

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

Draft Report

January 4, 2016

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 16-6 DRAFT

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 several states, the signatories or Commissioners to the Interstate Commission on the River Basin. No official endorsements should be inferred.

Suggested citation for this report

Smith, Zachary M., Claire Buchanan, and Andrea Nagel. 2016. DRAFT. Refinement of the Basin-Wide Benthic Index of Biotic Integrity for Non-Tidal Streams and Wadeable Rivers in the Chesapeake Bay Watershed. ICPRB Report 16-6. 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. The index is derived from macroinvertebrate data collected by federal, state, and local stream monitoring programs in the Chesapeake Bay region. The index was 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.

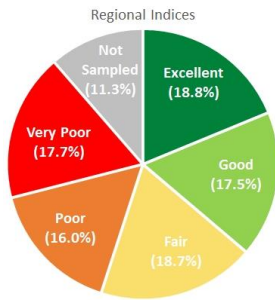
The analysis database contained 25,311 sampling events. Sampling sites in 1st to 4th order streams were classified into five categories based on habitat and water quality information: Reference (best quality), Minimally Degraded, Mixed (indeterminate quality), Moderately Degraded, and Degraded (poorest quality). Key attributes of the stream macroinvertebrates (taxonomic serial number, functional feeding group, habit, pollution tolerances) were reviewed and updated. Eighty-five metrics were calculated from the raw counts of March to November samples. Metrics were scored with a reference-based 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 causes of degradation and 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 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 in the raw counts. Order-level indices are less sensitive, but they do not require laboratory enumeration and are suited for rapid screening in the field. Family-level indices 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 for the purpose of comparing stream health across jurisdictional boundaries in the Chesapeake watershed. The 50th, 25th, and 10th percentiles of each version’s index scores in Reference stream 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.

A simple count of the narrative ratings indicates biological integrity is Very Poor or Poor at 46% of sampling sites and Fair, Good, or Excellent at 54% 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% sites and Fair, Good, or Excellent at 46% sites.



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

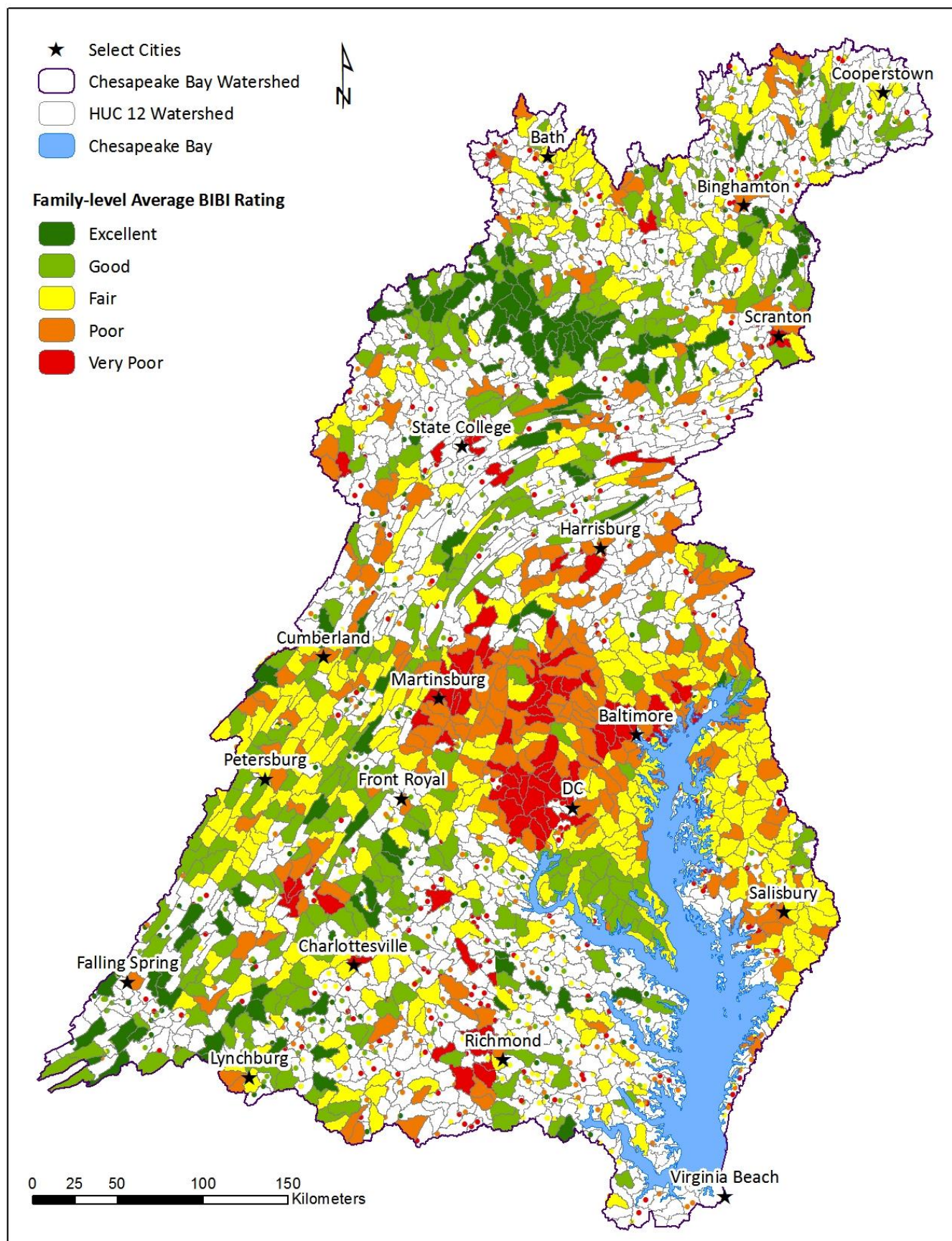
Straightforward counts such as these are misleading, however, because some areas—especially urban areas around Washington, D.C.—are more frequently sampled than others in the Chesapeake watershed. When index scores are weighted by the proportion of their local (HUC12) watershed area that they represent, the results indicate stream health is likely Very Poor or Poor in 34% of the Chesapeake watershed area; Fair, Good, or Excellent in 55% of the watershed; and not known in 11% of the watershed. Many unsampled streams are in predominantly agricultural or forested HUC12 watersheds and, when sampled, may improve percentages of the better ratings. Area-weighted ratings provide a better starting point than simple counts for measuring trends over time.

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 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 appeared 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 Rapid Bioassessment protocols (Barbour et al. 1999). We recommend stronger efforts to ensure that a standard suite of habitat and water quality measurements are collected 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 Regional Index) ratings for streams and small rivers in the Chesapeake Bay watershed. When sufficient data ($n = 5$) are available, HUC12 watersheds are colored according to the rating of their average index score. Otherwise, individual sampling locations are indicated and colored according to their ratings.

DRAFT REPORT



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 B-IBI 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 diligence in providing 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 B-IBI 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 2016 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	7
C. Environmental Data Preparation	8
i. Habitat	8
ii. Water Quality	8
D. Stream Condition Classification	9
E. Spatial Classification	10
F. Metric Testing	15
i. Metric Sensitivity	15
ii. Range and Variability	18
iii. Redundancy Analysis	18
G. Metric Scoring Approach	19
H. Index Construction	20
I. Index Classification Efficiency	22
J. Taxonomic Tiers	22
K. Delete-d Jackknife Validation	22
L. Delete-d Jackknife Precision	24
M. Narrative Rating Categories	24
N. Area-Weighting of Rating Results	26
III. Results	27
A. Index Construction	28
B. Chesapeake-Wide Index	29
C. Two Region Indices	30
i. Order-Level Indices	30

ii. Family-Level Indices.....	30
iii. Genus-Level Indices.....	30
D. Bioregion Indices	34
i. Order-Level Indices.....	34
ii. Family-Level Indices.....	34
iii. Genus-Level Indices.....	34
E. Index Validation and Precision.....	50
F. Narrative Ratings	50
G. Chesapeake Watershed Stream Health.....	54
IV. Discussion.....	56
A. Spatial scales	56
B. Taxonomic Versions.....	57
C. Distributions of Index Scores	59
D. Narrative Ratings.....	59
E. Chesapeake Bay Assessments	60
V. Conclusions and Recommendations	61
VI. Citations.....	64

Appendices

A. Taxonomic Classification (27 pgs)
B. Taxonomic Attributes (24 pgs)
C. Taxonomic Standardization (5 pgs)
D. Rarefaction (3 pgs)
E. Biological Metric Descriptions (4 pgs)
F. Abiotic Parameters for Evaluating Stream Environment (4 pgs)
G. Stream Classification (23 pgs)
H. HUC12 Watershed Characteristics in Bioregions (10 pgs)
I. Repeat Sampling Events at the Same Station (TBD)
J. Index Methodologies (17 pgs)
K. Index Performance, Accuracy, and Precision (12 pgs)
L. Index Rating (17 pgs)
Appendix Citations (2 pgs)

Refinement of the Basin-Wide Benthic 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, although 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. For example, 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 Geological Survey (USGS) and Environmental Protection Agency (USEPA) have used macroinvertebrates to evaluate US waterbodies on a national scale. As part of its 2004-2005 the Wadeable Stream Assessment (WSA), the USEPA collected 63 stream macroinvertebrate samples in the 165,760 km² Chesapeake Bay watershed (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, 2007, Friberg et al. 2006, Gerth and Herlihy 2006, Herbst and Silldorff 2006, Southerland et al. 2006, 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 km² 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 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 users. 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-eight federal, state, county, and non-profit agencies/programs that collect samples within the Chesapeake Bay basin (Table 1). A total of 25,311 samples collected with various methods between 1992 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-eight 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	607

DRAFT REPORT

	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	595
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_WSA	EPA - Wadeable Stream Assessment Program	7/20/2004	11/10/2004	63
24	USFS_SA	National Forest Service Stream Assessment	5/18/2000	5/8/2003	7

	AGENCY/PROGRAM CODE	AGENCY/PROGRAM NAME	START DATE	END DATE	TOTAL COUNT
25	USGS_NAWQA	National Water Quality Assessment Program	6/2/1993	8/16/2012	254
26	VADEQ_SA	Virginia DEQ Benthic Monitoring Program	5/20/1992	11/28/2014	4,650
27	VCU_INSTAR	Interactive SStream Assessment Resource	6/11/1999	11/3/2011	772
28	WVDEP_SA	West Virginia Dept. of Environmental Protection, Div. of Water and Waste Management	8/19/1996	10/1/2014	1,276
TOTAL					25,311

B. Taxonomic Data Preparation

i. Taxonomic Classification and Hierarchy

The taxonomic status of the benthic macroinvertebrates 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. 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. Municipal waste tolerance values, functional feeding groups (FFG), and habits were assigned from available sources (Barbour et al. 1999, NRSA 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; assigning new attributes is beyond the scope of this study. In other cases, taxa have multiple assigned attributes from multiple sources. If multiple sources provided municipal waste 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 of the variables had the same total count. Some of our attribute sources (NRSA 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. However, we argue that this can create odd distributions within an attribute category (i.e., FFG or Habit) and can create a bias toward taxon with multiple attributes. Taxa that have been assigned multiple attributes will incorrectly inflate the percentage of each metric class, resulting in a total percentage greater than 100% within a metric class (e.g., the sum of all percent FFG metrics). The sum of the resulting percentages within a metric class will be greater than 100% because some of the taxa have been represented more than once. 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.

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. Additionally, we limited our analyses to a Strahler stream order ≤ 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 from the analysis data set.

Undocumented differences in the laboratory procedures for enumerating stream macroinvertebrates can create bias. For example, some laboratories fail to explain their taxonomic rules beyond “the taxa were identified to the genus level or the lowest possible taxonomic resolution.” To reduce

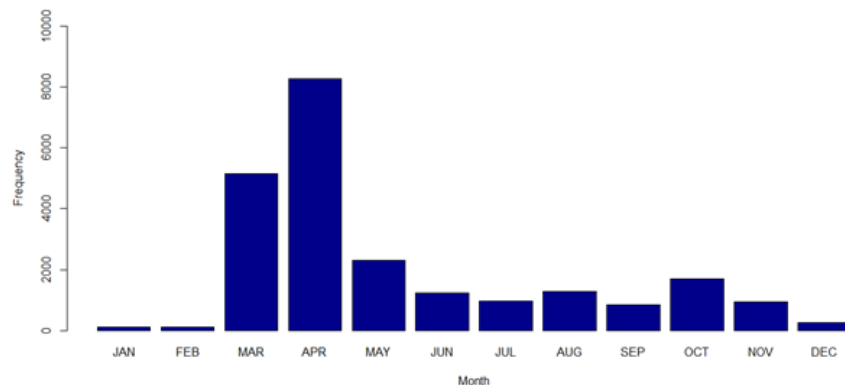


Figure 1. All of the kick-net samples in the Chessie BIBI database were aggregated together and the frequency of unique sample events were plotted for each month.

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. Often it was apparent that some agencies/programs identified a taxon while others excluded the taxon from their assessment. For example, MD_MBSS and NYSDEC did not include 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, and 3) the mean relative abundance of the taxon was low, 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 assemblages, 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, Gastropoda, Oligochaeta, and Trepaxonemata taxa were all rolled-up to the class-level, despite

some agencies/programs identifying these organisms to the species-level. Additionally, Entognatha, Lepidoptera, Neuroptera, and Neophora 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

The majority of 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).

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 of the

agencies/programs within 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 predicted, just as a rarefied richness can be predicted. 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 provides a repeatable estimate of taxonomic composition, whereas, rarefaction typically produces different results with each iteration. We used probabilistic rarefaction for calculations of richness and diversity measures in this study.

v. Biological metrics

Eighty-five biological metrics were identified in the literature (GADNR 2007, Pond et al. 2008, Carter and Resh 2013, Smith 2016). Additionally, the percentage of individuals in each Phylum, Subphylum, Class, Subclass, Order, Suborder, Family, Tribe, and Genus were

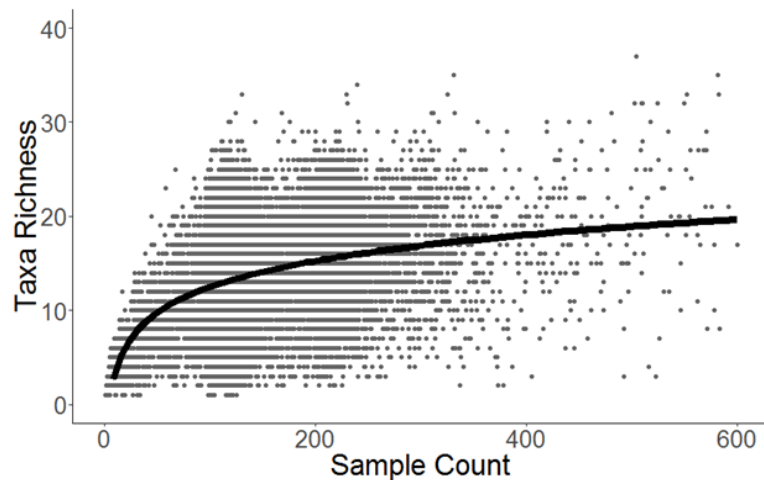


Figure 2. Sample with counts less than or equal to 600 ($n = 22,240$) were plotted against family richness. A base 10 logarithmic curve was generated with the available data.

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 could not be 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.

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, are given in Appendix E.

C. Environmental Data Preparation

Environmental data collected with the macroinvertebrate samples were used to create a standardized gradient of stream 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 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. As a result, only nine habitat parameters were measured consistently and frequently (i.e., more than 75% of sampling events); none of these parameters were collected at all sampling locations. One parameter, the velocity/depth ratio, was excluded because it was not scored on the same 0-20 standard scale used to score other habitat parameters. 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 to classify stream condition.

ii. Water Quality

Eighty-four 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 stream condition.

D. Stream 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. 2008) (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. 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 $\geq 75\%$ of the available habitat scores were ≥ 16 and none were < 12 , and the sum of the three water quality scores was zero. Degraded sites had $\geq 50\%$ of all available habitat scores ≤ 6 and the sum of the three water quality scores > 1 . The Reference and Degraded stream condition classifications (Table 3) were the only categories used to test metric sensitivity and index classification efficiency during the development of the indices. Three intermediate categories were also defined, i.e., Minimally Degraded, Mixed (includes all sites with insufficient data to classify condition), and Moderately Degraded. These categories were used as a visual validation that index was appropriately detecting ecological response on a declining gradient from Reference to Degraded.

Table 2. The criteria below were used to assign degradation scores to each sampling event.

Score	Specific Conductivity	pH	Dissolved Oxygen
0	$x \leq 500.0$	$6.0 \leq x \leq 8.5$	$x > 5.0$
1	$500.0 < x < 750.0$	$5.0 \leq x < 6.0$ or $8.5 < x \leq 9.0$	$x \leq 5.0$
2	$750.0 \leq x < 1000.0$	$4.0 \leq x < 5.0$ or $9.0 < x \leq 9.5$	
3	$x \geq 1000.0$	$x < 4.0$ or $x > 9.5$	

Table 3. Sites were classified in to 1 of 5 classes based on habitat and water quality requirements. If more habitat scores were missing, the site could not be classified.

Site Class	Habitat Requirements	Water Quality Requirements
Reference	<ul style="list-style-type: none"> • $\geq 75\%$ of available habitat scores ≥ 16 • No habitat scores < 12 	<ul style="list-style-type: none"> • The sum of the assigned water quality scores equals 0
Minimally Degraded	<ul style="list-style-type: none"> • $\geq 66\%$ of available habitat scores ≥ 16 • $< 75\%$ of available habitat scores ≥ 16 	<ul style="list-style-type: none"> • The sum of the assigned water quality scores equals 0
Mixed	<ul style="list-style-type: none"> • Does not meet the requirements of the other site classes 	<ul style="list-style-type: none"> • Does not meet the requirements of the other site classes
Moderately Degraded	<ul style="list-style-type: none"> • $\geq 50\%$ of available habitat scores ≤ 12 	<ul style="list-style-type: none"> • The sum of the assigned water quality scores is < 2
Degraded	<ul style="list-style-type: none"> • $\geq 50\%$ of available habitat scores ≤ 6 	<ul style="list-style-type: none"> • The sum of assigned water quality scores is > 1

E. Spatial Classification

Classifying nominally undisturbed streams into homogeneous spatial units 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 1,500 mi (2,400 km) 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 km², 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, including 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 region 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 (Appendix G) (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 4 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 often differ from Reference conditions in another bioregion despite both meeting the eight habitat and three water quality criteria. The anthropogenic influences are not necessarily evident in the five stream condition categories (above). They include streamflow flashiness due to urbanization and 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 HUC-12 watersheds are shown by bioregion in Appendix H.

Table 4. Hydrologic and physiographic features used to delineate the twelve bioregions of the Chesapeake Bay basin.

Bioregion Code	Bioregion Name	Area (km ²)	Subregion (HUC 4)	Additional Distinctions	EPA Level III Ecoregion	EPA Level IV Ecoregion
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
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
CA	Central Appalachians	5,986			69 Central Appalachians	69a Forested Hills and Mountains 69b Uplands and Valleys of Mixed Land Use
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
SEP	Southeastern Plains	16,464			65 Southeastern Plains	65n Chesapeake Rolling Coastal Plain 65m Rolling Coastal Plain
BLUE	Blue Ridge	5,175			66 Blue Ridge	66a Northern Igneous Ridges 66b Northern Sedimentary and Metasedimentary Ridges
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

DRAFT REPORT

Bioregion Code	Bioregion Name	Area (km²)	Subregion (HUC 4)	Additional Distinctions	EPA Level III Ecoregion	EPA Level IV Ecoregion
SRV	Southern Ridge and Valley	20,052	Potomac and Lower Chesapeake-James		67 Ridge and Valley	67a Northern Limestone/Dolomite Valleys
						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
UNP	Upper-Northern Piedmont	12,294	Susquehanna and Upper Chesapeake	Great Valley	64 Northern Piedmont	64a Triassic Lowlands
						64b Trap Rock and Conglomerate Uplands
						64c Piedmont Uplands
						64d Piedmont Limestone/Dolomite Lowlands
					67 Ridge and Valley	67a Northern Limestone/Dolomite Valleys
67b Northern Shale Valleys						
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
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
PIED	Piedmont	15,660			45 Piedmont	45c Carolina Slate Belt
						45e Northern Inner Piedmont
						45f Northern Outer Piedmont
						45g Triassic Basins
					58 Northeastern Highlands	58 Reading Prong

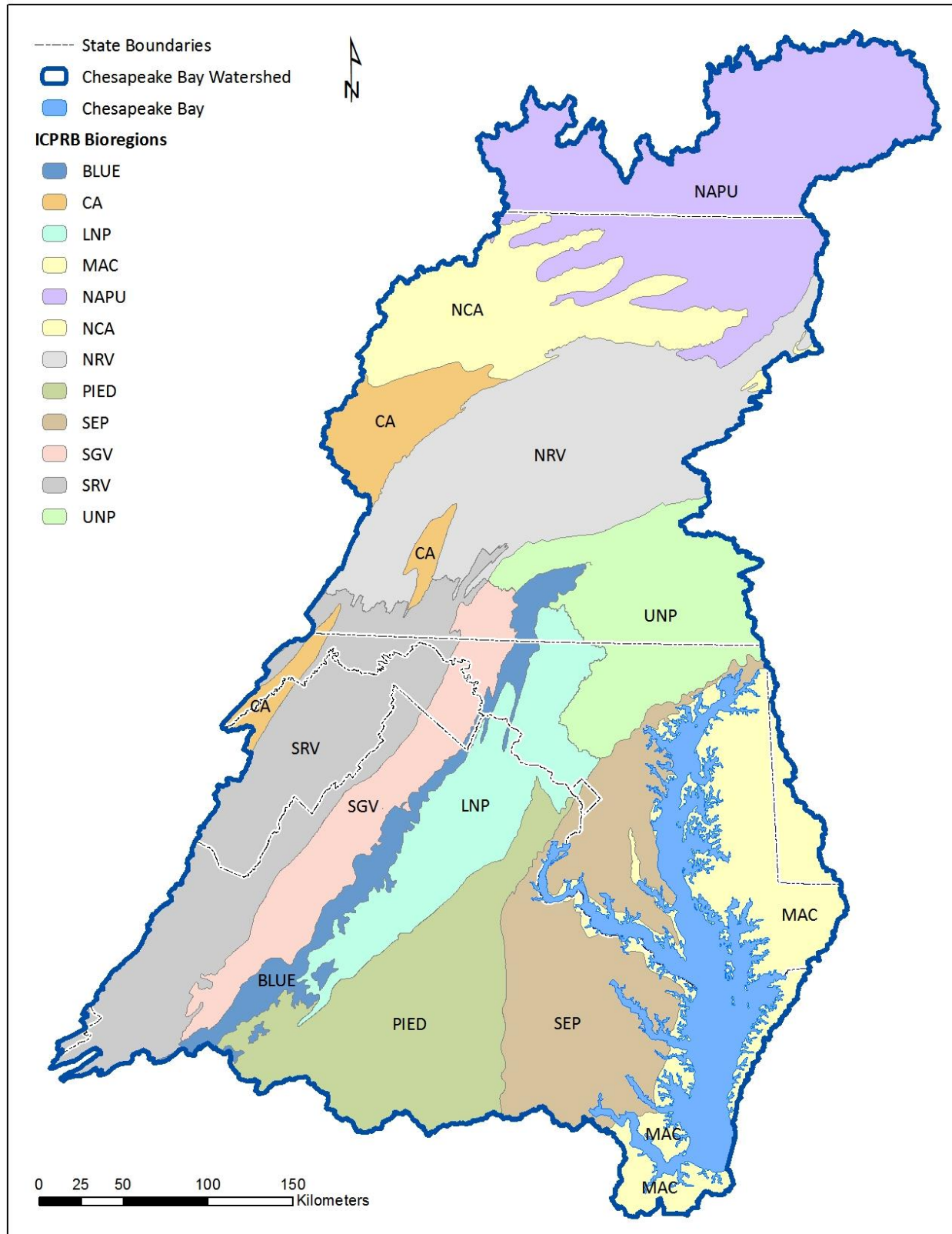


Figure 3. The Chesapeake Bay basin was divided into twelve bioregions for the Chessie BIBI refinement.

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 stream 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 stream 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 were 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 stream 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 the metric's Best Separation Point (BSP). The BSP was used as the threshold to calculate the Balanced Discrimination Efficiency (BDE) for each metric. The metrics with the greatest BDE's were considered for the final index. The BDE equation is effectively the same as Equation 1 for CE.

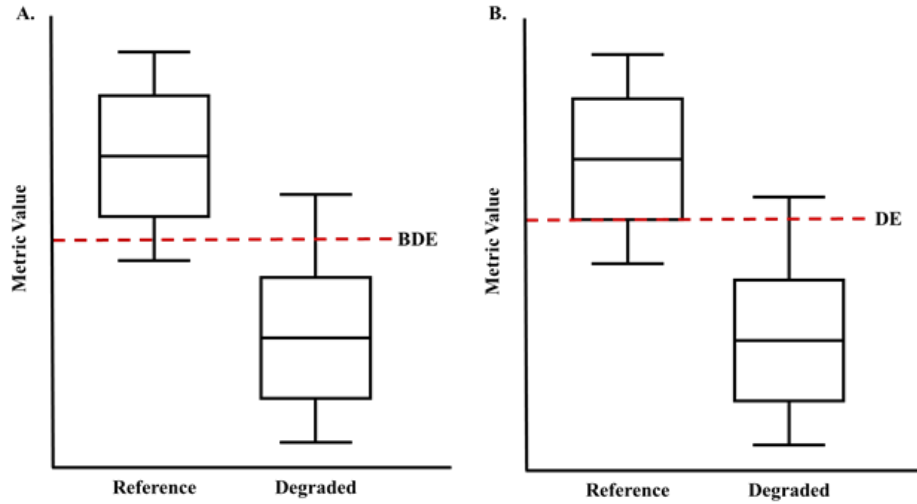


Figure 4. Balanced Discrimination Efficiency (BDE) generally measures metric sensitivity at a different threshold than Discrimination Efficiency (DE). The figures depict metrics that decrease with disturbance. BDE is based on the Best Separation Point (BSP), the point at which the percentage of Reference 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 Section 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 5 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.

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 interquartile range was not greater than the range between 0 to the Reference 25th percentile.

Table 5. 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

iii. Redundancy Analysis

Spearman correlation was used to assess metric redundancy. The final biological index is composed of multiple metrics that evaluate the response of different aspects of the biological assemblage to disturbance. Two significantly correlated metrics in final index is analogous to doubling the weight of a single metric in the index. Therefore, it is necessary to remove redundant metrics to reduce the metric bias in the final index. 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 metric values have a positive or negative relationship. Redundant metrics ($r \geq 0.85$ or $r \leq -0.85$) were compared pairwise using a Wilcoxon-Rank Sum test. No α -value was specified because the test was only used to indicate which metric had greater separation between the Reference and Degraded distributions. The metric with the lower p -value were retained. The metrics remaining after the redundancy analysis were considered for the final index.

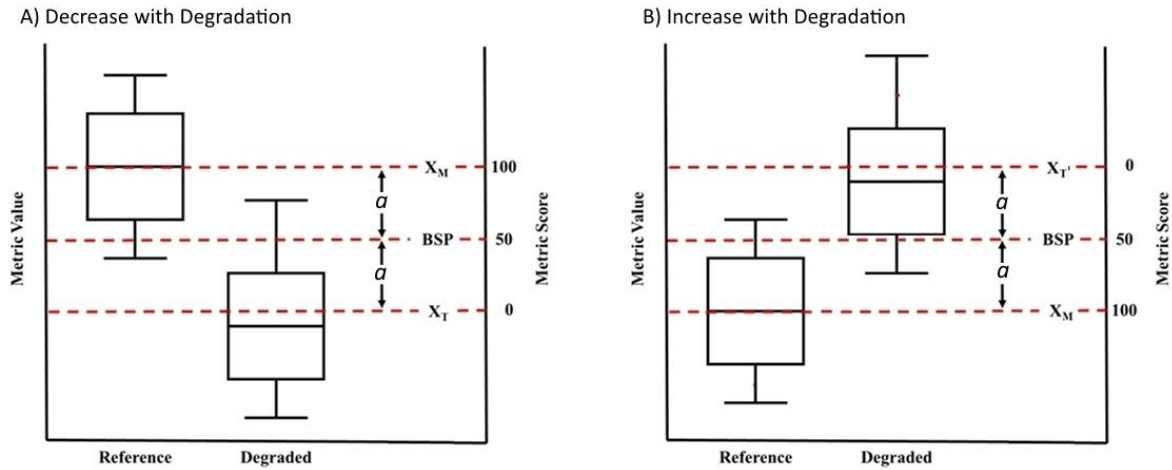


Figure 5. Metric scoring procedure. For metric values that decrease with degradation (Figure 5A), the Reference median (X_M) is the ceiling and X_T is the floor of the 0 – 100 point gradient. Values greater than X_M receive a score of 100; values less than X_T receive a score of 0; values in-between 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.

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. 2008). 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 a lot 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 range was expanded by setting the “floor” equal to twice the distance between the Reference median and the BSP. As before, the Reference median (X_M) was the “ceiling,” or threshold for determining a score of 100 (Figure 5). Metrics that decrease with disturbance received a score of 100 if the metric value was greater than or equal to the X_M threshold. Metrics that increase with disturbance received a score of 100 if the metric value was less than or equal to the X_M threshold. The second threshold (X_T) was derived from twice the distance between the Reference median and the BSP; BSP is the center point of the continuous scoring range, and therefore, is equivalent to a score of fifty. For metrics that decreased with disturbance, the “floor” was X_T , or X_M minus $2a$, where a equals X_M minus BSP (Figure 5A). Metric values less than the value of X_T received a score of zero. For

metrics that increase with disturbance, the “floor” was X_T , or X_M plus 2 a , where a equals BSP minus X_M (Figure 5B). Metric values greater than the value of X_T were assigned a score of zero. On rare occasions, the calculated values of the thresholds X_T and X_T' for percentage metrics (e.g., % EPT) 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.

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

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

Where:

X = the metric value.

X_T = the lower threshold (i.e. floor).

X_M = the upper threshold (i.e. ceiling).

Equation 6

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

Where:

X = the metric value.

$X_{T'}$ = the upper threshold (i.e. ceiling).

X_M = the lower threshold (i.e. floor).

The scores of multiple sampling events for a station were not averaged during index development. Approximately 2,154 stations in the analysis dataset have between 2 and 35 sampling events each. The index scores for the sampling events at one station could 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. We assumed, for the purpose of developing an index, that each sampling event represents the biological response to the immediate environment and is not heavily influenced by the previous year’s biological status. The issue was explored further in Appendix I (Results Pending).

H. Index Construction

Eight ways of constructing a multi-metric index from family-level metrics were examined (Appendix J). The purpose of the exercise was to gain insights into the consequences of choosing a particular 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. 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; then 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 J).

Family-, genus-, and order-level indices were developed for each of the three spatial scales: Chesapeake-wide, region, and bioregion. For the family- and genus-level indices, metrics 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. This means that 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.

Function 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. At least one richness/diversity metric, one composition metric, and one tolerance metric were required for order-level indices. Other sensitive metrics were then added if they improved index CE.

All of the Reference and Degraded samples available for a given bioregion were used to select metrics for that bioregion's indices. When pooled together to form the Inland and Coast data sets and the Chesapeake-wide data set, however, the larger sample sizes in some bioregions biased the metric selection process. We sought to avoid this bias by equalizing to the extent possible each bioregion's representation in the two regions. For stations with multiple sampling events, one sampling event was randomly selected to represent that station. Next, fifty Reference and fifty Degraded sites were randomly selected from each bioregion and aggregated to form the regional data sets. If fewer than fifty samples were found, the bioregion data were not subsampled. The process described above then was used to select metrics for the region indices.

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 earlier in the data preparation's probabilistic rarefaction step affected the metric selections and scoring thresholds. The probabilistic rarefaction process significantly 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). We identified the metrics that occurred most often in ten runs of the probabilistic rarefaction and metric selection steps, and incorporated them into the final indices. The five or more metrics that were selected in eight or more of ten runs were included in the final indices. If fewer than five metrics were selected in eight 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. For example, if the fifth metric was selected in six out of ten runs, all metrics with a frequency greater than or equal to 60% were selected. The means of the metric scoring thresholds calculated in the ten runs were used as the metric scoring thresholds in the final indices. 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.

I. Index Classification Efficiency

Scores of the metrics selected for the final index were averaged to calculate the index scores. The ability of the index to correctly classify Reference and Degraded sites was 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 or genus 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 each of the twelve bioregions and two regions (Inland and Coast). 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). Overfitting refers to the selection of metrics that appear to reflect an ecological response but in actuality 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 (e.g., Southerland et al. 2005, Pond et al. 2008). 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 correctly 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 as an estimate of the results of all of the possible combinations.

Samples were removed from the Reference and Degraded populations. 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 bioregions.

All of the available and applicable data within a bioregion was used to develop 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 of 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 of 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 is developed using all of the data and post-development the data is 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 is repeated many 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 of 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, Table 16) 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. Reference medians for the five large Inland (non-

coastal) bioregions ranged from 60.0 to 76.4 on the 0 – 100 scale in Buchanan et al. (2011) and averaged of 67.4. In the ten new bioregions for the same Inland region, Reference medians were higher, ranging from 70.6 to 88.6 on the 0 – 100 scale and averaging 80.3.

Preliminary results indicated that a common rating scale for the twelve new bioregions could create bioregion biases. Although each bioregion index is scored on a standard 0 - 100 scale, a score from one index may not be directly comparable to a score in another index (Pond et al. 2008). The influence of natural and anthropogenic factors become more pronounced at the smaller bioregion spatial scales (see above), and using a common scale will penalize bioregions that score low for causes that were not accounted for when Reference conditions were established. 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.

Due to their generally lower CEs, order-level indices were rated on a 3-category scale based on the Reference 50th percentile and half of the value of the 10th Reference percentile (Table 6). The family- and genus-level indices were rated on a finer, 5-category scale derived with the 50th, 25th, and 10th percentiles of Reference and half of the value of the 10th percentile (Table 7). 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.

Sampling events excluded because 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 stream condition classes there was no definitive pattern (Table 8). Although there were more Moderately Degraded and Degraded samples with low counts than Reference and Minor Degradation, the majority of samples were classified as Mixed.

Table 6. Thresholds separating the rating categories for order-level index scores. Scores less than or equal to half of the value of the 10th percentile were rated Poor; scores between the half of the value of the 10th Reference percentile and Reference 50th percentiles were rated Fair; scores greater than or equal to the Reference 50th percentile were rated Good.

Rating Threshold	Percentile
Poor Fair	½ of value of 10 th
Fair Good	Reference 50 th

Table 7. Thresholds separating the rating categories for the family- and genus-level index scores. Index scores were defined by the 50th, 25th, and 10th Reference percentiles and half of the value of the 10th percentile.

Rating Threshold	Percentile
Very Poor Poor	½ of value of 10 th
Poor Fair	Reference 10 th
Fair Good	Reference 25 th
Good Excellent	Reference 50 th

Table 8. The number of sampling events with less than or equal seventy individuals identified, aggregated by stream condition class.

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

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 of the data was 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). The 2000 – 2011 period dominates the results. The index scores of stations with multiple samples were averaged and the average rated in order to avoid giving any location disproportionate importance. The 21,552 sampling events in the

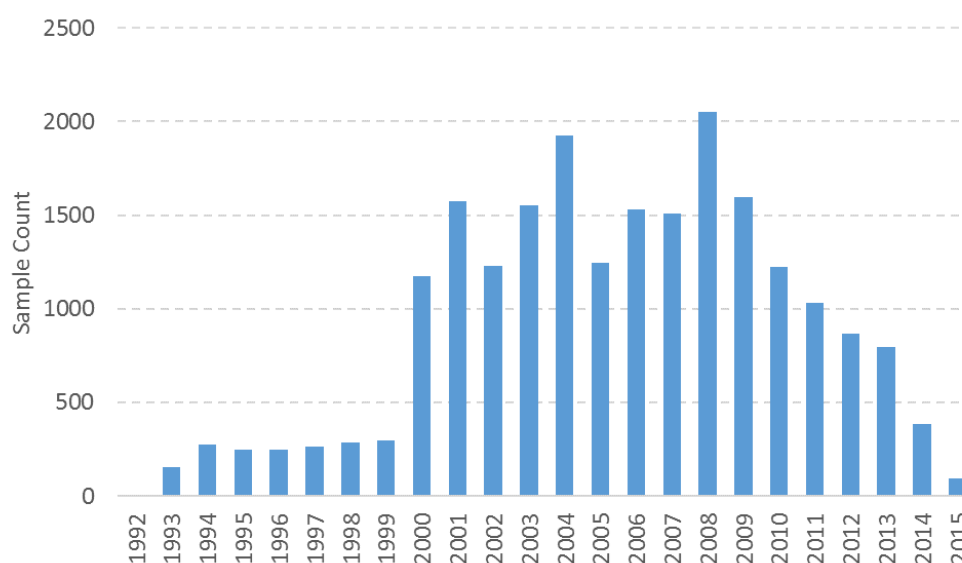


Figure 6. Sample count by year in the analysis database after the data preparation steps described in Methods were applied.

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.

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

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 km²) 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 HUC 12 were used to area-weight their respective scores. Each station rating was first multiplied by the appropriate factor for its HUC12-bioregion unit:

Equation 9

$$\text{Factor} = \frac{\text{area of [HUC 12 – Bioregion unit]}}{\text{number of stations in [HUC 12 – Bioregion unit]}}$$

The weighted ratings for each rating category (Excellent, Good, Fair, Poor, and Very Poor) in each HUC12 were then summed by bioregion and divided by that bioregion's total area. The weighted ratings can also be rolled up to region and basin scales.

For Coast and Inland, the two regional indices, station ratings are weighted by how much of their respective HUC12-region they represent. Most HUC12s (1,922) fall entirely within one region; only 59 overlap the Coast-Inland boundary. The weighted ratings for each category are summed and then 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 25,311 sampling events included in the stream macroinvertebrate database, 21,552 remained after applying the data preparation steps. A majority of sampling events in all bioregions represent intermediate stream conditions and were not used to develop the indices, i.e., Minimally Degraded, Mixed, and Moderately Degraded. The Mixed category also contains

samples with insufficient water quality and stream habitat data. For this study, 2,009 Reference and 1,198 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—1,719 Reference and 1,098 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 eight method comparison (Appendix J) 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 had an enormous impact on 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. Including metrics that are only moderately sensitive to degradation will tend to pull the distributions of Reference and Degraded index scores toward the middle. Metric range and variability criteria tended to protect against the zero inflation issue.

For the Chessie BIBI refinement, we used the index construction strategy that is similar to strategies described in the literature (Method A in Appendix J). 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 assemblage. 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. Function 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 five or six metrics, although some bioregions had as many nine or ten metrics. Order-level indices were much more variable, ranging from one metric (MAC) to eight (NAC, NRV).

B. Chesapeake-Wide Index

A preliminary watershed-wide index was developed using all Reference and Degraded sampling events regardless of their locations. Each taxonomic version of the index had relatively high classification efficiencies (CEs): 79.7% for order-level, 84.6% for family-level, and 83.9% for genus-level. Six, five, and eight metrics were included in the order-, family-, and genus-level versions, respectively. Reference sampling events on the Inland side of the watershed had a uneven influence in shaping the index because they were 26.5 times more numerous in the Chessie BIBI database (Table 9). We concluded this preliminary index was biased and misrepresented the Coast region.

To minimize bias, we randomly reduced numbers of Inland sampling events to 50 Reference and 50 Degraded per bioregion before running multiple iterations of the R-program that selects metrics and develops the index. The most frequently selected metrics (Table 10) were used to build the index. Classification efficiencies (CEs) were 80.4% for the order-level (BSP = 43), 84.7% for the family-level (BSP = 49), and 82.9% for the genus-level (BSP = 46).

Table 9. Macroinvertebrate sample numbers in the Coast and Inland stream condition categories (and percent of the Region's total sample number).

Region	Reference	Minimally Degraded	Mixed	Moderately Degraded	Degraded	Region Total
COAST	73 (1.4%)	99 (1.9%)	3,746 (71.9%)	856 (16.4%)	439 (8.4%)	5,213
INLAND	1,936 (11.8%)	583 (3.6%)	9,157 (56.0%)	4,004 (24.5%)	659 (4.0%)	16,339
Watershed Total						21,552

Table 10. 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
GOLD	Composition	X		X
PCT_ARTHROPODA	Composition	X		
PCT_COTE	Composition	X	X	
PCT_EPT	Composition	X	X	
PCT_HEXAPODA	Composition			X
PCT_INSECTA	Composition	X		
PCT_CLING	Habit		X	X
MARGALEFS	Richness/Diversity		X	
PIELOU	Richness/Diversity			X
RICH_EPHEMEROPTERA	Richness/Diversity		X	
RICH_EPT	Richness/Diversity		X	
PCT_DOM3	Tolerance		X	
PCT_INTOL_0_4	Tolerance			X

C. Two Region Indices

Uneven distributions of Reference and Degraded sites also can bias development of the two Region indices (i.e., Inland and Coast). Reference and Degraded samples are unevenly distributed across the bioregions within each region. The Inland region has ten bioregions and over half of its Reference samples are located in three bioregions: LNP, NCA, and SRV. The Coast has two bioregions and about two-thirds of Reference samples are located 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 well-sampled bioregions prior to index development.

i. Order-Level Indices

Redundancy, range, variability, and sensitivity assessments prevented the selection of a tolerance metric for the Coast order-level index, so that index was represented by four composition metrics and two richness/diversity metrics (Table 11). The Inland order-level index was represented by five composition metrics, one tolerance metric, and two richness/diversity metrics. Overall, the mean metric BDE was 66.4% and ranged from 54.2% to 74.5%. The Coast BSP was 58; the Inland BSP, 54. The Inland index CE was 76.9%; the Coast index, 75.5%. Figure 7 shows declining trends in index scores as degradation increases.

ii. Family-Level Indices

Both indices included metrics from four of the five metric categories (Table 12). Six metrics were included in both indices. The mean metric BDE for both regions was 72.5%. The Inland metric BDEs ranged from 66.6% - 78.3%, and the Coast metric BDEs ranged from 61.8% - 79.4%. The Inland index CE was 79.8%; the Coast, 78.6%. The Inland BSP of 52 did not differ greatly from the expected value of 50, but the Coast BSP of 59 was slightly elevated. Figure 8 shows declining trends in index scores as degradation increases.

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 13). At least four metric types were selected for both indices (Table 14). The metric BDEs in the Coast region ranged from 51.1% to 68.5% while BDEs in the Inland region ranged from 64.9% to 78.8%. Overall, the mean BDE for both regions was 68.3%. The Inland index CE was 78.9%; the Coast index, 70.8%. The BSP values for both indices did not differ much from the expected BSP value of 50. Figure 9 depicts a declining trend in the index scores with increasing degradation.

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

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
COAST	GOLD	Composition	Decrease	55.95	34.39	68.4
COAST	PCT_ARTHROPODA	Composition	Decrease	95.52	94.61	54.2
COAST	PCT_DIPTERA	Composition	Increase	42.15	52.15	64.7
COAST	PCT_INSECTA	Composition	Decrease	85.56	75.34	61.8
COAST	MARGALEFS	Richness/Diversity	Decrease	1.33	1.22	67.6
COAST	PIELOU	Richness/Diversity	Decrease	0.69	0.62	69.5
INLAND	GOLD	Composition	Decrease	79.68	64.68	70.9
INLAND	PCT_COTE	Composition	Decrease	59.99	34.35	72.8
INLAND	PCT_EPT	Composition	Decrease	67.75	41.89	74.9
INLAND	PCT_INSECTA	Composition	Decrease	99.11	97.91	63.9
INLAND	PCT_POTEC	Composition	Decrease	76.63	53.89	74.5
INLAND	MARGALEFS	Richness/Diversity	Decrease	1.08	1.08	58.4
INLAND	PIELOU	Richness/Diversity	Decrease	0.76	0.68	64.1
INLAND	PCT_DOM3	Tolerance	Increase	86.71	91.71	64.1
Mean						66.4

Table 12. 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	PCT_INSECTA	Composition	Decrease	89.82	76.07	61.8
COAST	RICH_COLLECT	Richness/Diversity	Decrease	7.00	4.98	72.7
COAST	PCT_COLLECT	FFG	Increase	82.52	89.62	65.5
COAST	ASPT_MOD	Tolerance	Decrease	4.35	2.78	70
COAST	HBI	Tolerance	Increase	5.17	6.12	77.8
COAST	PCT_TOLERANT_5_10	Tolerance	Increase	80.62	99.80	79.4
INLAND	PCT_COTE	Composition	Decrease	60.17	34.58	72.8
INLAND	RICH_COLLECT	Richness/Diversity	Decrease	7.86	6.03	66.6
INLAND	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	4.00	1.58	78.3
INLAND	RICH_EPT_NO_TOL	Richness/Diversity	Decrease	7.00	4.51	75.7
INLAND	PCT_CLING	Habit	Decrease	62.51	42.65	72.1
INLAND	PCT_INTOL_0_3	Tolerance	Decrease	50.18	19.67	77.8
Mean						72.5

Table 13. 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	72	96	3,328	849	439	4,784
INLAND	1,647	516	7,615	3,700	579	14,057
Total Count	1,719	612	10,943	4,549	1,018	18,841

Table 14. 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	1.06	68.5
COAST	GOLD	Composition	Decrease	56.57	34.72	68.4
COAST	PCT_INSECTA	Composition	Decrease	88.86	73.48	61.8
COAST	RICH_FILTER	Richness/Diversity	Decrease	2.86	1.05	57
COAST	PCT_COLLECT	FFG	Decrease	73.26	70.94	51.1
INLAND	GOLD	Composition	Decrease	78.30	61.36	72.4
INLAND	PCT_COTE	Composition	Decrease	59.01	28.69	76
INLAND	PCT_POTEC	Composition	Decrease	75.69	51.29	76.6
INLAND	PCT_EPT_RICH	Richness/Diversity	Decrease	63.95	26.18	71.1
INLAND	PIELOU	Richness/Diversity	Decrease	0.82	0.74	64.9
INLAND	PCT_CLING	Habit	Decrease	60.20	33.60	73.1
INLAND	PCT_INTOL_0_4	Tolerance	Decrease	66.85	37.85	78.8
Mean						68.3

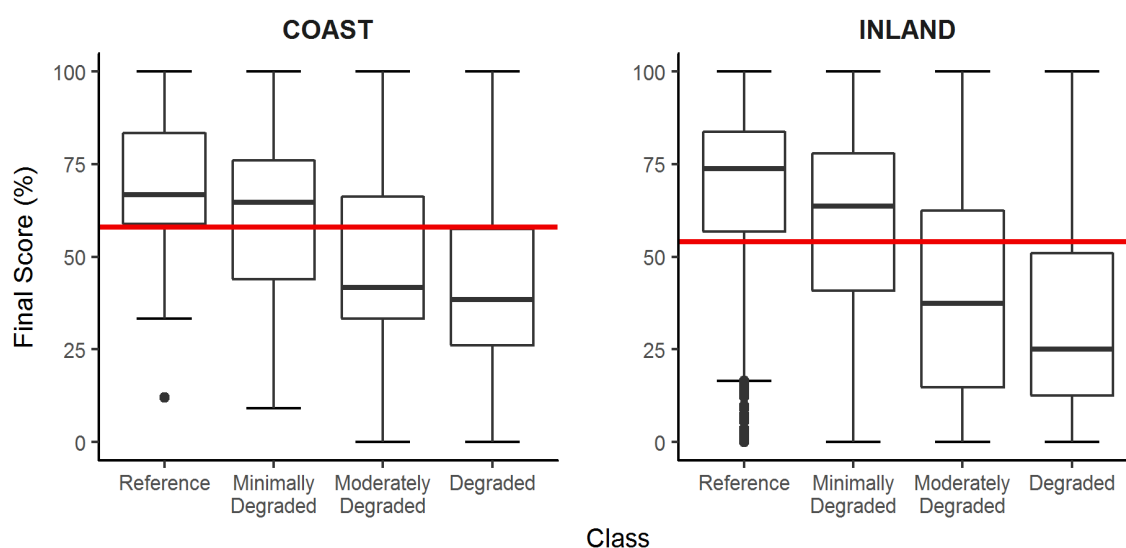


Figure 7. Distributions of index scores for the Coast and Inland order-level indices. The whisker lengths are designated by the interquartile range multiplied by 1.5. The horizontal red line represents the Best Separation Point (BSP) between the Reference and Degraded distributions. BSP was used as a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

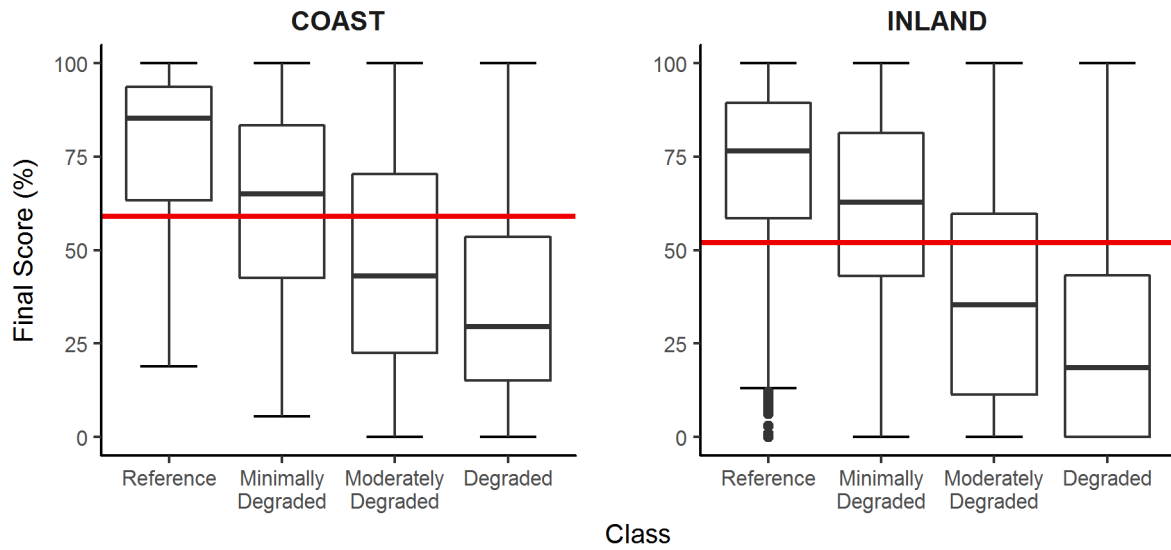


Figure 8. Distributions of index scores for the Coast and Inland family-level indices. The whisker lengths are designated by the interquartile range multiplied by 1.5. The horizontal red line represents the Best Separation Point (BSP) between the Reference and Degraded distributions. BSP was used as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

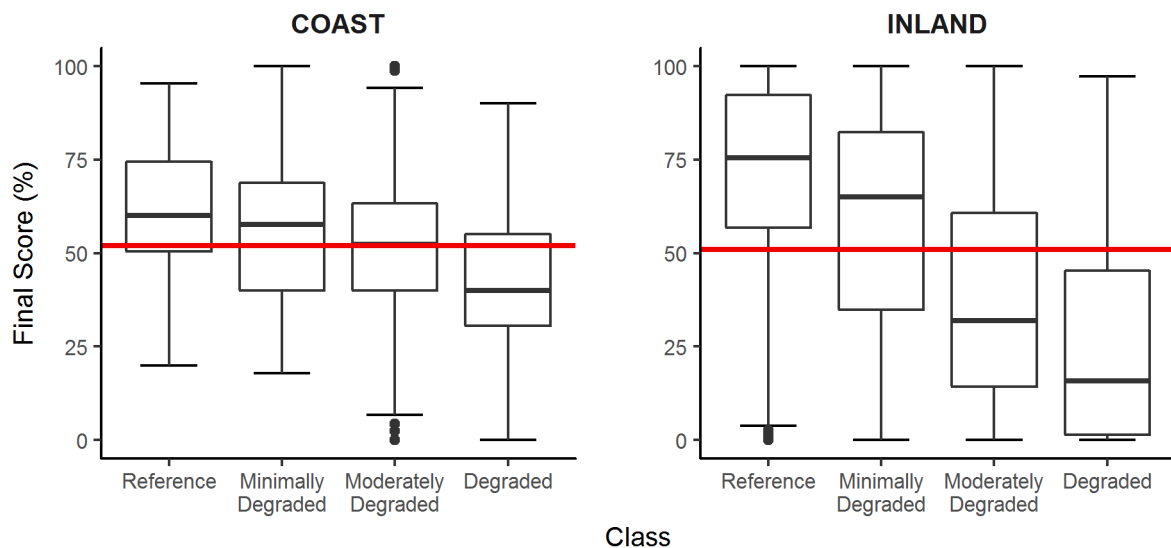


Figure 9. Distributions of index scores for the Coast and Inland genus-level indices. The whisker lengths are designated by the interquartile range multiplied by 1.5. The horizontal red line represents the Best Separation Point (BSP) between the Reference and Degraded distributions. BSP was used as a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

D. Bioregion Indices

The MAC and CA bioregions had fewer than 50 Reference sampling events; the Blue BLUE and PIED bioregions had fewer than 50 Degraded sampling events (Table 15). 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 $n = 372$; genus $n = 302$) and the LNP the best (order and family $n = 3,625$; genus $n = 3,224$).

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, five metrics were selected for each order-level indices (Table 16). Redundancy, range, variability, and sensitivity assessments limited the representation of these metric classes in the Mid-Atlantic Coastal Plain (MAC) and Southeastern Plains (SEP) bioregions. A single metric was selected for the MAC: PCT_DOM4 (the percentage of the assemblages comprised of the four most dominant taxa). Similarly, PCT_DOM5 and PCT_INSECTA were the only metrics selected for the SEP index. 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 70.4% but CE values ranged from 58.1%, or little better than a coin toss, in NAPU to 83.8% in BLUE. The mean index BSP ($\bar{x} = 52.5$) did not vary drastically from the expected value of 50.0 (Table 17). Index scores showed a weak descending trend from Reference to Degraded conditions (Figure 10 and Figure 11).

ii. Family-Level Indices

The average number of metrics selected for each bioregion was six; the greatest number of metrics was selected in the SEP bioregion (Table 18). Although a minimum of four metrics was nominally required for each index, redundancy, range, variability, and sensitivity tests prevented the selection of four metric categories in the NRV bioregion. The NRV index was composed of one composition metric, one FFG metric, and three richness/diversity metrics. The mean BDE of all of the metrics was 73.1%. CEs of the indices averaged 81.1%, and ranged from 69.3% in the NAPU to 89.0% in the LNP. The BSP's averaged 51.8, and ranged from 45 (SRV) to 61 (PIED) (Table 17). In general, a descending gradient as degradation increases was observed for each bioregion index (Figure 12 and Figure 13).

iii. Genus-Level Indices

Sampling events used during the analysis were required to have 90% or more taxa identified to the genus-level. A total of 2,929 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 19). Despite the reduction of Reference and Degraded samples, many of the bioregions retained a substantial sample size for index development. However, the CA and the MAC bioregions had fewer than 50 Reference samples, and the BLUE, CA, NAPU, PIED, and SRV had fewer than 50 Degraded samples.

Four out of the five metric types were typically represented in each index, although assessments of redundancy, range, variability, and sensitivity prevented the selection of more than three metric types in the UNP (Table 20). The fewest metrics ($n = 5$) were selected for BLUE, NAPU, and SEP. The most metrics ($n = 9$) were selected for PIED index. On average

approximately six metrics were selected for each bioregion. The mean BDE of all of the metrics selected within the basin was 75.2%. The mean CE was 82.5% but ranged from 69.9% (SEP) to 88.6% (NCA). The mean BSP (\bar{x} = 49.8) did not differ from the expected BSP of 50 (Table 17). The lowest BSP was observed in the NRV (x = 45), while the greatest BSP was found in the MAC (x = 61). Figure 14 and Figure 15 depict the final index scores of the sampling event categories.

Table 15. Order- and family-level Macroinvertebrate sample numbers in each bioregion's stream condition categories (and percent of the bioregion's total sample number).

Bioregion	Reference	Minimally Degraded	Mixed	Moderately Degraded	Degraded	Bioregion Count
BLUE	134 (30%)	27 (6.1%)	242 (54.3%)	37 (8.3%)	6 (1.3%)	446
CA	39 (10.5%)	17 (4.6%)	204 (54.8%)	71 (19.1%)	41 (11%)	372
LNP	357 (9.8%)	85 (2.3%)	1,819 (50.2%)	1,225 (33.8%)	139 (3.8%)	3,625
MAC	21 (1.3%)	32 (2%)	1,108 (69.7%)	216 (13.6%)	212 (13.3%)	1,589
NAPU	94 (6.7%)	56 (4%)	878 (62.5%)	326 (23.2%)	51 (3.6%)	1,405
NCA	318 (36.3%)	28 (3.2%)	385 (44%)	89 (10.2%)	55 (6.3%)	875
NRV	161 (14.7%)	32 (2.9%)	651 (59.4%)	198 (18.1%)	54 (4.9%)	1,096
PIED	135 (10.8%)	57 (4.6%)	676 (54.1%)	356 (28.5%)	25 (2%)	1,249
SEP	52 (1.4%)	67 (1.8%)	2,638 (72.8%)	640 (17.7%)	227 (6.3%)	3,624
SGV	183 (11%)	63 (3.8%)	681 (40.9%)	647 (38.9%)	91 (5.5%)	1,665
SRV	433 (19.5%)	152 (6.8%)	1,221 (55%)	365 (16.4%)	49 (2.2%)	2,220
UNP	82 (2.4%)	66 (1.9%)	2,400 (70.9%)	690 (20.4%)	148 (4.4%)	3,386
Total						21,552

Table 16. Metrics included in the bioregion-specific, order-level indices.

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
BLUE	PCT_COTE	Composition	Decrease	64.57	47.11	67.7
BLUE	PCT_EPT	Composition	Decrease	82.50	39.96	85.0
BLUE	PIELOU	Richness/Diversity	Increase	0.74	0.82	66.2
BLUE	PCT_DOM1	Tolerance	Decrease	48.26	34.60	68.8
BLUE	PCT_DOM3	Tolerance	Increase	87.61	87.61	50.0
BLUE	PCT_DOM4	Tolerance	Decrease	95.58	95.58	50.0
CA	PCT_POTEC	Composition	Decrease	77.29	73.37	58.8
CA	PIELOU	Richness/Diversity	Decrease	0.78	0.68	68.8
CA	SIMPSONS	Richness/Diversity	Decrease	0.71	0.63	70.0
CA	PCT_DOM4	Tolerance	Increase	95.69	98.79	66.3
LNP	PCT_COTE	Composition	Decrease	69.83	17.47	85.4
LNP	MARGALEFS	Richness/Diversity	Decrease	1.29	1.27	63.9
LNP	RICH	Richness/Diversity	Decrease	7.00	7.00	58.6
LNP	PCT_DOM1	Tolerance	Increase	46.00	56.66	65.0
MAC	PCT_DOM4	Tolerance	Increase	90.43	93.79	61.6
NAPU	PCT_COTE	Composition	Decrease	64.86	56.88	57.6
NAPU	PCT_EPT	Composition	Decrease	62.08	51.98	57.6
NAPU	PCT_POTEC	Composition	Decrease	73.25	63.79	60.7
NAPU	SIMPSONS	Richness/Diversity	Decrease	0.68	0.68	52.5
NAPU	PCT_DOM4	Tolerance	Increase	96.00	96.74	53.1
NCA	PCT_COTE	Composition	Decrease	64.19	48.37	68.5
NCA	PCT_EPHEMEROPTERA	Composition	Decrease	40.90	10.70	78.5
NCA	PCT_POTEC	Composition	Decrease	82.30	79.20	55.2
NCA	PIELOU	Richness/Diversity	Decrease	0.79	0.73	58.7
NCA	RICH	Richness/Diversity	Decrease	6.00	5.99	59.5
NCA	PCT_DOM1	Tolerance	Increase	46.03	55.29	63.3
NCA	PCT_DOM3	Tolerance	Increase	87.18	92.06	61.9
NCA	PCT_DOM4	Tolerance	Increase	95.69	98.69	67.6
NRV	GOLD	Composition	Decrease	81.08	72.56	57.6
NRV	PCT_COTE	Composition	Decrease	60.00	46.40	59.1
NRV	PCT_EPT	Composition	Decrease	64.09	49.87	61.6
NRV	PCT_POTEC	Composition	Decrease	74.29	61.87	61.0
NRV	RICH	Richness/Diversity	Decrease	6.00	6.00	61.0
NRV	PCT_DOM2	Tolerance	Increase	71.50	75.70	54.5
NRV	PCT_DOM4	Tolerance	Increase	93.43	97.91	63.5
PIED	PCT_ARTHROPODA	Composition	Decrease	98.11	94.31	63.9
PIED	PCT_COTE	Composition	Decrease	62.81	36.35	80.4
PIED	PCT_EPHEMEROPTERA	Composition	Decrease	34.07	14.61	76.5
PIED	HURLBERTS_PIE	Richness/Diversity	Decrease	0.75	0.67	70.4

DRAFT REPORT

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
PIED	PCT_DOM4	Tolerance	Increase	89.76	93.78	64.6
SEP	PCT_INSECTA	Composition	Decrease	84.77	81.81	57.7
SEP	PCT_DOM5	Tolerance	Increase	93.98	99.28	71.5
SGV	GOLD	Composition	Decrease	77.17	72.83	60.0
SGV	PCT_ARTHROPODA	Composition	Decrease	97.22	95.70	54.5
SGV	PCT_COTE	Composition	Decrease	65.79	43.47	72.6
SGV	PCT_HEXAPODA	Composition	Decrease	96.58	88.32	71.2
SGV	PIELOU	Richness/Diversity	Decrease	0.77	0.73	60.3
SGV	RICH	Richness/Diversity	Decrease	7.00	7.00	57.5
SGV	PCT_DOM4	Tolerance	Increase	92.93	96.25	58.6
SRV	PCT_ARTHROPODA	Composition	Decrease	99.15	98.99	53.0
SRV	PCT_COTE	Composition	Decrease	58.60	47.80	63.3
SRV	PCT_EPT	Composition	Decrease	68.04	56.44	59.6
SRV	PCT_HEXAPODA	Composition	Decrease	98.41	97.63	55.0
SRV	MARGALEFS	Richness/Diversity	Decrease	1.08	1.08	56.1
SRV	PIELOU	Richness/Diversity	Decrease	0.78	0.70	68.2
SRV	PCT_DOM2	Tolerance	Increase	68.94	78.54	65.7
SRV	PCT_DOM4	Tolerance	Increase	94.74	96.98	60.6
UNP	GOLD	Composition	Decrease	73.35	24.33	76.6
UNP	PCT_EPT	Composition	Decrease	57.21	1.09	81.4
UNP	PCT_HEXAPODA	Composition	Decrease	99.49	95.15	81.7
UNP	HURLBERTS_PIE	Richness/Diversity	Decrease	0.67	0.57	67.3
UNP	PCT_DOM4	Tolerance	Increase	95.61	96.65	56.4
Mean						63.9

Table 17. The Best Separation Point (BSP) for bioregion index scores The BSP is used as the threshold for calculating Classification Efficiency (CE).

Bioregion	<u>Order-Level Index</u>		<u>Family-Level Index</u>		<u>Genus-Level Index</u>	
	Index BSP	Index CE	Index BSP	Index CE	Index BSP	Index CE
BLUE	49	83.8	56	84.6	46	83.8
CA	52	72.5	48	80.0	50	83.0
LNP	50	64.4	52	89.0	51	86.8
MAC	53	61.9	58	82.2	61	81.3
NAPU	59	58.1	49	69.3	49	74.3
NCA	55	67.6	50	77.5	47	88.6
NRV	52	65.0	51	79.9	45	88.1
PIED	55	80.4	61	85.0	48	86.3
SEP	51	66.0	51	76.1	50	69.9
SGV	52	72.3	52	84.4	52	86.0
SRV	51	70.9	45	79.2	51	76.7
UNP	51	81.4	49	86.5	47	84.9
Mean	52.5	70.4	51.8	81.1	49.8	82.5

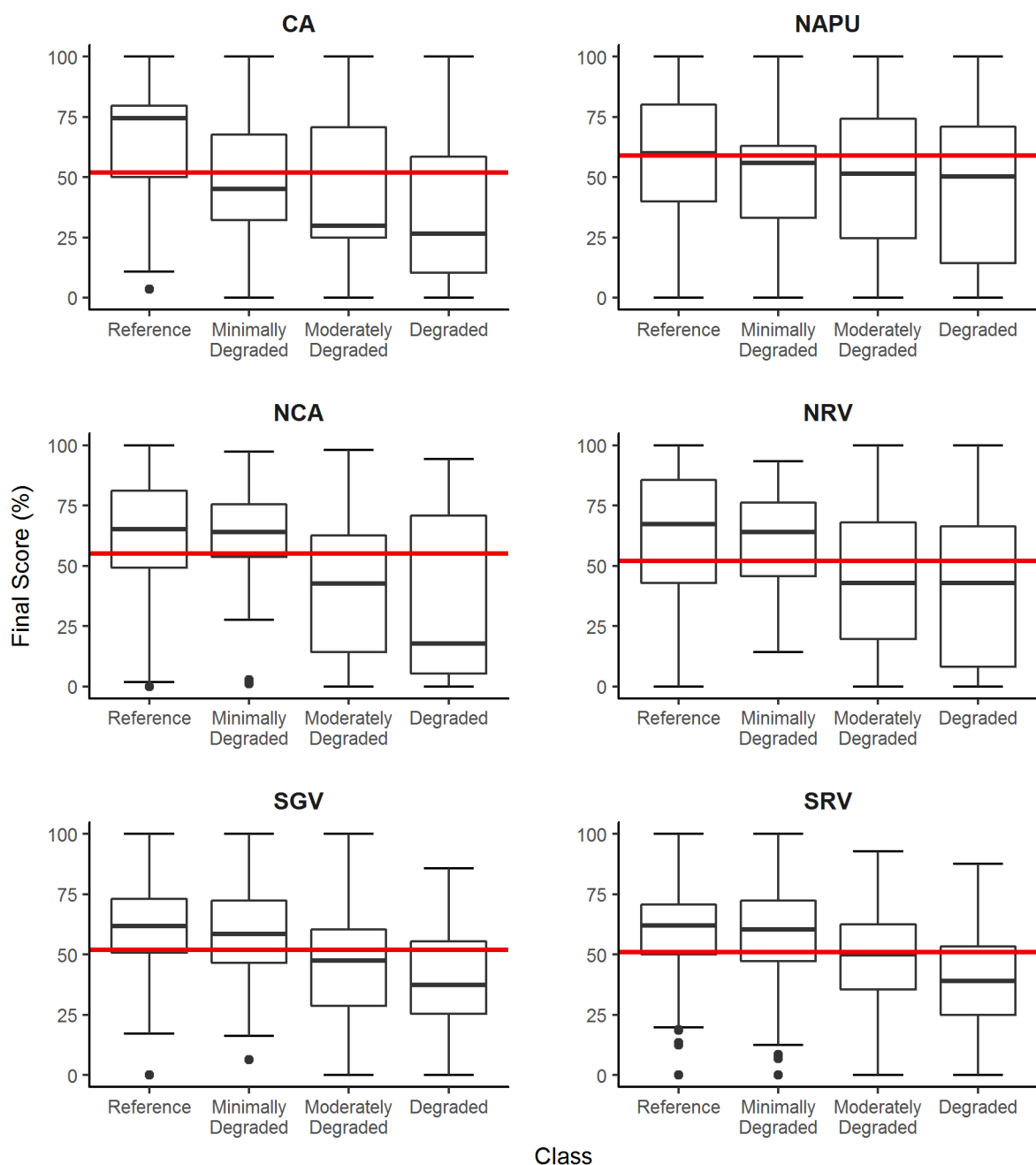


Figure 10. Distributions of index scores for order-level indices in six bioregions: Central Appalachians (CA), Northern Appalachian Plateau and Uplands (NAPU), Northern Central Appalachians (NCA), Northern Ridge and Valley (NRV), Southern Great Valley (SGV), and Southern Ridge and Valley (SRV). The whisker 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).

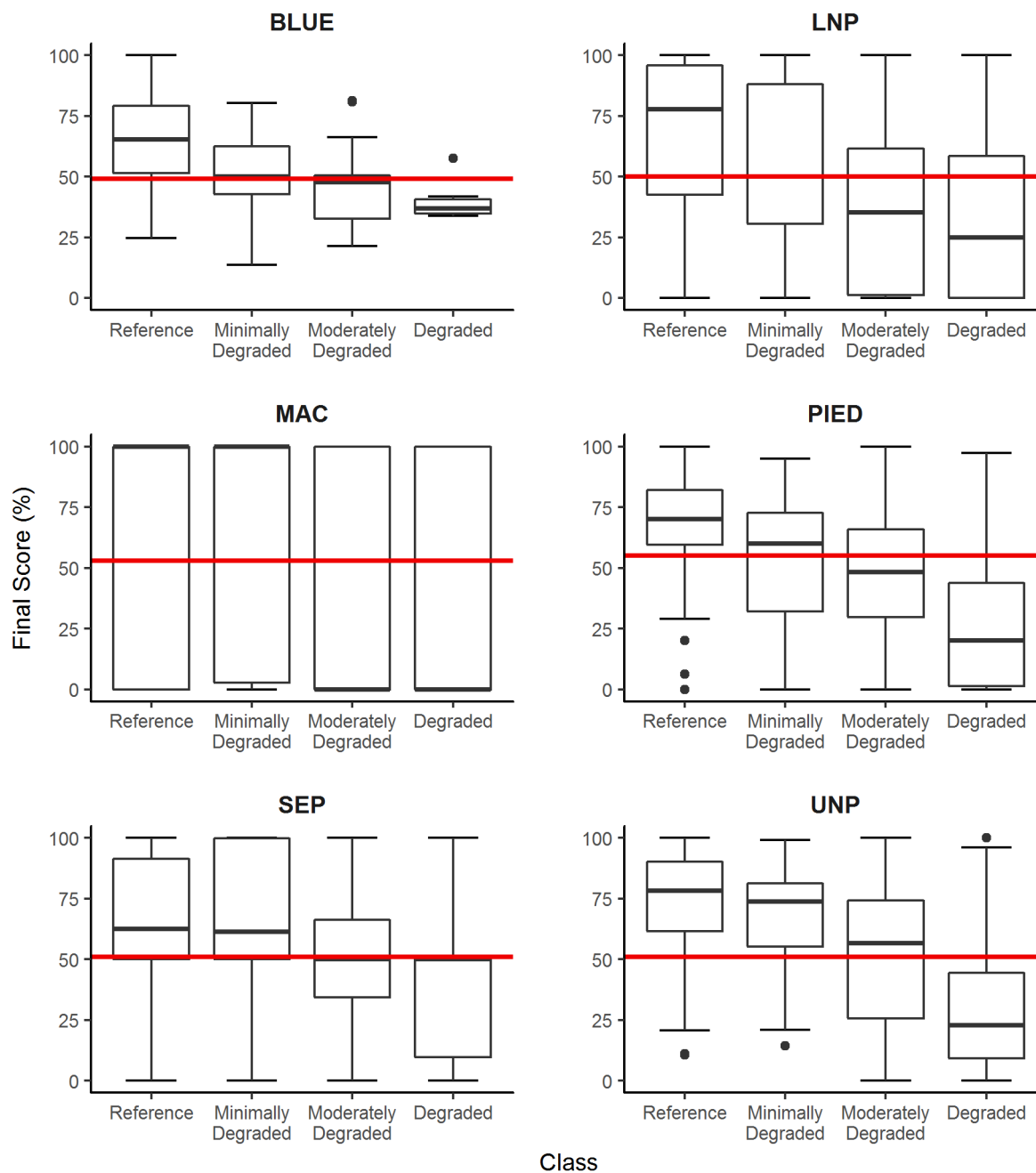


Figure 11. Distributions of index scores for order-level indices in six bioregions: Blue Ridge Mountains (BLUE), Lower Northern Piedmont (LNP), Middle-Atlantic Coast (MAC), Piedmont (PIED), Southeastern Plains (SEP), and Upper Northern Piedmont (UNP). The whisker lengths are designated by the Interquartile 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).

Table 18. Metrics included in the family-level bioregion indices.

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
BLUE	PCT_EPT	Composition	Decrease	82.50	39.96	85.0
BLUE	PCT_EPT_RICH	Richness/Diversity	Decrease	69.28	45.26	93.7
BLUE	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	4.00	1.06	84.2
BLUE	PCT_CLING	Habit	Decrease	75.36	58.32	66.9
BLUE	PCT_DOM5	Tolerance	Increase	73.96	92.00	83.1
CA	PCT_EPT_NO_HYDRO	Composition	Decrease	86.52	72.50	68.8
CA	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	3.00	1.43	68.5
CA	RICH_SHRED	Richness/Diversity	Decrease	3.00	1.40	68.6
CA	PCT_COLLECT	FFG	Increase	54.15	89.75	71.3
CA	ASPT_MOD	Tolerance	Decrease	6.13	5.33	68.8
CA	PCT_DOM2	Tolerance	Increase	45.76	73.60	81.3
CA	PCT_INTOL_0_3	Tolerance	Decrease	51.97	23.17	76.3
LNP	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	3.00	2.51	78.9
LNP	PCT_COLLECT	FFG	Increase	62.37	92.87	80.6
LNP	PCT_CLING	Habit	Decrease	72.73	23.67	85.2
LNP	PCT_DOM1	Tolerance	Increase	28.69	53.61	78.3
LNP	PCT_INTOL_0_3	Tolerance	Decrease	64.00	0.00	88.2
MAC	RICH_PREDATOR	Richness/Diversity	Decrease	3.00	1.40	63.8
MAC	PCT_COLLECT	FFG	Increase	71.43	92.21	70.4
MAC	PCT_BURROW	Habit	Decrease	59.22	55.94	56.9
MAC	ASPT_MOD	Tolerance	Decrease	3.22	2.98	61.6
MAC	PCT_TOLERANT_5_10	Tolerance	Decrease	98.17	98.13	60.9
NAPU	PCT_EPT_NO_HYDRO	Composition	Decrease	85.92	73.38	62.8
NAPU	PCT_EPT_HYDRO_BAETID	Composition	Decrease	70.38	57.04	66.8
NAPU	RICH_CLING	Richness/Diversity	Decrease	8.00	6.44	64.2
NAPU	PCT_CLING	Habit	Decrease	59.30	47.06	60.7
NAPU	PCT_MOD_TOL_4_6	Tolerance	Increase	56.41	71.87	61.7
NCA	PCT_EPHEMEROPTERA	Composition	Decrease	40.90	10.70	78.5
NCA	MARGALEFS	Richness/Diversity	Decrease	3.22	2.62	78.1
NCA	RICH_COLLECT	Richness/Diversity	Decrease	8.00	6.00	75.4
NCA	RICH_GATHER	Richness/Diversity	Decrease	5.00	5.00	74.0
NCA	PCT_GATHER	FFG	Decrease	42.24	26.14	68.5
NCA	PCT_TOLERANT_5_10	Tolerance	Increase	28.91	36.57	61.6
NRV	PCT_EPT_HYDRO_BAETID	Composition	Decrease	75.00	56.70	65.0
NRV	RICH_COLLECT	Richness/Diversity	Decrease	8.00	6.00	71.9
NRV	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	4.00	1.76	76.2
NRV	RICH_GATHER	Richness/Diversity	Decrease	5.10	4.30	71.9
NRV	PCT_COLLECT	FFG	Increase	76.12	83.30	59.1

DRAFT REPORT

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
PIED	PCT_COTE	Composition	Decrease	62.81	36.35	80.4
PIED	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	3.60	1.21	82.2
PIED	RICH_SCRAPE	Richness/Diversity	Decrease	2.00	0.86	78.7
PIED	PCT_CLING	Habit	Decrease	65.64	33.08	79.3
PIED	HBI	Tolerance	Increase	3.78	5.06	84.2
SEP	PCT_INSECTA	Composition	Decrease	84.77	81.81	57.7
SEP	RICH_CLING	Richness/Diversity	Decrease	4.95	1.54	78.3
SEP	RICH_FILTER	Richness/Diversity	Decrease	2.00	0.36	72.6
SEP	RICH_PREDATOR	Richness/Diversity	Decrease	3.00	1.77	62.1
SEP	PCT_COLLECT	FFG	Increase	80.67	88.27	61.4
SEP	ASPT_MOD	Tolerance	Decrease	4.95	2.57	67.4
SEP	HBI	Tolerance	Increase	5.01	5.95	74.1
SEP	PCT_DOM1	Tolerance	Increase	38.08	57.54	67.4
SEP	PCT_DOM5	Tolerance	Increase	82.95	94.49	76.6
SEP	PCT_TOLERANT_5_10	Tolerance	Increase	69.46	100.00	70.3
SGV	PCT_COTE	Composition	Decrease	65.79	43.47	72.6
SGV	RICH_SCRAPE	Richness/Diversity	Decrease	2.33	1.10	76.2
SGV	PCT_COLLECT	FFG	Increase	71.56	94.20	78.4
SGV	ASPT_MOD	Tolerance	Decrease	6.27	4.13	76.4
SGV	PCT_DOM4	Tolerance	Increase	71.52	86.02	76.4
SGV	PCT_INTOL_0_4	Tolerance	Decrease	51.69	10.09	82.2
SRV	PCT_COTE	Composition	Decrease	58.60	47.80	63.3
SRV	RICH_CLING	Richness/Diversity	Decrease	9.00	6.83	74.7
SRV	RICH_PREDATOR	Richness/Diversity	Decrease	3.00	2.65	70.8
SRV	RICH_SCRAPE	Richness/Diversity	Decrease	2.00	2.00	73.4
SRV	PCT_COLLECT	FFG	Increase	65.25	78.29	65.8
SRV	PCT_DOM2	Tolerance	Increase	46.67	62.05	73.7
UNP	GOLD	Composition	Decrease	73.35	24.33	76.6
UNP	PCT_EPT	Composition	Decrease	57.21	1.09	81.4
UNP	PCT_HEXAPODA	Composition	Decrease	99.49	95.15	81.7
UNP	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	3.00	1.99	75.9
UNP	PCT_GATHER	FFG	Increase	56.26	79.92	78.2
UNP	PCT_CLING	Habit	Decrease	63.32	17.36	76.9
UNP	PCT_URBAN_INTOL	Tolerance	Decrease	98.19	91.31	75.3
Mean						73.1

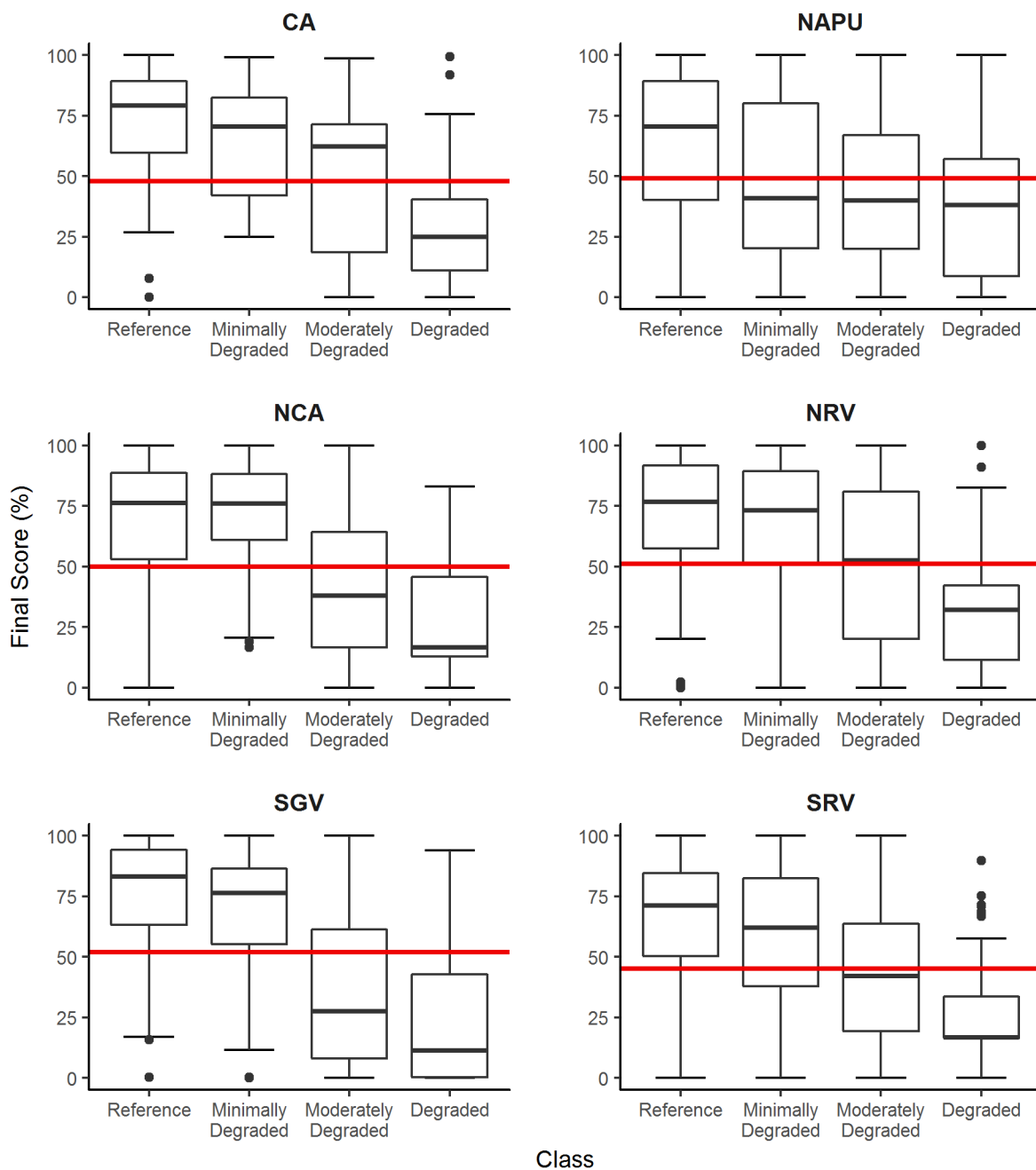


Figure 12. Distributions of index scores for family-level indices in six bioregions: Central Appalachians (CA), Northern Appalachian Plateau and Uplands (NAPU), Northern Central Appalachians (NCA), Northern Ridge and Valley (NRV), Southern Great Valley (SGV), and Southern Ridge and Valley (SRV). The 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).

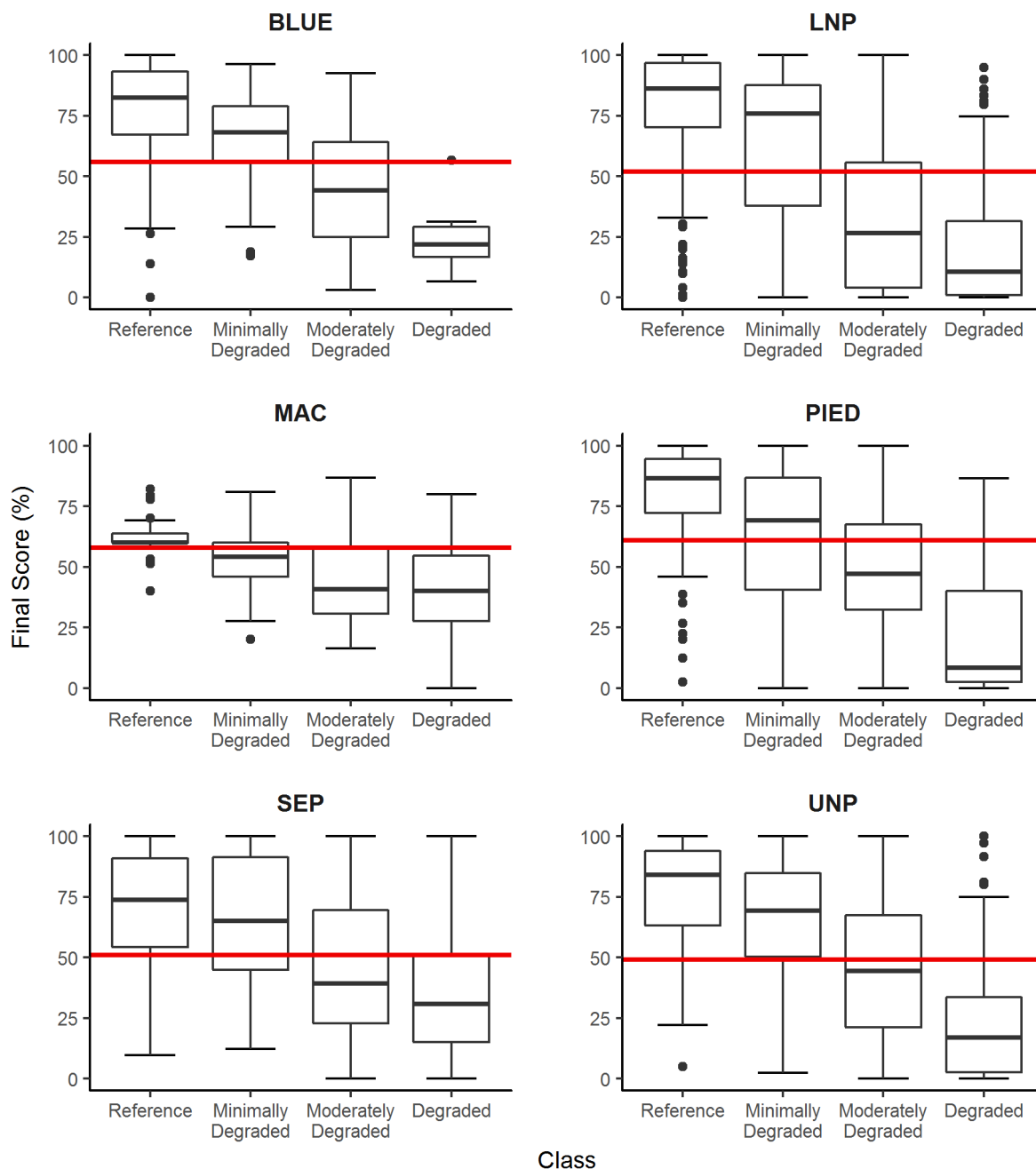


Figure 13. Distributions of index scores for family-level indices in six bioregions: Blue Ridge Mountains (BLUE), Lower Northern Piedmont (LNP), Middle-Atlantic Coast (MAC), Piedmont (PIED), Southeastern Plains (SEP), and Upper Northern Piedmont (UNP). The whisker lengths are designated by the Interquartile 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).

Table 19. 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	95 (29.9%)	22 (6.9%)	163 (51.3%)	32 (10.1%)	6 (1.9%)	318
CA	36 (11.9%)	15 (5.0%)	163 (54.0%)	65 (21.5%)	23 (7.6%)	302
LNP	215 (6.7%)	63 (2.0%)	1,606 (49.8%)	1,201 (37.3%)	139 (4.3%)	3,224
MAC	21 (1.4%)	32 (2.1%)	1,026 (68.1%)	216 (14.3%)	212 (14.1%)	1,507
NAPU	86 (7.7%)	51 (4.6%)	732 (65.9%)	211 (19.0%)	31 (2.8%)	1,111
NCA	318 (38.4%)	27 (3.3%)	368 (44.4%)	78 (9.4%)	37 (4.5%)	828
NRV	152 (16.0%)	31 (3.3%)	552 (58.2%)	172 (18.1%)	42 (4.4%)	949
PIED	109 (9.5%)	46 (4.0%)	637 (55.5%)	334 (29.1%)	22 (1.9%)	1,148
SEP	51 (1.6%)	64 (2.0%)	2,302 (70.2%)	633 (19.3%)	227 (6.9%)	3,277
SGV	165 (11.1%)	53 (3.6%)	577 (38.9%)	601 (40.5%)	88 (5.9%)	1,484
SRV	389 (20.3%)	143 (7.5%)	1,003 (52.5%)	334 (17.5%)	43 (2.2%)	1,912
UNP	82 (2.9%)	65 (2.3%)	1,814 (65.2%)	672 (24.2%)	148 (5.3%)	2,781
Total Count	1,719	612	10,943	4,549	1,018	18,841

Table 20. Metrics included in the bioregion-specific, genus-level indices.

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
BLUE	PCT_EPT_HYDROPSYCHE	Composition	Decrease	79.69	42.31	83.3
BLUE	MENHINICKS	Richness/Diversity	Decrease	1.93	0.84	83.3
BLUE	RICH_CLING	Richness/Diversity	Decrease	13.00	2.69	90.5
BLUE	PCT_CLING	Habit	Decrease	70.77	51.47	67.0
BLUE	HBI	Tolerance	Increase	2.83	3.63	67.0
CA	PCT_POTEC	Composition	Decrease	77.33	63.55	65.9
CA	PCT_EPT_RICH	Richness/Diversity	Decrease	67.88	34.08	83.0
CA	RICH_FILTER	Richness/Diversity	Decrease	3.60	0.72	74.6
CA	RICH_TRICHOPTERA	Richness/Diversity	Decrease	4.00	1.50	73.7
CA	PCT_CLING	Habit	Decrease	46.35	19.59	73.1
CA	PCT_DOM2	Tolerance	Increase	46.32	60.74	74.5
LNP	PCT_CHIRONOMINAE	Composition	Increase	0.00	0.20	78.0
LNP	GOLD	Composition	Decrease	89.90	37.76	83.2
LNP	PCT_COTE	Composition	Decrease	68.37	10.43	83.6
LNP	PCT_COLLECT	FFG	Increase	62.24	79.36	68.6
LNP	PCT_CLING	Habit	Decrease	66.20	22.78	80.6
LNP	PCT_INTOL_0_4	Tolerance	Decrease	68.57	7.53	86.2
MAC	PCT_CHIRONOMINAE	Composition	Increase	0.00	1.58	77.5
MAC	PCT_CAECIDOTEA	Composition	Increase	0.00	0.72	75.6
MAC	RICH_EPT	Richness/Diversity	Increase	0.00	1.60	68.5
MAC	RICH_INTOL	Richness/Diversity	Increase	0.00	1.60	69.7
MAC	PCT_BURROW	Habit	Decrease	55.65	22.69	75.4
MAC	PCT_DOM2	Tolerance	Decrease	90.34	50.00	71.8
NAPU	PCT_EPT_HYDRO_BAETID	Composition	Decrease	70.38	55.02	71.0
NAPU	PCT_EPT_RICH	Richness/Diversity	Decrease	69.34	42.87	74.2
NAPU	RICH_CLING	Richness/Diversity	Decrease	11.15	9.57	67.9
NAPU	PCT_COLLECT	FFG	Increase	54.97	66.45	61.5
NAPU	PCT_TOLERANT_5_10	Tolerance	Increase	40.09	53.69	71.0
NCA	PCT_COTE	Composition	Decrease	64.19	8.01	86.6
NCA	PCT_EPHEMEROPTERA	Composition	Decrease	40.90	0.00	87.1
NCA	RICH_EPHEMEROPTERA	Richness/Diversity	Decrease	7.00	1.19	88.2
NCA	RICH_SCRAPE	Richness/Diversity	Decrease	5.00	0.62	87.4
NCA	PCT_CLING	Habit	Decrease	62.87	36.99	78.7
NCA	PCT_DOM4	Tolerance	Increase	55.07	75.33	80.0
NRV	RICH_EPHEM_EPEORUS	Composition	Decrease	5.00	0.26	88.0
NRV	PCT_EPT_RICH	Richness/Diversity	Decrease	68.44	36.05	86.9
NRV	RICH_COLLECT	Richness/Diversity	Decrease	8.00	3.13	84.6
NRV	RICH_TRICHOPTERA	Richness/Diversity	Decrease	4.00	1.32	82.7

DRAFT REPORT

Bioregion	Metric	Metric Class	Influence of Disturbance	Reference Median	Bound	Metric BDE
NRV	PCT_CLING	Habit	Decrease	59.26	33.86	66.9
NRV	PCT_DOM1	Tolerance	Increase	28.18	56.62	77.8
PIED	PCT_EPT_CHEUMATOPSYCHE	Composition	Decrease	60.20	15.44	82.2
PIED	PCT_EPT_NO_HYDRO	Composition	Decrease	89.25	67.55	73.1
PIED	PCT_COTE	Composition	Decrease	61.36	24.42	85.4
PIED	PCT_EPHEMEROPTERA	Composition	Decrease	33.94	7.66	81.7
PIED	PCT_EPT_HYDRO_BAETID	Composition	Decrease	81.19	52.97	73.1
PIED	PCT_COLLECT	FFG	Increase	56.72	87.08	76.7
PIED	PCT_CLING	Habit	Decrease	62.96	34.06	76.3
PIED	PCT_MOD_TOL_4_6	Tolerance	Increase	50.41	70.43	71.2
PIED	PCT_URBAN_INTOL	Tolerance	Decrease	92.73	88.55	63.9
SEP	PCT_INSECTA	Composition	Decrease	84.62	81.40	57.1
SEP	RICH_TOL	Richness/Diversity	Increase	2.00	5.00	72.4
SEP	PCT_COLLECT	FFG	Decrease	74.75	67.89	55.9
SEP	HBI	Tolerance	Increase	5.00	5.70	70.8
SEP	PCT_TOLERANT_5_10	Tolerance	Increase	70.64	88.42	61.0
SGV	PCT_EPT_NO_HYDRO	Composition	Decrease	82.09	60.39	69.2
SGV	PCT_COTE	Composition	Decrease	64.36	40.22	75.1
SGV	PCT_EPHEMEROPTERA	Composition	Decrease	29.56	1.82	85.0
SGV	PCT_EPT_HYDRO_BAETID	Composition	Decrease	73.08	44.28	70.7
SGV	PCT_EPT_HYDROPSYCHE	Composition	Decrease	61.74	17.56	83.0
SGV	PCT_HEXAPODA	Composition	Decrease	95.68	88.04	70.1
SGV	PCT_COLLECT	FFG	Increase	66.67	80.39	64.8
SGV	PCT_CLING	Habit	Decrease	59.50	40.22	69.2
SGV	HBI	Tolerance	Increase	4.06	5.34	85.6
SRV	PCT_COTE	Composition	Decrease	58.47	47.13	65.1
SRV	MENHINICKS	Richness/Diversity	Decrease	1.78	1.22	64.8
SRV	RICH_CLIMB	Richness/Diversity	Decrease	1.00	0.57	65.1
SRV	PCT_CLING	Habit	Decrease	56.25	47.11	60.7
SRV	HBI	Tolerance	Increase	3.53	4.49	71.5
UNP	GOLD	Composition	Decrease	73.35	24.33	76.6
UNP	PCT_EPT	Composition	Decrease	57.21	1.09	81.4
UNP	PCT_HEXAPODA	Composition	Decrease	99.49	95.15	81.7
UNP	PCT_CLING	Habit	Decrease	61.58	25.04	74.7
UNP	PCT_INTOL_0_4	Tolerance	Decrease	57.25	14.45	80.5
UNP	PCT_URBAN_INTOL	Tolerance	Decrease	98.19	91.31	75.3
Mean						75.2

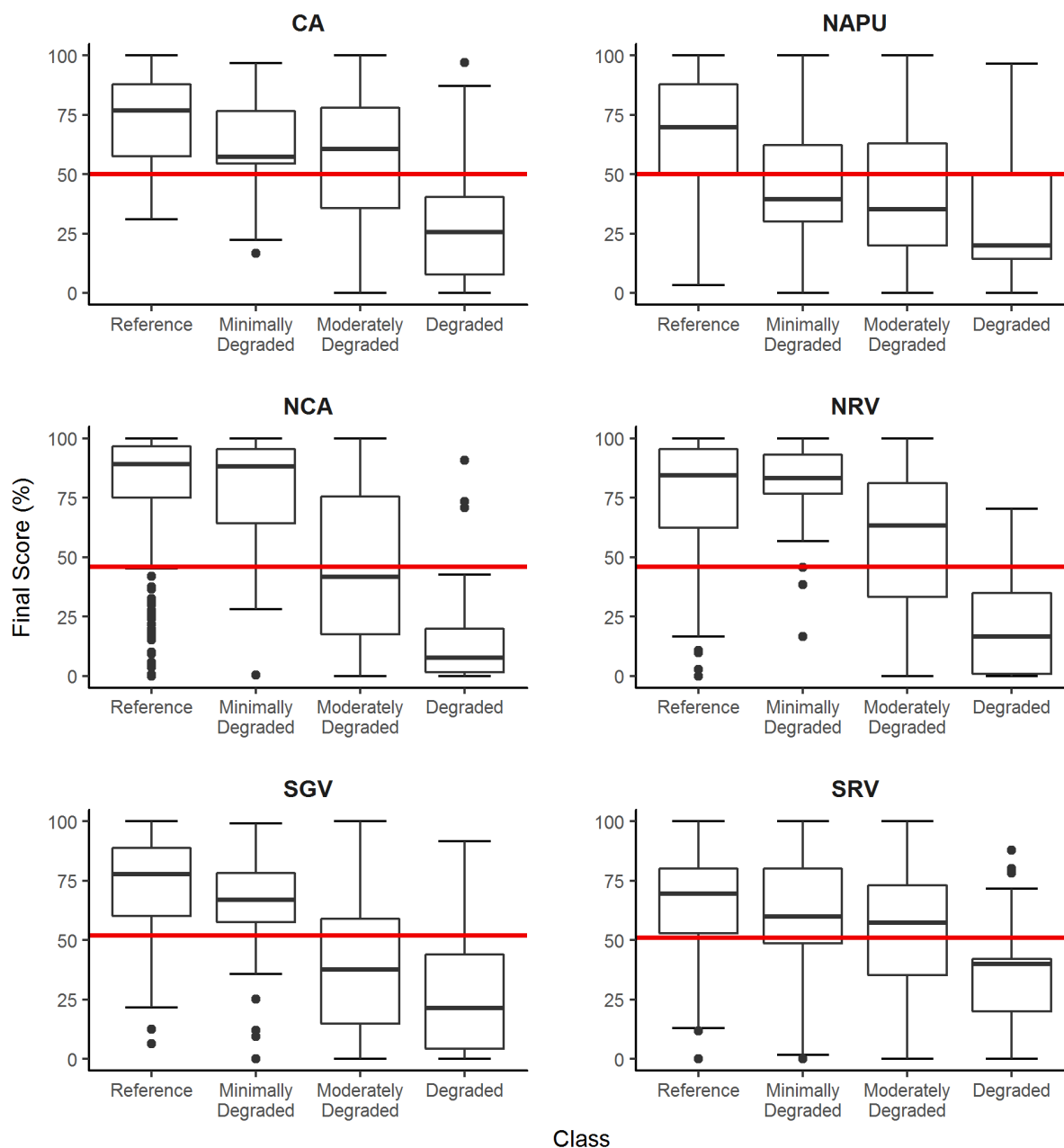


Figure 14. Distributions of index scores for genus-level indices in six bioregions: Central Appalachians (CA), Northern Appalachian Plateau and Uplands (NAPU), Northern Central Appalachians (NCA), Northern Ridge and Valley (NRV), Southern Great Valley (SGV), and Southern Ridge and Valley (SRV). The whisker 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).

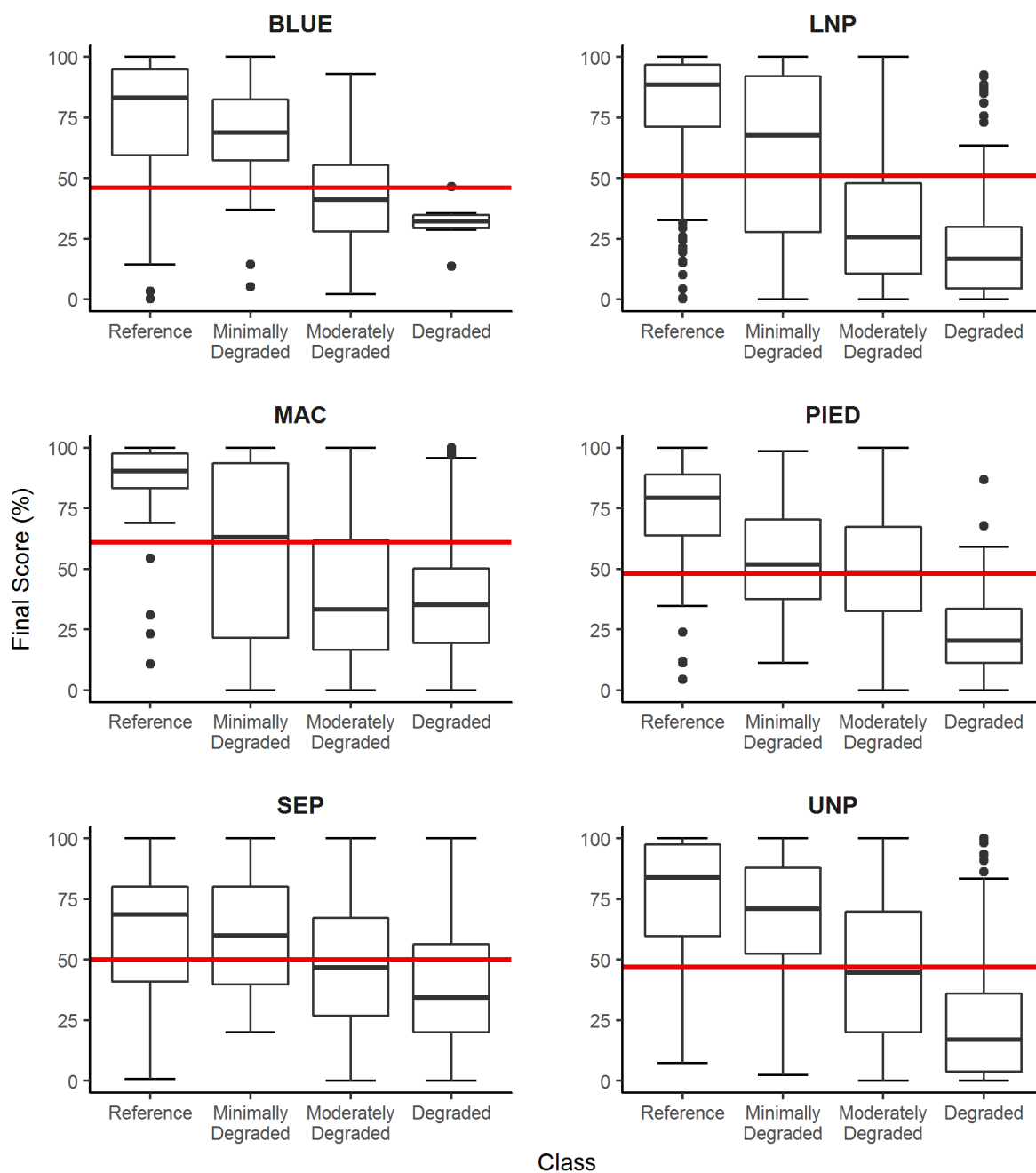


Figure 15. Distributions of index scores for genus-level indices in six bioregions: Blue Ridge Mountains (BLUE), Lower Northern Piedmont (LNP), Middle-Atlantic Coast (MAC), Piedmont (PIED), Southeastern Plains (SEP), and Upper Northern Piedmont (UNP). The whisker lengths are designated by the Interquartile 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 performed best at the family-level. At these spatial scales the family-level indices had high CEs and low variability in index accuracy (delete-d jackknife Cross Validation RMSE). Except for the Coast index, the family-level indices also had low variability associated with index precision (delete-d jackknife RMSE). Order-level Coast index precision (RMSE = 2.4) was lower than family-level precision (RMSE = 2.7).

In general, bioregion index performance improved as taxonomic resolution increased from order- to genus-level. However, the average CE and accuracy RMSE values depicted larger improvements from the order-level to the family-level, than from the family-level to the genus-level (Table K-4). Half of the bioregions (i.e., BLUE, LNP, MAC, SEP, SRV, and UNP) had higher CEs at the family-level as compared to the genus-level. The average RMSE, associated with index precision (delete-d jackknife RMSE), was smaller at the genus level, although, eight of the bioregions (i.e., CA, LNP, NAPU, PIED, SEP, SGV, SRV, and UNP) had smaller RMSE values at the family-level. Similarly, average index accuracy RMSE (delete-d jackknife cross validation RMSE) was lower at the genus-level but five of the twelve bioregions (i.e., CA, LNP, NAPU, PIED, SEP) had lower RMSE values at the family-level. 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 several changes in methodology affected ratings of individual samples. Buchanan et al. (2011) worked with 487 Reference sampling events located in five inland bioregions compared to the 1,936 Reference sampling events in ten inland bioregions used 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 differed slightly. Metrics in the current study were scored on a slightly 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 with the Reference 50th, 25th, and 10th percentiles, as well as, a value representing half of the 10th Reference percentile for each individual bioregion and region.

Using sample event ID, the 2011 family-level index ratings were paired with the corresponding bioregion family-level index ratings produced in this study. Table 21 shows the inland and coastal results separately. For the ten inland bioregions, 82.3% of the ratings matched exactly or differed by only one rating. For the two coastal bioregions, 83.7% matched exactly or differed by only one rating. Just 17.7% of the inland and 16.3% of the coastal 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 coastal ratings disagreed, this study's ratings tended to be somewhat poorer.

Table 21. 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.

		This Study's Family-Level Bioregion Ratings				
2011 Family-Level Ratings	Coast	Excellent	Good	Fair	Poor	Very Poor
	Excellent	14.2%	1.6%	0.1%	1.8%	0.1%
	Good	5.2%	5.5%	2.4%	2.9%	0.9%
	Fair	2.1%	3.1%	6.8%	8.2%	3.2%
	Poor	1.5%	0.4%	2.8%	8.1%	3.8%
	Very Poor	2.0%	0.0%	1.1%	10.4%	11.6%
	Inland					
	Excellent	10.6%	4.3%	1.4%	0.5%	0.1%
	Good	3.7%	4.0%	2.3%	1.2%	0.2%
	Fair	2.5%	4.3%	4.1%	2.5%	0.9%
	Poor	0.7%	2.4%	3.5%	4.0%	2.2%
	Very Poor	0.6%	2.7%	4.6%	11.0%	25.8%

There are connections between stream biota and watershed land use through stream habitat and water quality conditions. Many stream habitat and water quality parameters used to classify Reference conditions correlate positively with %Forest and negatively with %Urban (Table L-4). We compared the narrative ratings of the two spatial and three taxonomic versions of the index to six land use categories, to see if ratings of the different versions corresponded to increasing anthropogenic impacts in the watershed. Watershed land use characteristics were grouped into six categories by the percentages of forest, urban, and agricultural area in the HUC12 watersheds. 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 22 shows the percentages of Excellent, Good, and Fair (EGF) and Poor and Very Poor (PVP) ratings in each land use category. Results for the bioregion and regional indices are grouped by region for brevity in Table 22. Table L-5 shows results for the bioregion indices grouped by bioregion.

All spatial and taxonomic versions of the index were able to identify the highest quality watersheds. Large percentages of EGF were always associated with the least disturbed or minimally disturbed HUC12 watersheds. The family-level versions of the bioregion and region indices also identified the poorest quality watersheds well, with large percentages of PVP associated with the disturbed and most disturbed HUC12 watersheds. The genus-level bioregion and region indices could identify the poorest quality watersheds in the inland region but not the coastal region. The order-level bioregion indices could not identify the poorest quality watersheds but the order-level region indices could. Overall, the family-level indices were the

Table 22. Correspondence of narrative ratings for the two spatial and three taxonomic versions of the index with different intensities of disturbance in HUC12 watersheds. The bioregion index results are grouped by region to illustrate the regional differences. Large percentages of Excellent, Good, and Fair ratings (EGF) are expected in least disturbed watersheds; large percentages of Poor and Very Poor ratings (PVP) are expected in heavily disturbed watersheds.

Taxon. Version	Region	IBI Rating	Least Disturbed			Most Disturbed		
			>78%Forest <5%Urban	50%-78% Forest <5%Urban	Other	<25%Forest >50%Agric.	<25%Forest >20%Urban	<25%Forest >20%Urban >50%Agric.
Bioregional Indices	Genus	Coast	EGF	88.6%	74.2%	44.0%	56.2%	
			PVP	11.4%	25.8%	56.0%	43.8%	
		Inland	EGF	78.5%	74.4%	31.8%	13.3%	21.9%
			PVP	21.5%	25.6%	68.2%	86.7%	78.1%
	Family	Coast	EGF		83.7%	61.4%	23.6%	24.1%
			PVP		16.3%	38.6%	76.4%	75.9%
		Inland	EGF	83.3%	69.6%	53.1%	31.8%	11.9%
			PVP	16.7%	30.4%	46.9%	68.2%	88.1%
	Order	Coast	EGF		86.6%	79.6%	100.0%	61.2%
			PVP		13.4%	20.4%	0.0%	38.8%
		Inland	EGF	82.6%	77.6%	68.6%	64.4%	38.6%
			PVP	17.4%	22.4%	31.4%	35.6%	61.4%
Regional Indices	Genus	Coast	EGF		89.8%	70.9%	57.3%	40.8%
			PVP		10.2%	29.1%	42.7%	59.2%
		Inland	EGF	89.1%	77.6%	54.1%	37.6%	17.2%
			PVP	10.9%	22.4%	45.9%	62.4%	82.8%
	Family	Coast	EGF		84.3%	69.3%	57.5%	28.1%
			PVP		15.7%	30.7%	42.5%	71.9%
		Inland	EGF	83.9%	70.4%	45.0%	23.9%	8.5%
			PVP	16.1%	29.6%	55.0%	76.1%	91.5%
	Order	Coast	EGF		74.4%	47.4%	48.5%	19.7%
			PVP		25.6%	52.6%	51.5%	80.3%
		Inland	EGF	85.2%	80.4%	60.3%	45.0%	25.0%
			PVP	14.8%	19.6%	39.7%	55.0%	75.0%

most reliable indicators of watershed disturbance across the entire range of disturbance found in Chesapeake watersheds.

We cross-compared the narrative ratings produced for the two spatial and three taxonomic versions of the Chessie BIBI index in this study. This allowed us to determine which index versions could be used together. Ratings were paired on sample event ID. Percentages of the paired results that exactly match, nearly match (i.e., are adjacent ratings), and disagree (i.e., separated by one or more ratings) are shown in Table 23. Comparisons between the regional

Table 23. Cross-comparisons of ratings of the two spatial (region, bioregion) and three taxonomic (genus-, family-, order-) versions of the index. Agreement: match, exact agreement; near, differ by one rating (e.g., Excellent and Good); disagree, differ by more than one rating (e.g., Excellent and Fair). Highlighted cross-comparisons have greater than 80% matching or near matching ratings and less than 20% disagreeing ratings. Results are grouped by coastal 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 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 compared. D) Ratings produced by the family-level versions of the regional and bioregion indices are compared.

A) Region vs Bioregion Indices

Bioregion		Agreement	Region						
			Genus		Family		Order		
			Coast	Inland	Coast	Inland	Coast	Inland	
Genus	match		32.7%	49.2%	28.5%	49.4%	23.6%	36.5%	
	near		44.3%	40.4%	39.7%	37.4%	28.8%	41.9%	
	disagree		23.0%	10.4%	31.9%	13.2%	47.6%	21.7%	
	Family	match		25.0%	40.9%	49.8%	51.9%	41.6%	36.5%
		near		34.0%	33.5%	38.3%	36.9%	34.4%	43.8%
		disagree		32.8%	11.6%	11.9%	11.2%	24.1%	19.7%
	Order	match		22.6%	37.0%	31.7%	39.4%	36.5%	39.7%
		near		31.9%	37.0%	34.0%	36.0%	24.9%	42.3%
		disagree		45.5%	25.9%	34.3%	24.6%	38.6%	18.0%

B) Taxonomic Versions of the Region Indices

	<u>Agreement</u>	Genus		Family	
		Coast	Inland	Coast	Inland
Family	match	29.4%	52.3%		
	near	43.6%	38.1%		
	disagree	27.0%	9.5%		
Order	match	24.3%	40.2%	39.7%	37.8%
	near	34.4%	41.8%	37.9%	42.7%
	disagree	41.3%	18.0%	22.4%	19.5%

C) Taxonomic Versions of the Bioregion Indices

	<u>Agreement</u>	Genus		Family	
		Coast	Inland	Coast	Inland
Family	match	32.9%	54.6%		
	near	33.7%	35.5%		
	disagree	33.4%	9.9%		
Order	match	36.6%	39.1%	36.4%	39.5%
	near	29.7%	35.9%	32.6%	36.9%
	disagree	33.8%	25.0%	31.1%	23.7%

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

	CA	NAPU	NCA	NRV	UNP	BLUE	LNP	PIED	SGV	SRV	SEP	MAC
match	32.8%	40.4%	52.6%	50.8%	46.3%	47.5%	64.4%	56.3%	57.4%	44.9%	60.5%	25.4%
near	47.0%	38.9%	38.4%	35.2%	40.2%	40.4%	31.6%	34.6%	35.7%	39.5%	38.7%	37.2%
disagree	20.2%	20.6%	9.0%	14.0%	13.5%	12.1%	3.9%	9.1%	6.8%	15.6%	0.7%	37.4%

and bioregional indices (Table 23A) show the family-level versions have the least overall disagreement, with 11.9% in the coastal region and 11.2% in the inland region. Roughly half of the ratings matched exactly. The genus-level versions of the regional and bioregional indices agree well in the inland region (49.2% match, 10.4% disagree) but not the coastal region (32.7% match, 23% disagree). The order-level versions likewise agree in the inland but not the coastal region. Region and bioregion index ratings produced by the different taxonomic versions can sometimes show good agreement. For example, ratings produced by the genus-level region and family-level bioregion versions of the index compare well in the inland region (40.9% match, 11.6% disagree). Somewhat similar are the ratings of the family-level bioregion and order-level region versions of the index in the inland region (36.5% match, 19.7% disagree). Ratings of the genus- and order-level versions of the index always disagree to a large extent, particularly in the coastal region. For example, in the coastal region ratings for the order-level region index and genus-level bioregion indices disagreed 47.6% and ratings for the genus-level region index and order-level bioregion indices disagreed 45.5%. Closer examination of the results shows most of the disagreements occur in sampling events in the MAC bioregion.

Comparisons of the genus-, family-, and order-level versions of the regional indices show agreement between all three in the inland region but not the coastal region (Table 23B). The genus- and family-level Inland indices show the strongest agreement (52.3% match, 9.5% disagree). Comparisons of the taxonomic versions of the bioregional indices show agreement only between the genus- and family-level versions in the inland region (Table 23C).

Comparisons of the family-level versions of the regional and bioregion indices showed strong agreement in five bioregions (NCA, LNP, PIED, SGV, SEP), good agreement in four bioregions (NRV, UNP, BLUE, SRV), weak agreement in two CA and NAPU, and poor agreement in MAC. These results suggest family-level versions of the regional and bioregion indices can be used interchangeably in at least nine of the bioregions.

G. Chesapeake Watershed Stream Health

The area-weighted station ratings were mapped by HUC12 units to visually examine and evaluate the unbiased spatial distributions of the rating results of the various indices (Appendix L). The Chesapeake-wide indices indiscriminately rated the entire coastal region and broad swaths of urban and suburban lands as Poor or Very Poor despite a substantial number of Reference sites and several relatively undisturbed HUC12 units 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 coastal region that were consistent with the rating disagreements between and within the region and bioregion indices (Table 23). The family-level version of the Coast index classified the MAC bioregion as Fair - Poor (Figure L-7); the genus-level version of the Coast index classified MAC as mostly Good - Fair (Figure L-8); and the family- and genus-level versions of the bioregion indices rated the MAC bioregion as mostly Poor (Figure L-10 and L-11). The results are complicated by the fact that a range of index scores for Good ratings could not be established for the family-level MAC bioregion index (Table L-3). Rating discrepancies are also apparent in the CA and NAPU bioregions, but are not as striking as

in MAC. For the family-level versions, the Inland index yielded about 10% more ratings of Excellent, Good, and Fair than the CA and NAPU bioregion indices.

Area-weighting individual station 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 rolled up to estimate the percentages of each in the selected spatial scale. Figure 16 shows the family-level Chesapeake-wide index rolled up to basin, the family-level Coast and Inland indices rolled up to region, and the twelve family-level bioregion indices rolled up to bioregion. The region and bioregion results can be further rolled up to basin to compare results of the three index types on the whole Chesapeake watershed (Figure 17).

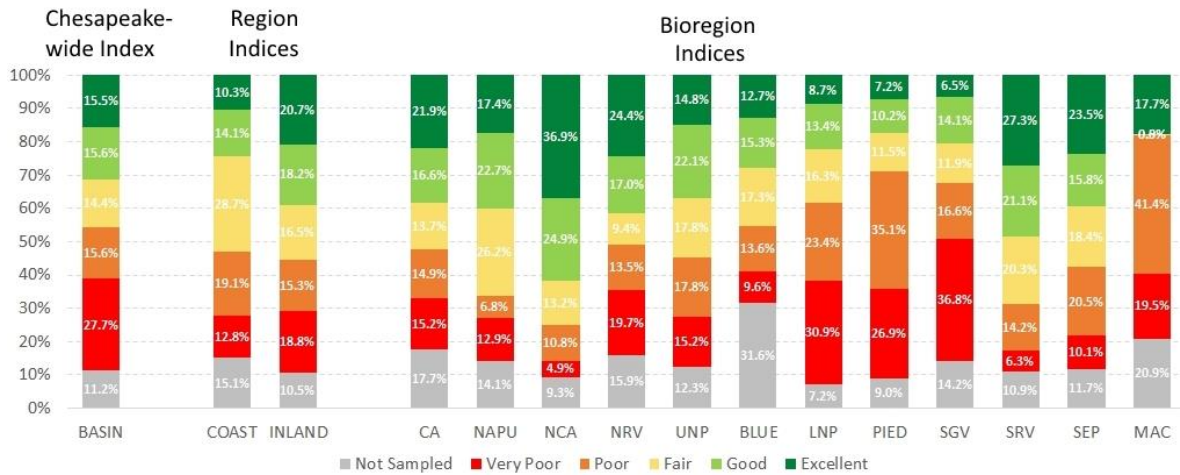


Figure 16. Area-weighted percentages of the five ratings derived with the family-level versions of the Chesapeake-wide index, the two Region indices, and the twelve Bioregion indices.

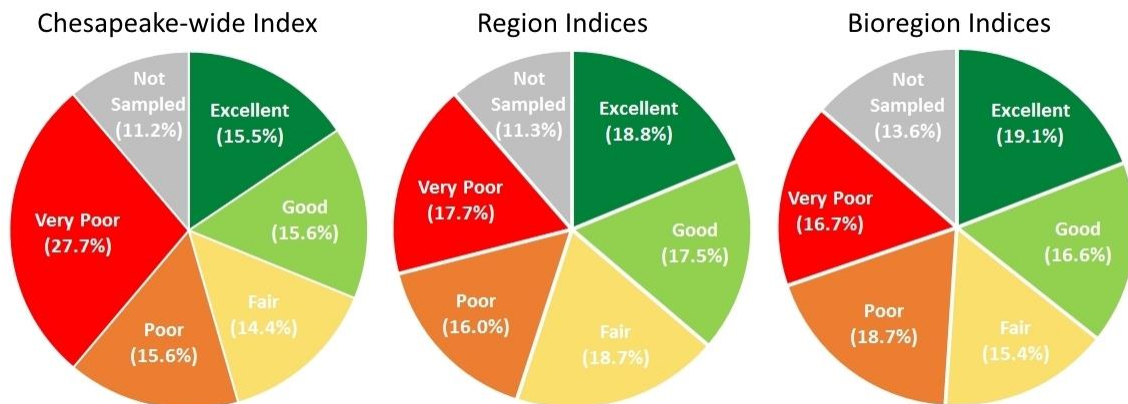


Figure 17. Area-weighted percentages of the five ratings derived with the family-level versions of the Chesapeake-wide, Region, and Bioregion indices and rolled up to the Chesapeake watershed scale.

The 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 or Poor. Results of the region and bioregion indices are more comparable, with 33.7% and 35.4% scoring Very Poor or Poor, respectively.

IV. Discussion

The objective of this study was to refine the Chessie BIBI, a stream benthic macroinvertebrate index of biotic integrity that evaluates stream biological communities on a common scale across the Chesapeake 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. Each version of the index is tailored to a specific spatial scale and taxonomic level, and is composed of the stream macroinvertebrate metrics determined to be the most sensitive to disturbance at that scale and level.

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 watershed's natural features or for the wide-spread anthropogenic degradation in the coastal region. Natural differences in the hydrologic and topographic features of the coastal and inland regions correspond to strong natural differences in stream macroinvertebrate assemblages (Appendix H and G). 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 coastal Reference sampling events in our analysis dataset were overshadowed by the more numerous inland Reference sampling events, effectively turning coastal sites into outliers during index development. 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 and Southeastern Plain 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 Coast and Inland region indices had CEs greater than 70.0%. Narrative ratings were developed independently for the indices using the 50th, 25th, and 10th percentiles of the index scores from each region's Reference sites (Tables L-2). To varying degrees, the ratings for all three taxonomic versions of the regional indices correspond to the range of landscape disturbance experienced in the Chesapeake watershed (Table 22). Ratings produced by the family-level indices correspond best to watershed disturbance. The Coast and Inland regional ratings can be used together to represent stream biological integrity across the entire Chesapeake watershed as long as taxonomic version is the same, e.g., the 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 genus- and family-

level versions of the Inland index suggest they could be used interchangeably in inland areas (Table 23B).

Further dividing the Chesapeake Bay basin into twelve bioregions provides more spatial resolution of the basin's complex natural features and accommodates some well recognized spatial differences in natural stream macroinvertebrate assemblages. The greater spatial resolution helps to avoid situations experienced in Buchanan et al. (2011) where high numbers of Reference and Degraded samples in one part of a bioregion dominated index development for the 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. 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. These two bioregions were divided into five 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), and the Southern Great Valley (SGV). 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.

Bioregion differences in the distributions of Reference index scores were large enough to discourage the use of a single rating system (Table L-3). Therefore, the 50th, 25th, and 10th percentiles of each bioregion's distribution of Reference index scores were used independently to establish ratings for the bioregion. These ratings can be used together to represent stream biological integrity across the entire Chesapeake watershed.

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. Both of the order-level region indices were sensitive to stream conditions ($CE \geq 70.0\%$). Only six of the twelve order-level bioregion indices were considered sensitive (i.e., BLUE, CA, PIED, SGV, SRV, and UNP) and most had higher amounts of uncertainty as measured by RMSE (Appendix K). The order-level versions of the regional indices were sensitive to watershed disturbance while order-level versions of the bioregion indices were not responsive (Table 22). Index development sometimes identified fewer than five metrics for inclusion in the order-level bioregion indices. For example, only one metric (PER_DOM4) met the strict selection criteria for inclusion in the order-level MAC index (Table 16), which partly explains the bioregion's particularly low CE (61.9%), greater uncertainty, and almost binary distribution of index scores in Reference conditions (Table L-3, Figure 11). Just two metrics were included in the order-level bioregion index for SEP, the other coastal bioregion. We recommend caution in using the order-level versions of both the region and

bioregion indices. However, we believe they eventually can 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 do not increase CEs substantially over the family-level versions in a majority of bioregions, and were actually 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. 2008) 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 (2008) 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.

We recommend as a general practice that the family-level index versions be used for assessments at the region and bioregion spatial scales. With the exception of MAC, ratings produced by the family-level regional indices correspond with those produced by the family-level bioregion indices (Table 23D). When sites are being assessed individually (e.g., to measure restoration “lift”), the genus-level versions of bioregion indices also may be useful if they are substantially more sensitive than their corresponding family-level versions. These include NAPU, NCA, and NRV (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.

We suggest caution when using the MAC and CA bioregion indices. These two regions have the lowest number of Reference samples and consequently have high RMSE. Validation estimates in Appendix K indicate the family-level MAC and CA indices are overestimating CE (delete-d jackknife cross validation procedure). The genus-level version of the CA index may be a reliable substitute for the family-level version. However, more Reference samples are needed in MAC before we can definitively rate this bioregion at either taxonomic level.

C. Distributions of Index Scores

Distributions of the bioregion index scores in Reference and Degraded conditions 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). Although water quality and stream habitat criteria for Reference or Degraded are all met, other environmental factors influence stream environments (Appendix H) and can cause percentiles of the Reference distributions to differ. These differences affect the narrative rating thresholds because the 50th, 25th, 10th, and half of 10th percentiles of each bioregion's Reference distributions are used as the rating thresholds. A high index score can be rated "Excellent" in one bioregion and something else in another, or a low index score can be rated "Very Poor" in one bioregion and something else in another. There is also less certainty and more variability in the values of the Reference percentiles when bioregion sample sizes are low (e.g., MAC and CA). The metrics selected for each bioregion index are those with the strongest and clearest biological responses to anthropogenic impacts in that bioregion's streams. The fact that distributions of Reference and Degraded scores vary so much indicates the standardized water quality and stream habitat criteria do not account for all of the local factors influencing macroinvertebrate assemblages.

Unlike the bioregion indices, Reference and Degraded distributions of index scores for the two region indices are very similar at the family-level (Figure 8) and somewhat alike at the genus-level (Figure 9). Bioregion differences have been subsumed and the metrics selected for the indices reflect the strongest and clearest biological responses for the entire region.

D. Narrative Ratings

Distributions of a 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. The largest differences occur between the Coast and Inland regions (Appendix G). 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 according to their bioregion's Reference and Degraded assemblages, bioregion differences carry through to the index scores. Consequently, an index score for one bioregion index is not necessarily equivalent to the same index score for another bioregion index (Pond et al. 2008).

The unevenness in the distributions of Reference index scores is eased when a common rating system is applied. The 10th, 25th, and 50th percentiles of each bioregion's Reference distributions were used in this study to establish rating thresholds. Thus, an Excellent rating by one index is roughly equivalent to an Excellent rating by another because both have index scores greater than the 50th percentile of their respective Reference distributions. With the exception of MAC, ratings of the family-level versions of the region and bioregion corresponded to each other (Table 23D). 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, also were comparable. Ratings matched exactly or differed by only one rating at 82.3% of inland sites and 83.7% of coastal sites. 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.

E. Chesapeake Bay Assessments

Although region and bioregion ratings are comparable overall, significant discrepancies in some areas raise the question of which spatial resolution provides the most accurate depiction of stream condition. The most obvious discrepancies exist in the MAC bioregion. MAC indices are limited by a low number of Reference sampling events ($n = 21$), which prevents the development of robust and sensitive indices. At the family- and genus-level, the Reference percentiles and ratings categories are unevenly distributed on the 0 – 100 scoring scale. The family-level index scores quickly jump to the extreme ends of the rating scheme while the genus-level index scores are dominated by the Fair category (Figure L-2). This makes the region and bioregion maps contradict one another in the Coast region (Figure L-7 *versus* L-10; Figure L-8 *versus* L-11). Despite CEs greater than 80% in MAC, CE variability (validation RMSE) calculated with delete-d jackknife cross validation is high. The MAC index ratings line up poorly with HUC12 land cover features. All of the HUC12 watersheds represented by MAC sampling sites in the analysis dataset have some level of disturbance, and none of the 21 MAC sites that technically meet the stream habitat and water quality criteria for Reference are located in these comparatively undisturbed HUC12s. Depicting MAC as a generally degraded area may be appropriate, however we have low confidence in the performance of the MAC bioregion indices.

When the MAC and SEP bioregions are combined in the Coast region index, the number of Reference samples available for index development increases to 73 (21 MAC, 52 SEP) and rating categories for the family-level index occur at more even intervals (Figure L-1). There is more concurrence between the Coast index ratings and the amount of HUC12 disturbance (Table 22). Combining the MAC and SEP bioregions into a single regional index currently provides the most robust index for the coastal area. Until more Reference sites are sampled in the MAC region, we suggest evaluating that bioregion with the Coast region indices.

If the objective is to assess the Chesapeake Bay basin as a whole, the index must be selected considering CE, precision, accuracy, and parsimony. When multiple indices provide roughly equivalent results, it is beneficial to select the most simplistic index or set of indices. Unnecessary complexity increases the potential of introducing error in future assessments. For each spatial resolution, the family-level indices were generally comparable to the genus-level indices, 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. 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 the basin-wide assessment. The region indices standardize metrics and scoring thresholds over large areas of the basin allowing for direct comparison of index scores among most sampling events. Region assessments provide the best results for long-term, large scale monitoring.

If the objective is to assess specific areas of the Chesapeake Bay basin (e.g., restoration sites), the bioregion indices may prove more useful. Indices developed for each of the twelve bioregions allow assessments to be made on relatively homogeneous geographic areas. We

argue the majority of the bioregion indices are better able to identify macroinvertebrate responses to specific disturbances and stressors characteristic of the bioregion because they are fine-tuned to account for the influence of natural factors within the bioregion. Bioregion indices may be beneficial in monitoring localized short-term trends. Ultimately, assessments made with multiple index spatial scales and taxonomic tiers may provide the most holistic review of stream condition within the Chesapeake Bay basin.

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 by HUC12 to reduce spatial bias. In the past, CBP has focused on the basin-wide percentage of streams categorized as Excellent, Good, or Fair. Using the area weighted method, we estimate approximately 55% (Region indices) or 51% (Bioregion indices) of streams in the Chesapeake basin were in Excellent, Good, or Fair condition over the 1992 – 2015 time period. An objective of the CBP is to develop a 2008 baseline against which they 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.

V. Conclusions and Recommendations

The Chessie BIBI is an ever evolving index. Admittedly, the data within the Chessie BIBI database is prone to bias due to differences in sampling technique, taxonomic resolution, habitat measures, and water quality measures among agencies/programs within the database. Several steps were taken to extensively groom the database and reduce the influence of any existing bias. Accumulating multiple data sources increases samples size, and therefore, statistical power. Additionally, the data set transcends geopolitical borders, allowing for contiguous analysis within regions deemed environmentally similar (i.e., regions and bioregions). All benthic macroinvertebrate assessments are 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).

One of the major difficulties in refining the Chessie BIBI was due to discrepancies in the habitat and water quality variables and not to differences associated with macroinvertebrate collection/identification methods. The most important aspect of IBI development is the stream condition classification procedure. The environmental factors used to classify samples as Reference and Degraded influence the metrics selected in the final index. For example, site classification based strictly on nutrient criteria and site classification based on habitat variables would likely produce two different IBIs despite the fact that the same data set was 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, and their data does not need to be compatible with other data sets. However, standardizing procedures makes it easier for analysts to collaborate, review, verify previous conclusions, and rapidly

improve the science. Collecting at a minimum the habitat parameters outlined in the Rapid Bioassessment protocol (EPA 1999) 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 nutrients 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.

Multiple bioregions (i.e., BLUE, CA, MAC, and PIED) were lacking adequate Reference and/or Degraded sample sizes ($n < 50$). Also, bioregions, such as NAPU, had low CE. Targeting specific streams expected to be of Reference or Degraded quality within each of these bioregions would most likely improve index performance in future efforts. A review of land use and USGS gauges or other water monitoring data will help to identify streams that meet these conditions. Spatial distribution of sample locations could also be factored into the 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 because Reference and Degraded samples would be more evenly distributed throughout the basin. 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.

Quantifying land use in the watershed above the sampling station may also improve stream condition classification. The HUC12 average land use variables in Appendix H provide an estimate of the land use within the watershed above the sampling station. However, we recommend watersheds delineated from the sampling station will provide values appropriate for stream condition classification. Applying the HUC12 estimates in place of the delineated watershed can introduce error. HUC12s represent sub-watersheds that when aggregated together represent larger hydrological delineations, such as watersheds and basins. Currently the Chessie BIBI database only allows the assignment of land use values from the HUC12 in which the sampling station is located. Therefore, large amounts of land use information in the watershed above the sampling station may be lost because two or more HUC12s would need to be aggregated together. Additionally, if the sampling station is located in the upper margin of the HUC12, a large amount of land use values downstream of the sampling station will be inappropriately used to classify stream condition. Delineating watersheds from the sampling location in ArcGIS will prevent these issues but the process will require extensive QA/QC. The most difficult aspect of the process would be to accurately place the sampling station on the appropriate stream. Due to low accuracy GPS equipment, a significant portion of the stations in the Chessie BIBI database do not fall directly on the stream layer in ArcGIS. ArcGIS tools can be used to snap the points to the stream layers 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 of the sampling stations in the database.

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 2016b). 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 2016b). 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).

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. Further efforts should be made to continue to update the current set of attributes/traits and to include additional attributes/traits from additional sources.

The family-level regional indices are recommended for the basin-wide assessment of the Chesapeake Bay basin. Regional indices represent large geographic areas within the basin. They provided high CE, reduced complexity, lower metric variability, and lower variability in scoring/rating thresholds relative to the bioregion indices. Area weighting the index scores prior to rating will reduce the spatial bias associated with heavily sampled geographic areas, enabling an accurate summary of the basin to be derived. The Chesapeake Bay Program (CBP) is seeking a 2008 baseline established by the Chessie BIBI scores and ratings to monitor restoration progress within the basin. Establishing the baseline will require additional attention and direction from CBP. To encompass spatial and temporal variability, the baseline will require data from multiple years. Data collected between 2000-2011 are the most prospective candidates for establishing the baseline because the majority of the Chessie BIBI data was collected during this time period.

Without a subset of repeatedly sampled stations, it will be difficult 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. Benthic macroinvertebrate samples could be collected at a predefined frequency (e.g., annually or 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* 31:670–682.
- 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. Virginia 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 12:8.
- 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. Macroinvertebrate Biological Assessment of Wadeable Streams in Georgia. Standard Operating Procedure, Georgia Department of Natural 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 macroinvertebrate community impairment to catchment characteristics 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* 31:656–669.
- 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.
- NRSA. 2008. NRSA 0809 Benthic Taxa List and Autecology - Data. Environmental Protection Agency.
- 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, AND B. LOWMAN. 2008. 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 [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 Department 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 Condition Gradient (BCG) for Fish and Benthic Macroinvertebrate Assemblages in the Northern Piedmont 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. 2012. Freshwater Biological Traits Database. United States Environmental Protection Agency.
- USEPA. 2016a. National Rivers and Streams Assessment 2008-2009: A Collaborative Survey.
- USEPA. 2016b. 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

DRAFT REPORT

Protection Agency, National Health and Environmental Effects Research Laboratory, Corvallis, Oregon. Report with map supplement, Scale 1:1–0.

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.