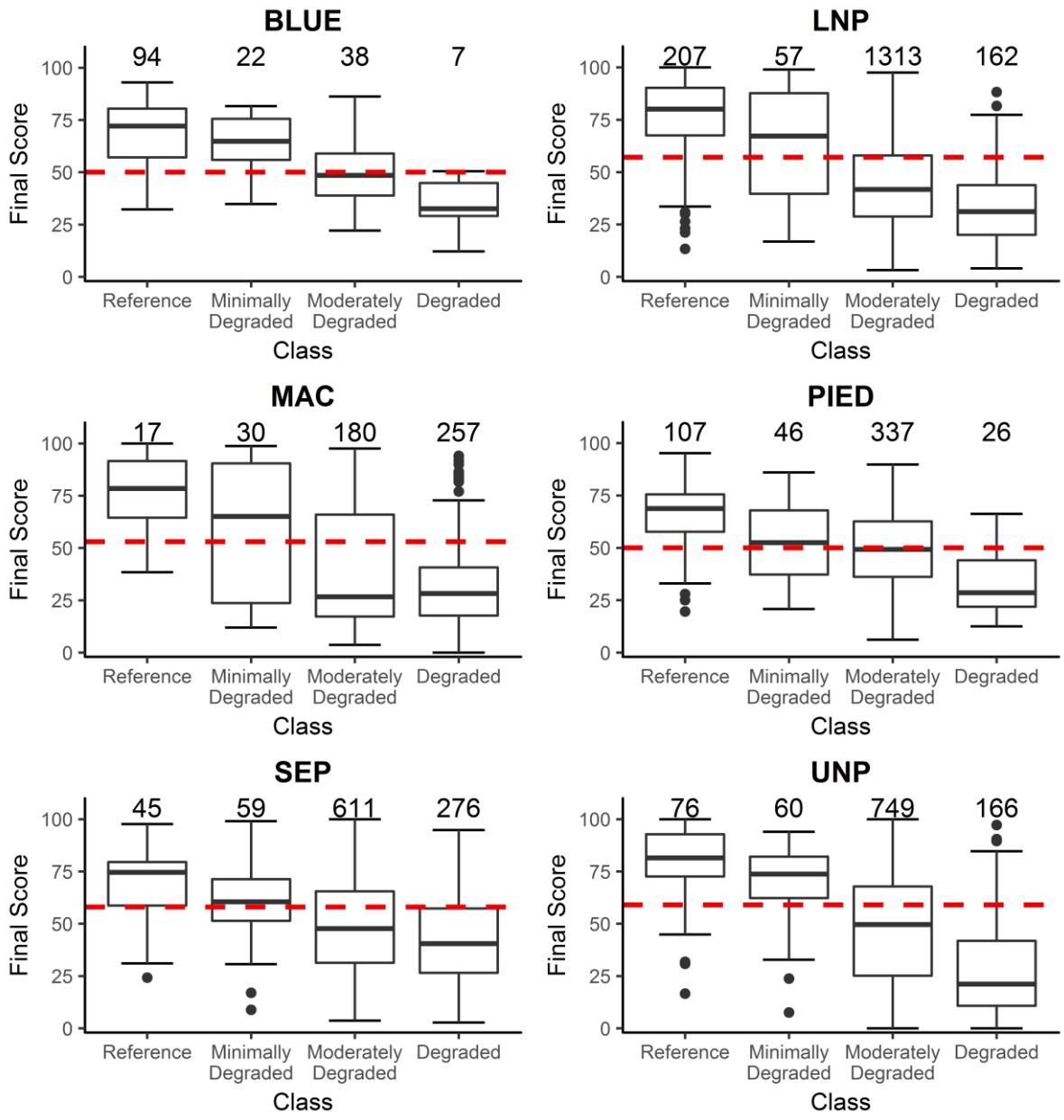


Figure 15. Distributions of index scores for genus-level indices in six bioregions: Blue Ridge Mountains (BLUE), Lower Northern Piedmont (LNP), Middle-Atlantic Coast (MAC), Piedmont (PIED), Southeastern Plains (SEP), and Upper Northern Piedmont (UNP). The whisker lengths are designated by the Interquartile Range multiplied by 1.5. The horizontal red line represents the Best Separation Point (BSP) between the Reference and Degraded distributions. BSP was used as a threshold for evaluating Classification Efficiency (CE).

E. **Index Validation and Precision**

Validation and precision tests on the various indices are described in Appendix K. The Chesapeake-wide and region indices had similar RMSE values at all three taxonomic resolutions (i.e., delete-d jackknife Cross Validation RMSE). The family-level was the preferred choice because its CE value was high (≥ 70%) and the family-level CE was generally greater than or comparable to the genus-level CE.

In general, the average RMSE associated with the bioregion index performance did not differ greatly among the three taxonomic resolutions (Table K-4). The average CE increased as taxonomic resolution increased from the order- to genus-level. Eight of the twelve bioregions (i.e., BLUE, CA, MAC, NAPU, NCA, NRV, PIED, and SEP) had higher CE at the Genus-level compared to the Family-level. However, the CE improvement from the family to the genus-level was minor ($\bar{x} = 1.7\%$), while the CE improvement from the order- to the family-level was larger ($\bar{x} = 7.4\%$). Overall, genus-level indices did not provide substantial improvement over the family-level indices.

F. **Narrative Ratings**

We compared this study's bioregion family-level Chessie BIBI ratings to the family-level ratings produced by Buchanan et al. (2011) to see how the multiple changes in methodology affected ratings of individual samples. Buchanan et al. (2011) worked with 487 Reference sampling events assigned to five inland bioregions compared to the 1,803 Reference sampling events assigned to ten inland bioregions in this study. They did not develop a Coast index, but instead used the existing Coastal Plain Macroinvertebrate Index (CPMI) developed by Maxted et al. (2000). Their water quality and stream habitat criteria for inland Reference and Degraded conditions were slightly different. Metrics in the current study were scored on a somewhat wider gradient, and CEs were determined with each index's calculated BSP rather than an assumed BSP of 50. Thresholds for narrative ratings in the 2011 report were determined using averages of the five inland bioregions' 50th, 25th, 10th, and 5th percentiles of Reference index scores. Thresholds in this study were determined for each individual bioregion and region with the Reference 50th, 25th, and 10th percentiles, and a value representing half of the 10th Reference percentile.

Using sample event ID, the 2011 family-level index ratings were paired with the corresponding family-level bioregion and region index ratings produced in this study. The twelve family-level bioregion indices compare well with the 2011 ratings (Table 18A). For the ten Inland bioregions, 84.6% of the ratings matched exactly or differed by only one rating (match = 47.2%, near = 37.4%). For the two Coast bioregions, 85.8% matched exactly or differed by only one rating (match = 50.2%, near = 35.6%). Just 15.4% of the Inland and 14.2% of the Coast ratings disagreed (differed by more than one rating on the 5-ratings scale). When the Inland ratings disagreed, this study's ratings tended to be somewhat better. When the Coast ratings disagreed, this study's ratings tended to be somewhat poorer.

The family-level region index ratings were also similar to the 2011 ratings (Table 18B). The Inland index ratings matched exactly or differed by only one rating 89.4% of the sampling events (match = 58.0%, near = 31.4%). For the Coast index, 93.2% of the ratings matched exactly or differed by one rating (match = 53.4%, near = 39.8%). Just 10.6% of the Inland and 6.8% of the Coast ratings disagreed by more than one rating.

Table 18. Comparison of narrative ratings for sampling events used in Buchanan et al. (2011). The 2011 ratings were determined with methods described in Buchanan et al. (2011). This study's ratings were determined with the methodology described above. Ratings that match exactly are highlighted in dark blue; those that differ by only one rating are highlighted in light blue.

A.

This Study's Family-Level Bioregion Ratings

Family-Level 2011 Bioregion Ratings: **Coast**	Excellent	Good	Fair	Poor	Very Poor
Excellent	8.3%	4.9%	1.3%	0.6%	0.0%
Good	1.1%	9.7%	3.7%	2.3%	0.4%
Fair	0.6%	2.9%	10.4%	5.8%	2.9%
Poor	0.5%	0.8%	4.3%	7.8%	3.8%
Very Poor	0.4%	0.8%	3.2%	9.1%	14.0%

Family-Level 2011 Bioregion Ratings: **Inland**	Excellent	Good	Fair	Poor	Very Poor
Excellent	8.7%	3.7%	1.1%	0.7%	0.0%
Good	3.0%	3.1%	2.2%	1.5%	0.1%
Fair	2.2%	3.2%	2.9%	4.2%	0.7%
Poor	1.1%	2.1%	2.6%	5.1%	2.1%
Very Poor	0.9%	2.0%	2.9%	16.4%	27.4%

B.

This Study's Family-Level Region Ratings

Family-Level 2011 Bioregion Ratings: **Coast**	Excellent	Good	Fair	Poor	Very Poor
Excellent	12.2%	2.8%	0.1%	0.0%	0.0%
Good	3.8%	10.9%	2.2%	0.4%	0.0%
Fair	0.2%	6.8%	8.0%	6.1%	1.5%
Poor	0.0%	1.7%	4.8%	8.2%	2.6%
Very Poor	0.0%	0.3%	2.5%	10.7%	14.1%

Family-Level 2011 Bioregion Ratings: **Inland**	Excellent	Good	Fair	Poor	Very Poor
Excellent	9.4%	3.0%	1.4%	0.4%	0.0%
Good	3.3%	3.0%	2.4%	1.1%	0.1%
Fair	1.6%	2.9%	3.9%	3.8%	1.0%
Poor	0.3%	1.4%	2.8%	5.6%	3.0%
Very Poor	0.1%	0.7%	2.5%	10.2%	36.1%

Many habitat and water quality parameters used to classify Reference conditions correlate positively with %Forest and negatively with %Urban at the HUC12-level (Table L-4). We compared the narrative ratings of the family-level bioregion and family-level region indices to six land use categories along a disturbance gradient, to see if ratings corresponded to increasing anthropogenic impacts in their watersheds. Land use characteristics of the HUC12 watersheds

(*Retrieved [02/24/2017], from USEPA Recovery Potential Screening Tools, https://www.epa.gov/rps/recovery-potential-screening-tools-downloadable-tools-comparing-watersheds.*) were grouped by their percentages of forest, urban, and agricultural area. The categories were: 1) >78%Forest + <5% Urban (least disturbed); 2) 50% - 78% Forest + <5% Urban (minimally disturbed); 3) Other; 4) <25% Forest + >50% Agriculture (disturbed); 5) <25% Forest + >20% Urban (disturbed); and 6) <25% Forest + >20% Urban + >50% Agriculture (most disturbed). The ratings of individual sample events were paired with their HUC12 watershed's land use category. Table 19 shows the percentages of Excellent, Good, and Fair (EGF), as well as, Poor and Very Poor (PVP) ratings of the family-level indices in each of the six land use categories. Results for the bioregion indices are grouped by region for brevity. The largest percentages of Excellent, Good, and Fair ratings (EGF) occurred in the least disturbed watersheds; large percentages of Poor and Very Poor ratings (PVP) occurred in the most disturbed watersheds.

*Table 19. Narrative ratings of the bioregion and region versions of the family-level index in six different intensities of watershed disturbance. Percentages reflect the proportions of EGF (excellent, good, fair) and PVP (poor, very poor) ratings in each disturbance gradient category. For, forest; Urb, urban; Agr, agriculture; Other, land uses do not meet criteria of other disturbance gradient categories. *The bioregion index ratings are grouped by region for comparison purposes.* ***Bolded****, percentages greater than 60%.*

		Least Disturbed			Most Disturbed		
		>78%For + <5% Urb	50%-78%For + <5%Urb	<25%For + >50%Agr	<25%For + >20%Urb	<25%For + >20%Urb + >50%Agr	Other (Moderate Disturbance)
Family-Level Bioregion Index*							
Coast	n	*0*	*161*	*805*	*907*	*0*	*3,331*
	EGF		**84.47%**	40.87%	27.34%		65.93%
	PVP		15.53%	59.13%	**72.66%**		34.07%
Inland	n	*2,407*	*1,673*	*593*	*1,050*	*154*	*10,232*
	EGF	**80.31%**	**68.86%**	19.22%	6.29%	1.95%	44.24%
	PVP	19.69%	31.14%	**80.78%**	**93.71%**	**98.05%**	55.76%
Family-Level Region Index							
Coast	n		*161*	*805*	*907*		*3,331*
	EGF		**90.68%**	58.51%	27.45%		67.31%
	PVP		9.32%	41.49%	**72.55%**		32.69%
Inland	n	*2,407*	*1,673*	*593*	*1,050*	*154*	*10,232*
	EGF	**82.47%**	**64.85%**	18.72%	5.71%	2.60%	40.94%
	PVP	17.53%	35.15%	**81.28%**	**94.29%**	**97.40%**	59.06%

We cross-compared the narrative ratings produced for two of the spatial indices (region, bioregion) and their three taxonomic versions (Table 20). Different versions of the index that show little disagreement (e.g., ≤ 10%) can potentially be used together. Ratings were paired on sample event ID and the percentages of paired ratings that exactly match, nearly match (i.e., are adjacent ratings), and disagree (i.e., separated by one or more ratings) were determined. Family-level versions of the bioregion index (results are grouped by region) and the region index show the least overall disagreement, with 8.7% in the Coast region and 7.4% in the Inland region (Table 20A). Roughly half of the ratings matched exactly. The genus-level bioregion index (grouped by region) also showed little disagreement (7.2%) with the Inland genus-level index. Other comparisons of the region and bioregion taxonomic versions were not as strong. For taxonomic versions of the region index, the Inland family- and genus-level indices compared well, with only 9.7% disagreement (Table 20B). For taxonomic version of the bioregion index (grouped by region), Coast order- and family-level indices and Inland family- and genus-level indices compared well, with only 9.1% and 8.6% disagreement, respectively (Table 20C). The family-level versions of the bioregion index and the region index (separated into bioregions) showed less than 10% disagreement in eight of the twelve bioregions: SEP, CA, NRV, LNP, SRV, NCA, SGV, and PIED (Table 20D).

Table 20. Cross-comparisons of ratings of the two spatial (i.e., region and bioregion) and three taxonomic (i.e., order-, family-, and genus-) versions of the index. Agreement: match, exact agreement (e.g., Fair and Fair); near, differ by one rating (e.g., Good and Fair); disagree, differ by more than one rating (e.g., Excellent and Fair). Highlighted cross-comparisons have 10% or fewer rating disagreements. Results are grouped by Coast and Inland regions in A – C; results are grouped by bioregion in D. A) Ratings produced by the three taxonomic versions of the two regional indices are compared by region to ratings produced by the three taxonomic versions of the twelve bioregion indices. B) Ratings produced by the genus-, family-, and order-level versions of the regional indices (Coast, Inland) are compared. C) Ratings produced by the genus-, family-, and order-level versions of the twelve bioregion indices are grouped by region for brevity and compared. D) Ratings produced by the family-level versions of the regional and bioregion indices are compared by bioregion.

A) Region vs Bioregion Indices

			Region Index					
			Order		Family		Genus	
		Agreement	Coast	Inland	Coast	Inland	Coast	Inland
Bioregion Index	Order	Match	33.3%	48.3%	43.1%	38.6%	25.1%	44.2%
		Near	39.3%	36.5%	41.1%	36.3%	29.7%	38.7%
		Disagree	27.4%	15.2%	15.9%	25.1%	45.2%	17.1%
	Family	Match	29.2%	47.1%	48.0%	56.3%	21.5%	51.0%
		Near	45.9%	37.8%	43.3%	36.2%	36.7%	37.8%
		Disagree	24.8%	15.1%	**8.7%**	**7.4%**	41.8%	11.3%
	Genus	Match	25.4%	44.1%	22.3%	45.6%	44.5%	54.3%
		Near	37.1%	41.8%	37.9%	42.5%	44.1%	38.5%
		Disagree	37.5%	14.1%	39.8%	11.9%	11.4%	**7.2%**

Table 20. Continued…

B) Taxonomic Versions of the Region Indices

			Region Index			
			Order		Family	
		Agreement	Coast	Inland	Coast	Inland
Region Index	Family	Match	32.1%	50.7%		
		Near	44.2%	36.5%		
		Disagree	23.6%	12.8%		
	Genus	Match	24.5%	52.7%	24.4%	54.3%
		Near	36.6%	37.1%	33.7%	36.0%
		Disagree	38.8%	10.2%	42.0%	**9.7%**

C) Taxonomic Versions of the Bioregion Indices

			Bioregion Index			
			Order		Family	
		Agreement	Coast	Inland	Coast	Inland
Bioregion Index	Family	Match	48.3%	43.0%		
		Near	42.7%	36.2%		
		Disagree	**9.1%**	20.9%		
	Genus	Match	28.5%	39.8%	26.7%	52.8%
		Near	33.3%	40.5%	40.4%	38.6%
		Disagree	38.2%	19.8%	32.9%	**8.6%**

D) Family-Level Versions of the Bioregion vs Region Indices

		Bioregion Index										
	Agreement	MAC	SEP	CA	NRV	UNP	LNP	SRV	NCA	NAPU	SGV	BLUE	PIED
Region Index	Match	30.3%	55.9%	52.9%	55.4%	49.3%	69.7%	59.6%	51.4%	43.2%	59.3%	40.9%	52.2%
	Near	47.0%	41.7%	39.2%	37.6%	37.3%	28.2%	37.0%	41.5%	44.4%	36.9%	40.7%	38.5%
	Disagree	22.7%	**2.4%**	**7.9%**	**7.0%**	13.4%	**2.1%**	**3.5%**	**7.1%**	12.4%	**3.8%**	18.4%	**9.3%**

Note: this table corresponds to the central cell of Table 20A above, where bioregion results are summed by region for brevity. MAC and SEP bioregions are in the Coast region; the remaining ten bioregions are in the Inland region.

G. Chesapeake Watershed Stream Health

A simple count of the narrative ratings in the Chesapeake Bay watershed indicates biological integrity is Very Poor or Poor at 49.5% of sampling sites and Fair, Good, or Excellent at 50.5% of sites in the entire, updated database (1992 – 2015). These straightforward counts are misleading because some areas—especially urban ones around Washington, D.C.—are more frequently sampled than others in the Chesapeake Bay watershed. To avoid this spatial bias, station ratings of the available data were weighted by the proportion of their local subwatershed (i.e., HUC12) area they represent, and the weighted ratings summed to bioregion or region (*II. N. Area-Weighting of Rating Results*).

The area-weighted station ratings were mapped by HUC12 units to visually examine and evaluate the spatial distributions of the rating results of the various indices (Appendix L). The three taxonomic versions of the Chesapeake-wide index indiscriminately rate the entire Coast region and broad swaths of urban and suburban lands as Poor or Very Poor despite a substantial number of identified Reference sites and several relatively undisturbed HUC12 watersheds in these areas (Figures L-3, L-4, and L-5). The region and bioregion indices were, to varying degrees, better able to identify disturbed and undisturbed areas (Figures L-6 to L-11).

We noted discrepancies in the Coast's MAC bioregion that were consistent with the relatively large rating disagreements between and within the region and bioregion indices (Table 20). The MAC bioregion was rated as a mix of Good, Fair, and Poor by the family-level version of the Coast region index (Figure L-7), Good by the genus-level version of the Coast region index (Figure L-8), mostly Poor by the family-level version of the MAC bioregion index (Figure L-10), and mostly Fair by the genus-level version of the MAC bioregion index (L-11). Rating discrepancies are also apparent in the Inland region's CA and NAPU bioregions, but are not as striking as in MAC.

Area-weighting index ratings by HUC12 area removes a bias created by the uneven spatial distribution of sampling stations. The weighted ratings for Excellent, Good, Fair, Poor, and Very Poor can then be aggregated to estimate the percentages of each in the selected spatial scale. Figure 16 shows the proportions of the area-weighted ratings for the family-level Chesapeake-wide index, the family-level Coast and Inland indices, and the twelve family-level bioregion indices. The region and bioregion results can be further aggregated at the basin-level to compare results of the three index types on the whole Chesapeake Bay watershed (Figure 17).

The relative insensitivity of the Chesapeake-wide index to the complex topography and hydrology of the Chesapeake watershed is evident in the Figure 17 comparison, where that index scores substantially more of the watershed as Very Poor. Results of the region and bioregion indices are more comparable, with 39.5% and 37.2% scoring Poor or Very Poor, respectively.

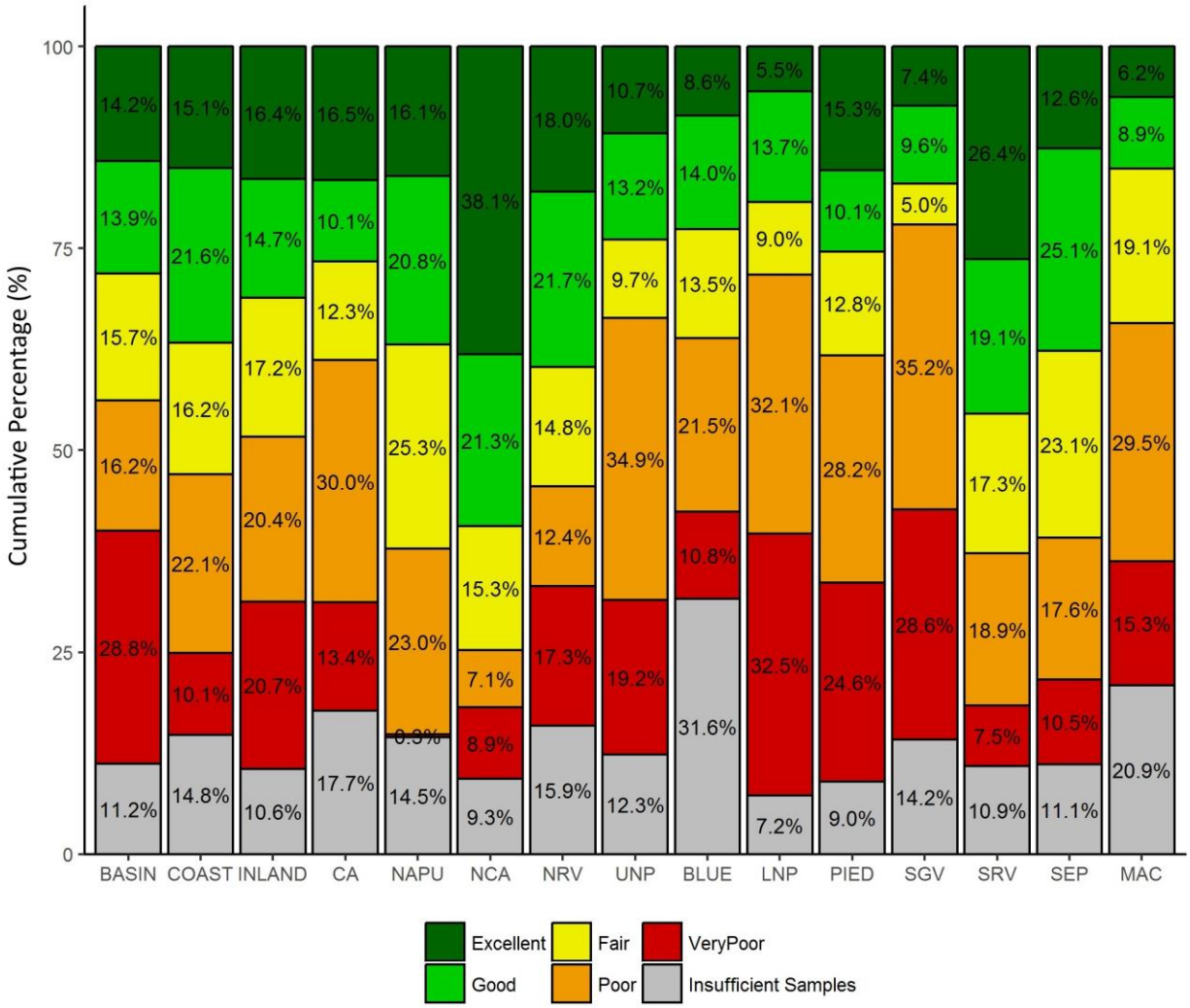
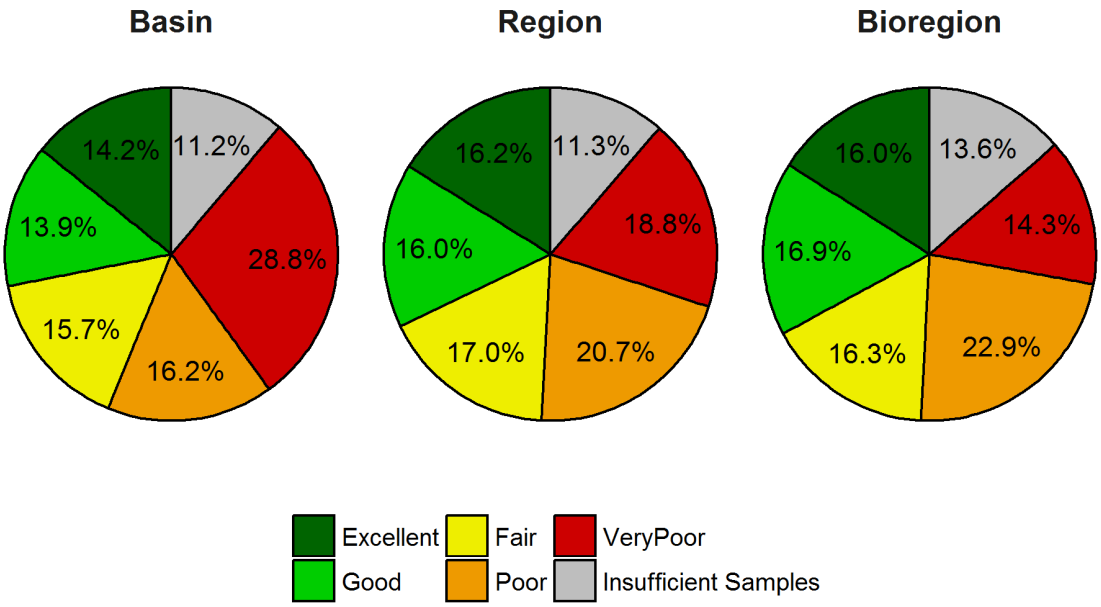


Figure 16. Area-weighted percentages of the five narrative ratings derived with the family-level versions of the Chesapeake-wide index (i.e., Basin), the two region indices (i.e., Coast and Inland), and the twelve bioregion indices (i.e., CA, NAPU, NCA, NRV, UNP, BLUE, LNP, PIED, SGV, SEP, and MAC).

A.



B.

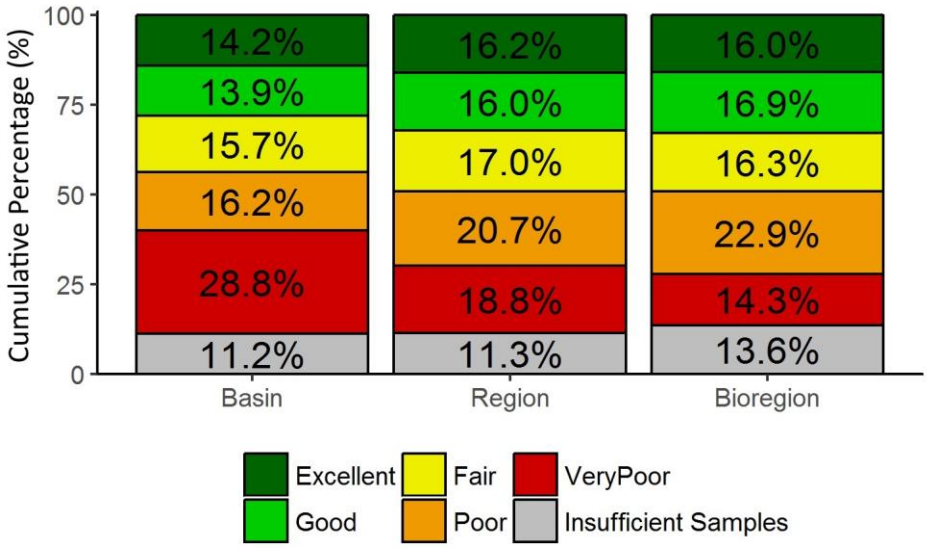


Figure 17. Area-weighted percentages of the five ratings derived with the family-level versions of the Chesapeake-wide, Region, and Bioregion indices and rolled up to the Chesapeake watershed scale. The data is plot using two graphing techniques (i.e., A. pie charts and B. stacked bar plots) to aid in visual interpretation of the results.

IV. Discussion

The objective of this study was to refine the existing Chessie BIBI, a stream benthic macroinvertebrate index of biotic integrity that evaluates stream biological communities with a common rating approach across the Chesapeake Bay watershed. The study is more rigorous than earlier attempts by Foreman et al (2008) and Buchanan et al. (2011), in part because of a recent update of the Chessie BIBI database. A version of the index was tailored to three spatial scales and three taxonomic levels using the stream macroinvertebrate metrics determined as most sensitive to disturbance at that scale and taxonomic level. Metric scores are average to obtain an overall index score for each spatial-taxonomic version of the index.

A. **Spatial Scales**

A single index that applies to all streams and wadeable rivers in the Chesapeake Bay basin would allow for an uncomplicated assessment of stream health. Although superficially promising, the single index does not account for regional differences in the Chesapeake Bay watershed's natural features or for the wide-spread landscape manipulation and degradation in the coastal region. Natural differences in hydrologic and topographic features of the coastal and inland regions correspond to strong natural differences in stream macroinvertebrate communities (Appendix G and H). The coastal region also has a paucity of high quality sites, a fact that is well recognized (e.g., Maxted et al. 2000). The few Reference sampling events in the coastal region were overshadowed by the more numerous Reference sampling events in the inland region, effectively turning coastal sites into outliers during index development of the Chesapeake-wide index. The Chesapeake-wide index does not fairly represent biological responses to stream degradation in the coastal region, and is not recommended.

Creating separate indices for the Coast and Inland regions produces more reliable assessments of stream biological condition in the Chesapeake watershed. The Coast region is defined by the Middle Atlantic Coastal Plain (MAC) and Southeastern Plain (SEP) Level III ecoregions and has mainly low elevation, low gradient, hydrologically connected streams with unconsolidated substrates (Woods et al. 1999). The Inland region is represented by a wide range of stream gradients, elevations, and substrates. It makes up a large portion of the Chesapeake Bay watershed, extending from southern Virginia to southern New York. All three taxonomic versions of the Inland region indices had CEs greater than or equal to 70.0% (Table K-2). The family- and genus-level Coast region indices were greater than or equal to 70.0% and the order-level CE fell just below this threshold (CE = 69.5%). Area-weighted ratings of the Coast and Inland region indices can be used together to represent stream biological integrity across the entire Chesapeake watershed (e.g., family-level Coast index and family-level Inland index). Sensitivity of the Coast index is currently impeded by a paucity of Reference sites in the MAC bioregion. The strong agreement between the family- and genus-level versions of the Inland index suggest they could be used interchangeably in Inland areas (Table 20B).

Further dividing the Chesapeake Bay basin into twelve bioregions provides more spatial resolution of the basin's complex natural features and accommodates some natural differences in least-disturbed stream macroinvertebrate assemblages (Appendix G). The greater spatial resolution helps to avoid situations experienced in Buchanan et al. (2011) where high numbers of Reference and Degraded samples in one area dominated index development in an entire bioregion. A good example is the original 2011 Piedmont bioregion, where most of the Reference and Degraded samples available at the time were from what is now the Lower

Northern Piedmont (LNP) bioregion (Buchanan et al. 2011). The LNP area contains the heavily sampled Washington D.C. suburbs, and development of the 2011 Piedmont index was dominated by that data. Similarly, indices for the 2011 Ridge and Valley bioregions were heavily weighted by Reference samples collected in their less disturbed, southern areas. The 2011 Ridge and Valley bioregions are both complex and large, bisecting the entire Chesapeake Bay basin on a northeast-southwest axis. In this study, these two bioregions were divided into six smaller, more homogeneous bioregions based on the Bray-Curtis Dissimilarity Index results of a stream classification analysis (Appendix G): Northern Ridge and Valley (NRV), Southern Ridge and Valley (SRV), Blue Ridge (BLUE), Central Appalachians (CA), the Southern Great Valley (SGV), and a portion of the Upper-Northern Piedmont (UNP). One bioregion that may benefit from subdivision in the future, when more data are available, is the Northern Appalachian Plateau and Uplands (NAPU). Reference sites in this bioregion are concentrated on the eastern side (Level IV Ecoregion 60b) and the bioregion's family-level index has a low CE, which suggests divergent landscapes are reducing the index's CE.

B. **Taxonomic Versions**

Bioregion and region indices derived from order-level metrics were developed in an attempt to create a coarse field method for rapidly screening stream condition. Functional feeding group (FFG), habit, and some tolerance metrics are inappropriate at the order-level, which leaves richness/diversity, composition, and a subset of tolerance metrics to be included in order-level versions of the indices. The Inland order-level index was considered sensitive to stream conditions (CE $\geq$ 70.0%) and the Coast order-level index CE was just below the sensitivity threshold (CE = 69.5%) (Table K-2). Nine of the twelve order-level bioregion indices had CE values greater than or equal to 70.0% (i.e., BLUE, CA, LNP, MAC, NRV, SEP, SGV, SRV, and UNP) but most had higher amounts of uncertainty as measured by RMSE relative to the Region order-level indices (Appendix K). Index development sometimes identified fewer than five metrics for inclusion in the order-level bioregion indices. For example, only four metrics met the criteria for inclusion in the order-level BLUE index (Table 13). This may reflect the low number of available degraded sampling events in BLUE. We recommend caution in using the order-level versions of both the region and bioregion indices. However, we believe with further work they can eventually be a resource for rapid *in situ* assessments and screening purposes—most likely conducted by non-experts. Volunteers and other non-experts can accurately identify benthic macroinvertebrates to the order-level with minimal training and equipment.

With few exceptions, the bioregion and region indices derived from family- and genus-level metrics had higher CEs than their order-level equivalents. Family- and genus-level taxonomic identification requires more extensive training and equipment to observe minute morphological features, such as gills, mouth parts, and leg segments. The genus-level versions did not increase CEs substantially over the family-level versions in most bioregions, and were lower in a few instances. The literature provides conflicting results on the benefits of identifying taxa to genus-level. Some studies have found that the genus-level provides additional information that allows for more accurate stream assessments (Lenat and Resh 2001, Pond et al. 2011) while other studies found that the genus-level provides minor improvements in the sensitivity compared to the family-level (Waite et al. 2004, Melo 2005, Corbi and Trivinho-Strixino 2006, Mueller et al. 2013). Genus-level identification produces a greater amount of information, allowing attributes to be more accurately assigned to the taxa and a better

representation of assemblage richness and diversity. However, increased resolution may also increase the amount of noise associated with the assemblages. A larger number of rare taxa are likely to appear at the genus-level resolution and their presence may reduce the ability to identify robust sensitive metrics. Van Sickle et al. (2007) found that O/E models improved with the exclusion of rare taxa and they suggest that a similar improvement will likely be observed with other biological assessment methodologies. In this study, variability in metric selection introduced by probabilistic rarefaction is directly tied to the presence of rare taxa (Appendix D). Additionally, seasonal differences are more apparent at the genus-level than the family-level. Season was not accounted for in the Chessie BIBI in order to maintain sufficiently large Reference sample sizes. Pond et al. (2011) found that the West Virginia Genus-Level Index of Most Probable Stream Status (GLIMPSS) performed well when separate indices were developed for samples collected in the spring and summer. Separate seasonal indices may improve the performance of the genus-level Chessie BIBI indices in future analyses.

C. **Distributions of Index Scores**

Distributions of an individual metric's values in Reference conditions are not identical across the twelve bioregions due in large part to the natural variation in the Chesapeake basin's stream macroinvertebrate assemblages (e.g., Appendix G and I). The largest differences occur between the Coast and Inland regions. For example, percentages of EPT in Reference conditions are typically greater than 50% in the ten inland bioregions and less than 40% in the two coastal bioregions. Differences are also evident between bioregions located within the same region. For the bioregions located within the Inland region, percentages of burrowers decrease as watershed mean slopes become steeper. Although metrics are scored per their bioregion's Reference and Degraded assemblages, bioregion differences carry through to the index scores. Consequently, the numeric score for one bioregion index is not necessarily equivalent to the same numeric score for another bioregion index (Pond et al. 2011).

Distributions of the bioregion index scores in Reference and Degraded conditions also are not the same across the twelve bioregions. This is true at the order-level (Figure 10 and 11), family-level (Figure 12 and 13), and genus-level (Figure 14 and 15), and occurs regardless of the methodology used to construct the index (Figure I-3). Likewise, for the two region indices, Coast and Inland, Reference and Degraded distributions of index scores are not identical (Figures 7, 8, and 9). Although the water quality and stream habitat criteria for Reference or Degraded conditions were applied consistently to all sites across the Chesapeake Bay watershed, other environmental factors influence stream biota (Appendix H) and can cause distributions of index scores in Reference conditions to shift. The standardized water quality and stream habitat criteria do not account for all the local factors influencing macroinvertebrate assemblages. Thus, the quality of Reference sites in one bioregion may in fact be higher or lower than that in another bioregion.

We recommend a certain amount of caution when using the BLUE, CA, and MAC bioregion indices. These bioregions had low numbers of either Reference or Degraded samples and consequently had high RMSE (Appendix K). Validation estimates indicate the family-level CA index overestimated CE (i.e., delete-d jackknife cross validation procedure). More samples are required in the BLUE, CA, and MAC bioregions to further refine these indices and verify their accuracy.

D. **Narrative Ratings**

The 50th, 25th, 10th, and half of the 10th percentiles of the Reference distributions, excluding outliers, in each spatial and taxonomic version of the index were used in this study to establish five ratings: Excellent, Good, Fair, Poor, and Very Poor (Table 5). This common narrative rating system reduces the unevenness of the distributions of Reference index score and allows narrative ratings to be compared across bioregions and regions. An Excellent rating derived from the family-level index in one bioregion should be roughly equivalent to an Excellent rating derived from the family-level index in another bioregion because both ratings represent index scores greater than the 50th percentile of their respective Reference distributions.

Ratings for this study's twelve-bioregion index and the Buchanan et al. (2011) six-bioregion index, which used some of the same percentiles to rate family-level index scores, were comparable with 84.6% (match = 47.2%, near = 37.4%) agreement in the Inland bioregions and 85.8% (match = 50.2%, near = 35.6%) agreement in the Coast bioregions (Table 18). It's worth recalling that Buchanan et al. (2011) used somewhat different Reference criteria and metric scoring approaches for the inland bioregions, and employed the Coastal Plain Macroinvertebrate Index created by Maxted et al. (2000) for the coastal region.

The natural features collectively represented in bioregion do not appear to substantially change the narrative ratings of the two family-level region indices relative to the twelve family-level bioregion indices (Appendix M). Most bioregions showed less than ± 10% difference between their region and bioregion index ratings in the frequency of desirable Excellent, Good, and Fair ratings (%EGF). Rating differences are also less than ± 10% in most of bioregions when scores and ratings of the region indices are compared by season and karst. These results corroborate direct comparisons of individual ratings which show less than 10% disagreement between the bioregion and region family-level index ratings (Table 20A). Overall, the bioregion, season, and karst differences are not large enough to dissuade use of the family-level region indices.

Discrepancies in three bioregions raise the question of which spatial resolution—region or bioregion index—provides the most accurate depiction of stream condition. The CA bioregion, which is split into three non-contiguous areas inside the Chesapeake Bay watershed (Figure 3), has a limited number of sampling events that meet the Reference criteria (n = 35). The MAC bioregion inside the Chesapeake watershed has even fewer Reference sampling events (n = 17). The BLUE bioregion has adequate numbers of Reference samples for index development (n = 133) but only seven samples representing Degraded conditions. Until CA and MAC have more Reference samples and BLUE has more Degraded samples, the family-level Inland and Coast indices produce more realistic evaluations of stream health in these three bioregions. To develop more robust indices, the CA and MAC bioregions may benefit from the incorporation of samples outside of the Chesapeake Bay watershed but still within their respective bioregion/ecoregion.

In several bioregions, bioregion index scores and %EGFs in the Reference conditions differ substantially from their region index counterparts (Appendix M). Data preparation and standardization steps minimize or eliminate most field and laboratory variables that may cause conflicting ratings. A consistent, methodical approach was used to develop each version of the index, and the bioregions in question have sufficient numbers of Reference samples. The differences may be caused by less stringent or different methods used by some monitoring

programs to evaluate stream habitat parameters (Appendix M) which result in higher habitat scores. These samples would technically meet the eight habitat parameter criteria for Reference (Table 3) and be included in the Reference pool. Reference distributions of the index scores determine the rating thresholds, and %EGF in Reference conditions will be dragged down if a significant portion of the Reference data pool is of a lower quality relative to other bioregions. At the regional scale, the influence of these samples is minimized by larger numbers of Reference samples collected by multiple programs. We believe the indices of at least NAPU, UNP and LNP—and possibly other bioregions—are affected by this issue.

When sites are being assessed individually with internally consistent methods (e.g., to measure restoration "lift"), the genus-level versions of bioregion indices may be useful if they are substantially more sensitive than their corresponding family-level versions (e.g., NRV in Table K-3). The inherent effect of seasonality on the genus-level metrics, however, may make the family-level indices more reliable to use when merging data from different monitoring programs. Good comparability in the ratings is found between the family-and genus-level versions of the Inland region and bioregion indices (Table 20B and C). Ratings of the family- and genus-level versions agree 90.3% in the Inland region index (Table 20B) and 91.4% in Inland bioregion indices (Table 20C). We view cautiously the apparent good agreement in ratings of the order- and family-level coastal bioregion index (Table 20C) given the poor agreement in other coastal index comparisons.

V. Conclusions and Recommendations

A. Sample Collection

A major difficulty in refining the Chessie BIBI was the discrepancies in the habitat and water quality variables rather than differences in macroinvertebrate collection and identification methods. Environmental condition classification is a major aspect of IBI development. The water quality and habitat factors used to classify samples as Reference and Degraded influence which metrics are selected in the final index. For example, site classification based strictly on nutrient criteria and site classification based on habitat variables are likely to produce two different IBIs even though the same data were utilized. None of the habitat or water quality variables in our database were measured at all sampling events, and a subset of frequently measured variables was used in the development of the indices. Agencies/programs can argue that their protocols are tailored to address specific needs, so their data do not need to be compatible with other data. However, standardizing procedures makes it easier for programs to collaborate, review, verify previous conclusions, and rapidly improve the science. Collecting at a minimum the habitat parameters outlined in the visual-based Rapid Bioassessment protocol (Barbour et al. 1999), in a consistent fashion, and specific conductivity, pH, DO, and temperature—measurements easily collected with the average sonde—would be very beneficial. The collection of ancillary parameters, such as total nitrogen, would further improve our understanding of aquatic ecosystems but these parameters should be collected in addition to the standard parameters, not in place of the standard parameters. A standard set of procedures would also make it easier to perform cooperative assessments.

The CA, MAC, BLUE, and PIED bioregions lack adequate Reference and/or Degraded sample sizes for robust index development. Other bioregions, such as NAPU, showed unexpectedly low CEs. Targeting specific streams expected to be of Reference or Degraded

quality within each of these bioregions would most likely improve index performance in future refinements. A review of land use and USGS gauges or other water monitoring data could help identify streams that meet these condition categories. Additionally, it may be beneficial to include data collected outside of the Chesapeake Bay watershed. Many of the bioregions or ecoregion level III's within the basin extend beyond the limits of the basin. Including samples outside of the basin increase sample sizes and would especially benefit poorly represented bioregions. For example, the MAC bioregion represents the northern extent of Level III Ecoregion 63 (Woods et al. 1999). Expanding the MAC area to include portions of North Carolina and South Carolina represented by ecoregion 63 would increase sample size and may ultimately improve index performance. The samples outside of the basin would only be used during index development. Chesapeake Bay watershed assessments would remain limited to the samples located within the basin.

Spatial distribution of Reference and Degraded sample locations could be factored into a program's sampling design to further reduce the potential of spatial bias. Although this effort would be conducted at the bioregion level, the performance of region indices would also improve, or at the very least, confidence in the current indices would improve. Furthermore, there are many HUC12s within the basin that are underrepresented (Appendix L). Collecting samples from these locations would benefit basin-wide assessments but should be considered a secondary goal behind the primary effort to target Reference and Degraded sites.

B. **Data Analysis**

Quantifying land use in the watershed above a sampling station may improve the classification of samples along the disturbance gradient. There are connections between stream biota and watershed land use through stream habitat and water quality conditions. The HUC12 land use variables in Appendix H provide an approximation of the watershed conditions above the sampling station. We recommend watersheds be delineated from the sampling station to aid in environmental condition classification. Delineating watersheds from the sampling location in ArcGIS will improve land use estimates in the catchments above sample locations, but the process will require extensive QA/QC. The most difficult aspect will be accurately placing the sampling station on the appropriate stream. Low accuracy in GPS equipment and inaccuracies associated with GIS stream layer location caused a significant portion of the stations in the Chessie BIBI database to fall outside the stream lines in ArcGIS. ArcGIS tools can be used to snap the points to the stream lines but many streams in a small geographic area may result in the sampling station snapping to the wrong stream. Therefore, effort would be required to validate that watersheds have been accurately delineated for all the sampling stations in the database.

An important issue that became apparent during the refinement of the Chessie BIBI was the absence of tolerance value, functional feeding group (FFG), and habit assignments for a subset of taxa. Tolerance, FFG, and Habit metrics are considered essential for diversifying and creating a robust IBI. Error is introduced when calculating metrics from these three categories if a taxon or taxa are not assigned the appropriate numerical or categorical variable. Assigning new tolerance values, FFGs, and habits requires extensive investigation and was beyond the scope of this study. Collaboration among multiple agencies/programs could be helpful in the assignment of future traits. Often taxa are not assigned traits because they occur infrequently but a large cohesive dataset may provide enough data for an accurate assignment of taxonomic attributes to these rare taxa. An effort was made to summarize taxonomic traits from multiple

sources. We concluded that aggregating assignments from multiple categories provided the best representation of the taxa of interest. We recommend that efforts be continued to update the current set of attributes/traits and to include additional attributes/traits from additional sources.

The classification system used to identify Reference, Degraded, and intermediate site conditions (Table 2 and Table 3) forms a step-wise, or non-continuous, stressor axis based on commonly measured physical and chemical parameters. Distinctly different biological metric values, metric scores, and Chessie BIBI index scores are found at each step along the axis. This stressor-response relationship is reminiscent of the Biological Condition Gradient (BCG), a conceptual framework relating six tiers of biological responses to a gradient of increasing stress on aquatic ecosystems (USEPA 2016c). The BCG is intended to more precisely define and measure biological status, better recognize the quality of reference sites, document the effectiveness of restoration efforts, identify anthropogenic stressors, and help establish biocriteria in water quality standards. The BCG stressor axis represents the cumulative effects of all physical, chemical, and biological factors adversely affecting aquatic biota whereas this study's stressor axis only reflects eight physical and three chemical factors. Rough approximations can still be drawn between the two. Biota in this report's Reference conditions equate approximately to BCG Level 2, which has biological "structure and function similar to natural community with some additional taxa and biomass, and ecosystem level functions are fully maintained" (USEPA 2016c). Likewise, populations in Degraded conditions are roughly equivalent to BCG Level 5, where sensitive taxa are markedly diminished, distributions of the major taxonomic groups are conspicuously unbalanced, and ecosystem functions show reduced complexity and redundancy. Results of this study could be used to quantify some of the biological attributes needed to construct regional BCGs. For example, attributes II (highly sensitive taxa), III (intermediate sensitive taxa), IV (intermediate tolerant taxa), and V (tolerant taxa). Study results could also support BCGs already developed for parts of the Chesapeake watershed (e.g., Stamp et al. 2014).

C. **Assessments**

To assess stream health in the Chesapeake Bay basin as a whole, the selected index should consider CE, precision, accuracy, and parsimony. Unnecessary complexity increases the potential of introducing error in future assessments. When multiple taxonomic and spatial versions of the index show roughly the same sensitivity, it is beneficial to select the simplest index or set of indices. For region and bioregion indices, CEs of the family-level versions were generally comparable to the genus-level version (Table K-2 and Table K-3), therefore, the family-level assessments were selected as the parsimonious taxonomic level. The Chesapeake-wide index was the most simplistic index developed but results suggested that underlying environmental factors, independent of stream condition, were confounding results. CEs of the region indices were generally comparable to the bioregion indices, and thus, the region indices were the most parsimonious spatial resolution. The family-level region indices are recommended for Chesapeake watershed-wide assessments. They standardize metrics and scoring thresholds over large areas of the basin, allowing for direct comparison of index scores among most sampling events. They can provide a robust indicator of stream health at either the region and bioregion scales.

We recommend using the region and bioregion indices in concert to assess local (e.g., restoration) sites in the Chesapeake watershed. Both spatial resolutions contain varying amounts of error. Scores and ratings provided by the regional indices can be directly compared across

large portions of the basin; for example, scores in BLUE can be compared directly to scores in UNP. Scores and ratings from the bioregion indices may be more sensitive to localized nuances *and* more affected by differences in monitoring program methodology. Since the region and bioregion indices were developed separately, they can be treated as independent measures. Confidence in the results improves when both spatial resolutions are in agreement (Table 18). If both spatial resolutions of the index classify a sampling station as "Very Poor," then the sampling station may be a prime candidate for restoration activities. If both spatial resolutions classify the station as "Excellent," then the station may benefit from conservation actions. Disagreement between the ratings will require the sampling stations to be review on a case by case basis. In general, the most conservative action would be to assume that lower of the two ratings is the most accurate. The area will then need to be reviewed further before making a final determination. A more in-depth comparison of the two spatial resolutions may be beneficial but could be erroneous. For example, the region index may classify a station as "Poor" and the bioregion index may classify the station as "Good." This result could be interpreted as the sample is considered to be "Poor" relative to the majority of the Chesapeake Bay watershed but relative to the samples within the bioregion the station tends to have a higher score. Therefore, this station may benefit from some restoration activity but relative to the other samples in the bioregion it should not be considered a top priority. This approach should be used with caution because the difference could be due to a weakness or error in the index and may not reflect differences in spatial resolution.

The Chesapeake Bay Program (CBP) is seeking a measure of stream health to monitor restoration progress within the Chesapeake basin. Because the sampling stations are not evenly distributed throughout the basin, we recommend that the samples be area-weighted. Area-weighting the index scores reduces the spatial bias associated with heavily sampled geographic areas, enabling an accurate summary of the basin to be derived. In the past, CBP has focused on the basin-wide percentage of streams categorized as Excellent, Good, or Fair. Using the area-weighting method, we estimate approximately 49.2% (region indices) and 49.2% (bioregion indices) of streams assessed in the Chesapeake basin were in Excellent, Good, or Fair condition over the 1992 – 2015 time period. Approximately 11.3% (region indices) and 13.6% (bioregion indices) have insufficient samples and could not be evaluated. An objective of the CBP is to develop a 2008 baseline against which CBP can measure progress in restoring stream health. This study provides the framework and the necessary data for the Stream Health Workgroup to construct the 2008 baseline.

The Chessie BIBI is an ever-evolving index. Admittedly, the biological data in the Chessie BIBI database are prone to bias due to differences in sampling technique, enumeration technique, and taxonomic resolution among agencies/programs. Several steps were taken to extensively groom the database and reduce the influence of any existing, known bias. We believe the benefits of combining multiple data sets outweighs the inherent biases. Sample size, and therefore, statistical power increases. The data set can transcend geopolitical borders, allowing for contiguous analysis within geographic areas deemed environmentally similar (i.e., regions and bioregions). Benthic macroinvertebrate assessments are often limited by time and funding. Many agencies/programs struggle to find an adequate sample size to develop or update an IBI in their region and they are constrained by political borders not observed by the fauna. We strongly recommend that IBIs be developed in cooperative manner between agencies/programs. Larger, cohesive data sets will improve statistical power, and the effort will be divided among multiple partners. Collaboration will also provide a succinct set of results that

will be more readily interpretable by non-experts; as opposed to differing index values and ratings for the same general area reported by multiple agencies/programs (Maxted et al. 2000).

The Chesapeake Bay Program (CBP) is seeking a 2008 baseline of Chessie BIBI scores and ratings for monitoring restoration progress in the basin. Establishing the baseline requires attention and direction from CBP. The baseline will have to be derived from multiple years of data to overcome spatial and temporal gaps. Data collected between 2000-2011 are the most prospective candidates for establishing the baseline because most of the Chessie BIBI data currently in the database were collected during that time period. Stations that are periodically sampled provide the best data source for determining trends. Without a subset of repeatedly sampled stations, it is difficult—but not impossible—to determine if observed trends are a response to restoration efforts and not the result of temporal or spatial variability inherent in the random sampling design applied by many agencies/programs. The CBP objective to document trends in stream health would benefit from agencies/programs periodically returning to existing stations. To measure trends, we recommend long-term monitoring programs should collect benthic macroinvertebrate samples at a predefined frequency (e.g., every 5 years) from stations that represent the range of stream conditions.

VI. Citations

ASTIN, L. E. 2006. Data synthesis and bioindicator development for nontidal streams in the interstate Potomac River basin, USA. Ecological Indicators 6:664–685.

ASTIN, L. E. 2007. Developing biological indicators from diverse data: The Potomac Basin-wide Index of Benthic Integrity (B-IBI). Ecological Indicators 7:895–908.

BARBOUR, M. T., J. GERRITSEN, G. E. GRIFFITH, R. FRYDENBORG, E. MCCARRON, J. S. WHITE, AND M. L. BASTIAN. 1996. A framework for biological criteria for Florida streams using benthic macroinvertebrates. Journal of the North American Benthological Society:185–211.

BARBOUR, M. T., J. GERRITSEN, B. D. SNYDER, AND J. B. STRIBLING. 1999. Rapid bioassessment protocols for use in wadeable streams and rivers. Periphyton, Benthic Macroinvertebrates, and Fish (2nd edn). US Environmental Protection Agency, Office of Water, Washington, DC EPA.

BILTON, D. T., J. R. FREELAND, AND B. OKAMURA. 2001. Dispersal in freshwater invertebrates. Annual review of ecology and systematics:159–181.

BLOCKSOM, K. A. 2003. A performance comparison of metric scoring methods for a multimetric index for Mid-Atlantic Highlands streams. Environmental Management 31:0670–0682.

BLOCKSOM, K. A., AND B. R. JOHNSON. 2009. Development of a regional macroinvertebrate index for large river bioassessment. Ecological indicators 9:313–328.

BOLLMAN, W., J. BOWMAN, AND D. WINTER. 2010. Analysis of Biological Samples: District of Columbia Phytoplankton, Zooplankton, and Benthic Macroinvertebrate Samples: 2005-2009. Rhithron Associates, Inc., Missoula, Montana.

BUCHANAN, C., K. FOREMAN, J. JOHNSON, AND A. GRIGGS. 2011. Development of a Basin-wide Benthic Index of Biotic Integrity for Non-tidal Streams and Wadeable Rivers in the Chesapeake Bay Watershed: Final Report to the Chesapeake Bay Program. Interstate Commission on the Potomac River, ICPRB Report:11–1.

BUNGE, J., AND M. FITZPATRICK. 1993. Estimating the number of species: a review. Journal of the American Statistical Association 88:364–373.

BUTCHER, J. T., P. M. STEWART, AND T. P. SIMON. 2003. A benthic community index for streams in the northern lakes and forests ecoregion. Ecological indicators 3:181–193.

CARTER, J. L., AND V. H. RESH. 2013. Analytical approaches used in stream benthic macroinvertebrate biomonitoring programs of State agencies in the United States. US Geological Survey.

CHALFANT, B. 2009. A benthic index of biotic integrity for wadeable freestone streams in Pennsylvania. Pennsylvania Department of Environmental Protection, Division of Water Quality Standards. URL: http://www. depweb. state. pa. us/watersupply/lib/watersupply/ibi_rifflerun. pdf, accessed 20.

CHESAPEAKE BAY PROGRAM. 2015. Stream Health Outcom Management Strategy.

CORBI, J. J., AND S. TRIVINHO-STRIXINO. 2006. Influence of taxonomic resolution of stream macroinvertebrate communities on the evaluation of different land uses. Acta Limnologica Brasiliensia 18:469–475.

DAIL, M. R., J. R. HILL, AND R. D. MILLER. 2013. The Virginia Coastal Plain Macroinvertebrate Index. Virgina Department of Environmental Quality, Roanoke, VA 24019.

FAUSCH, K. D., J. R. KARR, AND P. R. YANT. 1984. Regional Application of an Index of Biotic Integrity Based on Stream Fish Communities. Transactions of the American Fisheries Society 113:39–55.

FEMINELLA, J. W. 2000. Correspondence between stream macroinvertebrate assemblages and 4 ecoregions of the southeastern USA. Journal of the North American Benthological Society 19:442–461.
FENNEMAN, N. M. 1917. Physiographic subdivision of the United States. Proceedings of the National Academy of Sciences 3:17–22.
FOREMAN, K., C. BUCHANAN, AND A. NAGEL. 2008. Development of ecosystem health indexes for non-tidal wadeable streams and rivers in the Chesapeake Bay basin. Report to the Chesapeake Bay Program Non-Tidal Water Quality Workgroup 12:08.
FRIBERG, N., L. SANDIN, M. T. FURSE, S. E. LARSEN, R. T. CLARKE, AND P. HAASE. 2006. Comparison of macroinvertebrate sampling methods in Europe. Pages 365–378 The Ecological Status of European Rivers: Evaluation and Intercalibration of Assessment Methods. Springer.
GADNR. 2007. Macroinverterbate Biological Assessment of Wadeable Streams in Georgia. Standard Operating Procedure, Georgia Department of Natrual Resources/ Environmental Protection Division/ Watershed Protection Branch.
GERRITSEN, J., J. BURTON, AND M. T. BARBOUR. 2000. A stream condition index for West Virginia wadeable streams. US EPA Region 3.
GERTH, W. J., AND A. T. HERLIHY. 2006. Effect of sampling different habitat types in regional macroinvertebrate bioassessment surveys. Journal of the North American Benthological Society 25:501–512.
GIBSON, G. R., M. T. BARBOUR, J. B. STRIBLING, J. GERRITSEN, AND J. R. KARR. 1996. Biological Criteria: Technical guidance for streams and small rivers. Environmental Protection Agency, Washington, DC (United States). Office of Water.
GOTELLI, N. J., AND R. K. COLWELL. 2011. Estimating species richness. Biological diversity: frontiers in measurement and assessment 12:39–54.
HAWKINS, C. P., R. H. NORRIS, J. GERRITSEN, R. M. HUGHES, S. K. JACKSON, R. K. JOHNSON, AND R. J. STEVENSON. 2000. Evaluation of the use of landscape classifications for the prediction of freshwater biota: synthesis and recommendations. Journal of the North American Benthological Society 19:541–556.
HAWKINS, D. M. 2004. The problem of overfitting. Journal of chemical information and computer sciences 44:1–12.
HERBST, D. B., AND E. L. SILLDORFF. 2006. Comparison of the performance of different bioassessment methods: similar evaluations of biotic integrity from separate programs and procedures. Journal of the North American Benthological Society 25:513–530.
HUGHES, R. M., P. R. KAUFMANN, A. T. HERLIHY, T. M. KINCAID, L. REYNOLDS, AND D. P. LARSEN. 1998. A process for developing and evaluating indices of fish assemblage integrity. Canadian Journal of Fisheries and Aquatic Sciences 55:1618–1631.
JOHNSON, J. M. 2013. Non-tidal benthic monitoring database. Chesapeake Bay Program.
KARR, J. R. 1981. Assessment of biotic integrity using fish communities. Fisheries 6:21–27.
KARR, J. R. 1991. Biological integrity: A long-neglected aspect of water resource management. Ecological applications 1:66–84.
KENNEN, J. G. 1999. Relation of macroinverterbate community impairment to catchment charactersitics in New Jersey Streams. Wiley Online Library.
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 OTHERS. 2003. Development and evaluation of a macroinvertebrate biotic integrity index (MBII) for

regionally assessing Mid-Atlantic Highlands streams. Environmental Management 31:0656–0669.

LENAT, D. R., AND V. H. RESH. 2001. Taxonomy and stream ecology—the benefits of genus-and species-level identifications. Journal of the North American Benthological Society 20:287–298.

MAXTED, J. R., M. T. BARBOUR, J. GERRITSEN, V. PORETTI, N. PRIMROSE, A. SILVIA, D. PENROSE, AND R. RENFROW. 2000. Assessment framework for mid-Atlantic coastal plain streams using benthic macroinvertebrates. Journal of the North American Benthological Society 19:128–144.

MCCUNE, B., J. B. GRACE, AND D. L. URBAN. 2002. Analysis of ecological communities. MjM software design Gleneden Beach, OR.

MELO, A. S. 2005. Effects of taxonomic and numeric resolution on the ability to detect ecological patterns at a local scale using stream macroinvertebrates. Archiv für Hydrobiologie 164:309–323.

MINNS, C. K., V. W. CAIRNS, R. G. RANDALL, AND J. E. MOORE. 1994. An index of biotic integrity (IBI) for fish assemblages in the littoral zone of Great Lakes' areas of concern. Canadian Journal of Fisheries and Aquatic Sciences 51:1804–1822.

MUELLER, M., J. PANDER, AND J. GEIST. 2013. Taxonomic sufficiency in freshwater ecosystems: effects of taxonomic resolution, functional traits, and data transformation.

NAGEL, A. 2016. 2015/2016 Update of the Watershed Wide Benthic Invertebrate Database. Interstate Commission on the Potomac River Basin, 16-6.

OKSANEN, J., F. G. BLANCHET, R. KINDT, P. LEGENDRE, P. R. MINCHIN, R. B. O'HARA, G. L. SIMPSON, P. SOLYMOS, H. H. STEVENS, AND H. WAGNER. 2016. vegan: Community Ecology Package.

OMERNIK, J. M. 1987. Ecoregions of the conterminous United States. Annals of the Association of American geographers 77:118–125.

OSTERMILLER, J. D., AND C. P. HAWKINS. 2004. Effects of sampling error on bioassessments of stream ecosystems: application to RIVPACS-type models. Journal of the North American Benthological Society 23:363–382.

PETERSEN, I., Z. MASTERS, A. G. HILDREW, AND S. J. ORMEROD. 2004. Dispersal of adult aquatic insects in catchments of differing land use. Journal of Applied Ecology 41:934–950.

POND, G. J., J. E. BAILEY, B. LOWMAN, AND M. J. WHITMAN. 2011. West Virginia GLIMPSS (genus-level index of most probable stream status): a benthic macroinvertebrate index of biotic integrity for West Virginia's wadeable streams. West Virginia Department of Environmental Protection, Division of Water and Waste Management, Watershed Branch, Charleston, WV, USA.

R CORE TEAM. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

REHN, A. C., P. R. ODE, AND C. P. HAWKINS. 2007. Comparisons of targeted-riffle and reach-wide benthic macroinvertebrate samples: implications for data sharing in stream-condition assessments. Journal of the North American Benthological Society 26:332–348.

Retrieved [02/24/2017], from USEPA Recovery Potential Screening Tools, https://www.epa.gov/rps/recovery-potential-screening-tools-downloadable-tools-comparing-watersheds. (n.d.). .

Retrieved [06/01/2016], from the Integrated Taxonomic Information System On-Line Database, http://www.itis.gov. (n.d.). .

SEABER, P. R., F. P. KAPINOS, AND G. L. KNAPP. 1987. Hydrologic unit maps: US Geological Survey water supply paper 2294. US Geological Survey.

SHAO, J. 1989. The efficiency and consistency of approximations to the jackknife variance estimators. Journal of the American Statistical Association 84:114–119.

SHAO, J., AND C. J. WU. 1989. A general theory for jackknife variance estimation. The Annals of Statistics:1176–1197.

SMITH, A. J. 2016. Standard Operating Procedure: Biological Monitoring of Surface Waters in New York State. Standard Operating Procedure, New York State Departement of Environmental Conservation, Division of Water.

SOUTHERLAND, M. T., M. J. KLINE, D. M. BOWARD, G. M. ROGERS, R. P. MORGAN, P. F. KAZYAK, R. J. KLAUDA, AND S. A. STRANKO. 2005. New Biological Indicators to Better Assess the Condition of Maryland Streams. Versar Inc., University of Maryland Appalachian Laboratory, Maryland Department of Natural Resources, Annapolis, MD.

SOUTHERLAND, M., J. VØLSTAD, L. ERB, E. WEBER, AND G. ROGERS. 2006. Proof of concept for integrating bioassessment results from three state probabilistic monitoring programs. EPA/903/R-05/003. Office of Environmental Information, US Environmental Protection Agency, Fort Meade, Maryland.

STAMP, J., J. GERRITSEN, G. J. POND, S. K. JACKSON, AND K. VAN NESS. 2014. Calibration of the Biological Condtion Gradient (BCG) for Fish and Benthic Macroinvertebrate Assemblages in the Northern Piemont region of Maryland. Page 44. USEPA Office of Water and Montgomery County Department of Environmental Protection.

USEPA. 2006. Wadeable Streams Assessment: A Collaborative Survey of the Nation's Streams.

USEPA. 2008. NRSA 0809 Benthic Taxa List and Autecology - Data. Environmental Protection Agency.

USEPA. 2012. Freshwater Biological Traits Database. United States Environmental Protection Agency.

USEPA. 2016a. National Rivers and Streams Assessment 2008-2009: A Collaborative Survey. Office of Water and Office of Research and Development, Washington, DC.

USEPA. 2016b. Field-based methods for developing aquatic life criteria for specific conductivity.

USEPA. 2016c. A Practitioner's Guide to the Biological Condition Gradient: A Framework to Describe the Incremental Change in Aquatic Ecosystems. U.S. Environmental Protection Agency, Washington, DC.

VAN SICKLE, J., D. P. LARSEN, AND C. P. HAWKINS. 2007. Exclusion of rare taxa affects performance of the O/E index in bioassessments. Journal of the North American Benthological Society 26:319–331.

WAITE, I. R., A. T. HERLIHY, D. P. LARSEN, N. S. URQUHART, AND D. J. KLEMM. 2004. The effects of macroinvertebrate taxonomic resolution in large landscape bioassessments: an example from the Mid-Atlantic Highlands, USA. Freshwater Biology 49:474–489.

WOLF, J. 2008. Benthic macroinvertebrate impairments, freshwater streams and rivers health assessment. Chesapeake Bay Program.

WOODS, A. J., J. M. OMERNIK, AND D. D. BROWN. 1999. Level III and IV ecoregions of Delaware, Maryland, Pennsylvania, Virginia, and West Virginia. US Environmental

Protection Agency, National Health and Environmental Effects Research Laboratory, Corvallis, Oregon. Report with map supplement, Scale 1:1–000.
WVDEP. 2015. Watershed Assessment Branch Benthic Macroinvertebrate Taxa Autecology Data Table. West Virginia Department of Environmental Protection, Division of Water and Waste Management, Watershed Branch, Charleston, WV.