Creating a stream health baseline for the Chesapeake basin from monitoring and model data

Claire Buchanan[1], Kelly Maloney[2], Zachary Smith[3], Andrea Nagel[1], and John Young[2]

[1] Interstate Commission on the Potomac River Basin
[2] U. S. Geological Survey Leetown Science Center
[3] New England Interstate Water Pollution Control Commission

Abstract

This report describes how monitoring and model data were analyzed and combined to generate a preliminary estimate of acceptable stream health in the Chesapeake Bay basin for the 2006 – 2011 baseline period. Streams in about 73% of the basin's 64,020 mi^2 drainage area were evaluated with monitoring results, and output from a predictive model was used to estimate stream health in the remaining 27%. Stream health was measured with the bioregion, family-level version of the "Chessie BIBI," a multi-metric index for stream macroinvertebrate communities. Index scores are normally expressed as one of five index ratings: Excellent or Good (well-functioning), Fair (considered satisfactory), and Poor or Very Poor (stressed or poorly-functioning). Four versions of the predictive model were developed and tested, and the selected version outputs results as three-ratings: Excellent/Good, Fair, and Poor/Very Poor. The five ratings in the monitoring data were re-grouped to match the three ratings of the selected predictive model. The monitoring- and model-based ratings were then area-weighted to reduce bias caused by uneven sample densities and aggregated to the Chesapeake basin scale, with monitoring results given preference. The combined results suggest approximately 60% of the basin's *area* had acceptable stream ratings (Excellent, Good, or Fair) during 2006 – 2011. This estimate is a preliminary baseline for the Chesapeake Bay Program's stream health goal. A final baseline estimate will be produced after a higher resolution stream layer becomes available and acceptable stream health can be estimated as a percent of the basin's *stream miles*.

Introduction

This is a first attempt to develop a baseline for non-tidal stream conditions in the Chesapeake Bay drainage area, or basin. The baseline addresses a 2014 goal of the Chesapeake Bay Program (CBP) partnership which calls for improving stream health and function in ten percent of stream miles above a "2008 baseline" for the Chesapeake basin (CBP 2014). Biological communities are considered the definitive measure of stream health (Law et al. 2015), and CBP selected the Chessie BIBI, or Chesapeake Basin-wide Index of Biotic Integrity, as its indicator of stream health (CBP 2015). The index is calculated from stream macroinvertebrate data collected by federal, state, county, and volunteer monitoring programs across the Chesapeake basin. A common assessment method is applied to the data which allows comparisons of stream biological communities anywhere in the basin based on five ratings: Excellent (E), Good (G), Fair (F), Poor (P), and Very Poor (VP) (Smith et al. 2017). The 6-year period from 2006 to 2011 was selected as the baseline at a 2018 workshop (CBP 2018). This period encompasses all sampling designs of the basin's various monitoring programs. Workshop participants selected the bioregion version of the family-level Chessie BIBI index for CBP reporting purposes. Twelve distinct

bioregions are found in the Chesapeake basin (**Figure 1**) and have inherent characteristics that affect stream macroinvertebrate populations, such as topography, climate, elevation, soils, and vegetation. Bioregion-specific adjustments to the Chessie BIBI index are expected to make the index more attuned to natural landscape differences than the regional version of the index, and better able to detect improvement. For watersheds with little or no stream biological data, workshop participants recommended using output from a Random Forest model that uses landscape variables to predict Chessie BIBI ratings for individual catchments (Maloney et al. 2018). The mix of monitoring results and model predictions is not ideal, but it is currently the best solution for filling gaps in stream monitoring coverage.

Data Analysis

This report section describes some of the steps taken to generate a baseline from the combined monitoring and model results. Steps included identifying a suitable analysis scale (hydrologic unit), area-weighing index ratings to avoid spatial bias caused by non-random monitoring station distributions, quantifying uncertainty (variance) relating to sample size and the Chesapeake basin's heterogenous landscape, and determining how to fill monitoring gaps with model results. The baseline remains preliminary until the Random Forest model has been re-calibrated with a 2008 land use layer for the basin (although not much change is expected between the 2006 and 2008 land use layers), and a high-resolution stream layer becomes available. At that time, the basin's percentage of *healthy stream miles* will be calculated instead of the basin's percentage of *area with healthy streams* as reported here.

Hydrologic unit size

The basin's non-tidal stream database as of June 2017 contained 8,720 sampling events collected at 7,292 monitoring stations on 1st to 4th order streams during the 2006 – 2011 baseline period. Chessie BIBI index scores calculated from these data previously have been analyzed and mapped at a scale determined by bioregion and hydrologic unit code 12 (HUC12) (Smith et al. 2017). When bioregions are intersected with the 1971 HUC12 subwatersheds in the Chesapeake basin, the result is 2,547 HUC12-Bioregion subwatersheds averaging 25.1 mi^2 and ranging in size from <1 to 67.1 mi^2. The HUC12-Bioregion subwatershed scale facilitates research and resource management at a local level while also considering geographic differences in the natural factors influencing stream macroinvertebrates. For the baseline period, 1,472 of these HUC12-bioregion subwatersheds had monitoring data and 1,075 did not. Station densities in monitored subwatersheds ranged as high as 100 but most (88.7%) had between 1 and 10 stations.

The Random Forest model predicts Chessie BIBI ratings at the much smaller catchment scale. Catchments obtained from the Spatial Hydro-Ecological Decision System (SHEDS, www.ecosheds.org) database and less than 77.2 mi^2 (200 km^2) were used to develop the Chessie BIBI predictive model (Maloney et al. 2018). The model was developed at this scale because the smaller average sizes and higher resolution of the SHEDS catchments were considered more capable of accurately predicting Chessie BIBI ratings than the HUC12-Bioregion subwatersheds. There are over 100,000 catchments in the Chesapeake basin, averaging about 0.6 mi^2. Each is mapped at the 1:24,000 scale and has associated information about the land cover, natural characteristics, climate, dams, and soils in 2006. Model results were generated for 95,877 catchments or about 91.9% of the basin's area.

To calculate the Chesapeake basin's stream health baseline, index ratings from the monitored HUC12-bioregion subwatersheds and the modeled catchments were separately grouped by HUC12. HUC12 is one of the smaller U. S. Geological Survey's hydrologic units, and averages 32.5 mi^2 in the Chesapeake basin (range, <1 to 72.7 mi^2). In many cases, HUC12's are true watersheds, or land areas with interconnected streams, small rivers and other waterbodies draining to a defined outlet (e.g.

stream or river confluence). They are a convenient and manageable size for many restoration and protection applications. For the initial baseline estimate, monitoring results were given preference over model results. Ratings for the monitored HUC12's and unmonitored, or modeled, HUC12's were then merged, area-weighted (see below), and summed to obtain the baseline expressed as a percentage of the Chesapeake basin area supporting healthy streams.

Area-Weighting to Avoid Spatial Bias

Station density during the 2006 – 2011 baseline period averaged one station for every 8.8 mi^2 in the Chesapeake basin but sampling effort was not randomly distributed (**Figure 2**). Stream data collected by both state and county monitoring programs in the Washington, D.C. and Baltimore metropolitan region resulted in very high station densities in that region while roughly two in every five HUC12-Bioregion subwatersheds along the basin's western, more rural and forested side were not sampled. This bias is apparent when **Figure 2** is compared to a land use map of the basin (**Figure 3**). Maryland's Biological Stream Survey (MBSS) and its associated volunteer program had relatively high station densities during the baseline period, with about one station for every 1.8 mi^2. In Virginia, almost no stations were sampled in the Mid-Atlantic Coastal (MAC) bioregion during the baseline period, but the state's Piedmont bioregion was well sampled by the Department of Environmental Quality (VADEQ) and the INSTAR program run by Virginia Commonwealth University, resulting in one station for every 5.8 mi^2. Distributions in Pennsylvania were also uneven, with Pennsylvania's Department of Environmental Protection (PADEP) and the Susquehanna River Basin Commission (SRBC) apparently focused on the state's forested Northern Central Appalachian (NCA) bioregion where the average was one station for every 7.5 mi^2.

The high concentration of monitoring stations in the Washington, D.C. and Baltimore metropolitan region gives admirable coverage in that region's HUC12-Bioregion subwatersheds but can bias stream health estimates for the basin (**Figure 4A**). This spatial bias is somewhat corrected if each station's rating is "area-weighted" by a proportion of its local subwatershed area. Area weights are calculated by dividing a subwatershed's total area by its number of sampled stations. The weights are then applied to each rating in a subwatershed (see example calculation in Appendix A). Ratings from densely sampled subwatersheds receive smaller weights than ratings from sparsely sampled subwatersheds. Basin-wide percentages changed 8.1% after area-weighting, with the percentages of Poor and Very Poor (%PVP) dropping to 42.2% and the percentages of Excellent, Good, and Fair (%EGF) rising to 57.8% (**Figure 4B**). Unsampled HUC12-Bioregion subwatersheds cover 30.4% of the Chesapeake basin and when this gap in coverage also is accounted for, the *known* %EGF is 40.2% and %PVP 29.4% (**Figure 4C**). Many of the unsampled subwatersheds are forested or agricultural, so we cannot assume the area-weighted ratings shown in **Figure 4B** are accurate for the entire Chesapeake basin. Hence, the need for model results to fill the gaps.

All HUC12-Bioregion subwatersheds with sampled stations were used in the analysis above, including those with just one station. One station per HUC12-Bioregion subwatershed is generally viewed as inadequate for characterizing individual subwatersheds of that size. Our objective, however, is to characterize streams across the much larger Chesapeake basin. We investigated the impact on basin estimates of stream condition when a minimum number of stations is required for a HUC12-Bioregion subwatershed to be included in basin-wide estimates. When the required minimum increased from 1 to 10, the number of monitored subwatersheds meeting the requirement dwindled from 1,472 to 187 and %EGF dropped from 57.8% to 47.1%. The drop is largely explained by the fact that subwatersheds meeting higher minimum requirements are found primarily in the highly developed Washington-Baltimore metropolitan region. Each of the 187 subwatersheds with ten or more stations (**Figure 6**) may be considered adequately characterized, but they are spatially clumped and account for

only 10.4% of the total Chesapeake basin area. We conclude minimum station requirements for subwatersheds may not be necessary and could in fact be counter-productive when the goal is a basin-wide estimate of acceptable stream health.

Unlike the densely sampled subwatersheds, HUC12-Bioregion subwatersheds with just one station appear to be distributed randomly across the basin with the possible exception of the Washington-Baltimore metro area (**Figure 7**). They comprise 36.8% of the 1,472 sampled subwatersheds in the baseline period. If we look at just the single-station HUC12-Bioregion subwatersheds—effectively imposing a *maximum* station limit of one station per subwatershed—and calculate stream health for the basin, the estimated %EGF is 62.8% and the %PVP 37.2%. These values change less than 3% as the maximum station limit per subwatershed is increased incrementally from one to ten (**Figure 8**). The stability in the estimated %EGF suggests the sparsely sampled subwatersheds are randomly distributed. The Moran I statistic, calculated with the spatial statistics toolbox in ArcGIS (10.5), confirms that 1- and 2-station HUC12-Bioregion subwatersheds are, in fact, randomly distributed across the basin (Moran's Index = 0.000998, z-score = 0.465, p-value = 0.642). The random distribution of 1- and 2-station subwatersheds is not surprising given that many state monitoring programs have adopted probability-based sampling designs.

This randomness is not repeated at the smaller catchment scale. Catchments with only one and two stations are very common (n = 5,365) because their smaller sizes tend to preclude multiple sampling stations. As a result, a map of the 1- and 2-station catchments reflects the higher sampling intensities in Maryland and the Baltimore-Washington area (**Figure 9**).

This brief analysis reveals the consequences, sometimes unintended, of some common "analysis rules" when estimating percentage of healthy streams in the Chesapeake basin. The number of stations, their distribution across the heterogeneous Chesapeake basin landscape, and the subwatershed size that each station represents (i.e., catchment vs subwatershed) need to be carefully considered when calculating basin estimates of %EGF from monitoring data. We believe HUC12 subwatersheds with as few as one sample can be included in *basin* estimates of stream health, assuming station ratings are appropriately weighted by a portion of their subwatershed's area. This greatly expands the area of the basin with useable monitoring data. Results from the sparsely sampled subwatersheds suggest a baseline %EGF for the *basin* should be in the neighborhood of 60%.

Variability in Monitoring-Based Ratings

Taking advantage of the random distribution of the sparsely sampled subwatersheds, we looked at the effect of sampling effort on variation around estimates of the basin's %EGF. We randomly selected subwatersheds without replacement from the subset of monitored single-station HUC12-Bioregion subwatersheds and calculated their overall %EGF. To account for the lack of single-station subwatersheds in the Washington-Baltimore metropolitan region and to some extent the Pennsylvania NCA bioregion, we created and added to the analysis dataset the ratings of the average BIBI scores from 43 randomly picked HUC12-Bioregion subwatersheds containing ten or more stations. With these additional subwatersheds, the analysis dataset totaled 584 HUC12-Bioregion subwatersheds and had a %EGF of 61.0%. We assume for purposes of this analysis that the basin's true %EGF was 61.0% and variation around this value reflects the inherent heterogeneity of the Chesapeake basin landscape. In other words, relatively homogeneous landscapes require fewer sampled subwatersheds to reach a desirable confidence interval in the basin's %EGF estimate while heterogeneous landscapes require more sampled subwatersheds. Fifty random draws of the analysis dataset were done for each of several sample sizes. Standard deviation dropped rapidly from $\pm$ 8.1% to $\pm$ 3.8% as the number of sampled subwatersheds increase from 35 to 100. Standard deviation then dropped more slowly as subwatershed numbers continue to increase (**Figure 10**). Standard error follows a similar trajectory. Sample sizes

greater than 100 showed little variation in %EGF, ranging from 60.2% to 61.1% around the true value of 61%. When subwatersheds are randomly distributed across the Chesapeake basin, random samples of 100 or more subwatersheds appear to capture much of the variability relating to the Chesapeake basin's current landscape. Confidence in the %EGF estimate improves as sample size increases above 100.

Model-Based Ratings

To fill gaps in the basin's monitoring coverage for the 2006 – 2011 baseline period, the Random Forest model developed by Maloney et al. (2018) was re-run with 2006 – 2011 Chessie BIBI data. When developing the model, we used Chessie BIBI ratings of 2,815 density-corrected (to remove effect of high sample density in Maryland region), independent stations on 1^{st} – 4^{th} order streams as the response variable in the model. We used both a training data set which consisted of 75% of the samples (n = 2,111 samples) and independent validation data set (n = 704) to assess model performance. Predictor variables in the model included cumulative drainage area, mean upstream elevation; upstream sandy soils, upstream soils in hydrologic groups D plus A, B, C with high water tables, upstream total precipitation, latitude, longitude, dominant upstream bioregion, upstream SO_4 deposition, mean % upstream land surface cover as impervious surface (2006 NLCD), mean % upstream land surface in tree canopy cover (2006 NLCD), mean % upstream agriculture cover (2006 NLCD) and number of dams within the local catchment (see SHEDS data base for more detailed description at www.ecosheds.org).

Chessie BIBI ratings were predicted by the model for 95,877 catchments in the basin. These catchments covered 91.9% of the basin's total area. The model response variable was structured in four ways: as raw scores, as the five original Chessie BIBI ratings (E, G, F, P, and VP), as three rating categories (EG, F, and PVP), and as two rating categories (EGF and PVP). Straightforward tallies of the area-unweighted ratings show the basin %EGF estimate from the predicted raw scores was 64.9% (**Table 1**). Basin %EGF estimates for the five-, three-, and two-rating model versions were lower and within a few percentage points of each other, ranging from 60.0% to 62.5%. When modeled catchment results are area-weighted, the %EGFs changed slightly: 64.6% (raw), 59.4% (five-rating), 61.8% (three-rating), and 60.2% (two-rating). These estimates are close to the %EGF of ~60% generated from the monitoring data pool of randomly distributed 1- and 2-station subwatersheds (above).

Ratings predicted by the Random Forest model can be directly compared to monitoring-based ratings in 4,888 catchments (**Table 2**). (For the 1,143 catchments with two or more monitoring stations, index scores were averaged, and the average given its appropriate family-level bioregion rating.) The model version that predicts raw scores substantially underestimates VP (-10.8%) and E (-10.1%) and overestimates P (+10.0%), F (+7.4%), and G (+3.5%). When both monitoring and modeling raw scores are grouped into the three-rating categories, the PVP percentages compare well but the model overestimates F (+7.4%) and underestimates EG (-6.6%). When EG and F are combined, model and monitoring %EGFs are only 0.8% different. The raw score model is capable of distinguishing EGF from PVP on a gross level but does not do well distinguishing the five or three individual ratings. Therefore, it may not be able to detect shifts from VP to P or from G to E.

Of the three models that predict ratings instead of raw scores, the two-rating version had the highest percentage of disagreement (16.5%). The five-rating model version had a somewhat lower percentage of disagreement (13.6%) but did poorly at exactly matching ratings (63.5%) and had the largest number of "Near Match" ratings (22.9%). The three-rating model version had the lowest disagreement (9.1%), did well matching the ratings exactly (75.7%), had the lowest number of "Near Match" ratings (15.2%), and its percentages of EG, F, and PVP were very comparable to the monitoring results. Overall, output from the three-rating model clearly separates the best and worst Chessie BIBI ratings and represents the Fair category well.

For the preliminary baseline, we decided to use the three-rating model to fill gaps in the basin's monitoring coverage. This was chiefly based on the strong similarities in the monitoring and three-rating model results when the two are directly compared (75.7% exact matches, 9.1% disagreement), and on the closeness of their %F values. Fair is indicative of streams with uncertain or intermediate status, and an accurately predicted F rating might be helpful in identifying and planning restoration projects with significant potential for "lift."

The Preliminary 2006 – 2011 Baseline

The five ratings of the monitoring-based Chessie BIBI results were re-grouped into the three rating categories to make them directly comparable to output from the predictive model's three-rating version. The three-rating monitoring and model results were each grouped by HUC12 and area-weighted with weights derived from the HUC12 subwatershed area and either the number of sampled stations or the number of modeled catchments. The area-weighted HUC12 results were then combined, with monitoring ratings given preference and model-based ratings included only if the HUC12 subwatershed was unsampled. HUC12 subwatersheds with monitoring data account for 73% of the basin area and HUC12s with model results cover 27% of the basin area. The combined monitoring and model HUC12 results were then rolled up to the Chesapeake basin scale to calculate percentages of Excellent/Good (EG), Fair (F), and Poor/Very Poor (PVP).

An estimated 43% of the Chesapeake basin's area supported streams with EG ratings in the 2006 – 2011 baseline period. EG ratings indicate well-functioning biological communities that strongly resemble ones in undisturbed streams (Reference). Another estimated 40% of the basin's area supported streams with PVP ratings, indicating stressed or poorly-functioning biological communities. The remaining 17% of the basin area supported streams with the intermediate Fair rating. Communities in Fair streams are not clearly reference-like or degraded but may still function at satisfactory levels. A previous Chesapeake Bay Program goal (FLCCB 2014) identified Fair as an acceptable stream condition. Therefore, we summed the EG and F percentages to determine the percentage of acceptable stream conditions. The overall %EGF was estimated to be 60% of the Chesapeake basin area during the baseline period.

Management Implications

The Chesapeake Bay Watershed Agreement (2014) states the CBP partnership will "*Continually improve stream health and function throughout the watershed. Improve health and function of ten percent of stream miles above the 2008 baseline for the Chesapeake Bay watershed*." The Chessie BIBI was selected as the initial CBP indicator of stream health because it was ready to use and it represents the ecologically important and frequently monitored stream macroinvertebrate community (CBP 2015). This report developed a preliminary "2008 baseline" with Chessie BIBI ratings derived from monitoring results and model predictions. **Several lines of evidence indicate about 60% of the Chesapeake basin's *area* supported healthy communities of stream macroinvertebrates between 2006 and 2011, the 6-year period selected as the baseline.** A final baseline estimate will be produced after a higher resolution stream layer becomes available and stream health can be estimated as a percent of the basin's *stream miles* instead of percent of the basin's area.

The intent of this baseline is to provide an overall estimate of healthy streams across the Chesapeake basin during the 2006 – 2011 period. Different analysis approaches are required to characterize individual subwatersheds or map the basin results for that period. Additional research may be needed to find the optimal number of stations to characterize an individual subwatershed.

Other biological communities are good indicators of stream health, but they could not be used in this baseline period because they either were not monitored throughout the basin or a basin-wide index of their status had not been developed yet. Ideally, indices of other biological communities—or even indicator species such as Brook Trout—should be included in future basin-wide evaluations of stream health. Each community responds differently to stress and varies in their resilience to change. **Going forward, a mix of biological indices can provide a more nuanced and complete picture of stream health in the basin.**

Restoration efforts can significantly improve many stream functions that are expressed in the physical and chemical features of stream corridors (e.g., riparian buffer, energy level, water storage, pH, connectivity, sediment yield and substrate character). However, recovery of stream health following the implementation of management actions often lags improvements in stream function or is not observed. Biological recovery may be limited if restoration efforts focus on just one or two physical or chemical stream features. Recovery also can be stymied in highly developed areas by the lack of nearby sources of biota to colonize newly renovated streams. Biological recovery of a stream has the greatest potential to succeed if the needs of the biological community are met; that is, functions that support the biological community are present. Assessing stream improvement with biology alone (stream health) may provide a restricted view of the stream's recovery potential. The 2014 Agreement calls for "ten percent" improvement in stream function as well as stream health. A variety of methods are already available that can facilitate assessments of a broader suite of stream functions to evaluate stream restoration effectiveness at different scales (e.g., Dolloff et al. 1993, NCSU 2006, MDDNR 2003, EPA 2006, USC 2017). The CBP Stream Health Work Group includes in its workplan the development of practicable metrics that are consistent with the verification of stream restoration best management practices and their nutrient and sediment benefits along with other stream functional improvements. **An index, or suite of practicable metrics, of Chesapeake stream *functions* which incorporates physical and chemical metrics measured basinwide could help determine the effectiveness of diverse stream restoration and other management efforts in rehabilitating stream health**.

The reader is reminded that the Chessie BIBI index is not used by the Chesapeake basin's six states and the District of Columbia to assess impairment for state 303(d) listing purposes. Each jurisdiction has its own stream assessment methods and biocriteria. The Chessie BIBI index is intended to be used by CBP partners as a tool for regional planning, for tracking progress in the basin, and for CBP reporting.

References

Federal Leadership Committee for the Chesapeake Bay (FLCCB). 2014. Combined FY 2014 Action Plan and FY 2013 Progress Report Strategy for Protecting and Restoring the Chesapeake Bay Watershed.

Chesapeake Bay Program (CBP). 2014. Chesapeake Bay Watershed Agreement.

Chesapeake Bay Program (CBP). 2015. Stream Health Outcome Management Strategy. CBP Stream Health Workgroup.

Chesapeake Bay Program (CBP). 2018. Workshop to develop a 2008 baseline for the CBP stream health outcome indicator, Cacapon Resort State Park, WV, 5th – 6th April 2018. Workshop report.

Dolloff, C. A., D. G. Hankin, and G. H. Reeves. 1993. Basinwide Estimation of Habitat and Fish Populations in Streams. U.S. Forest Service, Southeastern Forest Experiment Station, General Technical Report SE-83.

Environmental Protection Agency (EPA). 2006. Wadeable Streams Assessment: A Collaborative Survey of the Nation's Streams. EPA 841-B-06-002.

Law, N., B. Stack, R. Starr, and E. Yagow. 2015. Designing sustainable stream restoration projects within the Chesapeake Bay watershed. STAC Publication Number 15-003, Edgewater, MD. 50 pp.

Maloney, K., Z. M. Smith, C. Buchanan, A. Nagel, and J. A. Young. 2018. Predicting Stream Biological Conditions for Small Headwater Streams in the Chesapeake Bay Watershed. Freshwater Science 37(4):795-809. DOI: 10.1086/700701.

Maryland Department of Natural Resources (MDDNR). 2003. A Physical Habitat Index for Freshwater Wadeable Streams in Maryland, Final Report. CBWP-MANTA-EA-03-4.

North Carolina State University (NCSU). 2006. Stream Restoration Evaluation Assessment Form. NCSU Water Quality Group. http://www.ncsu.edu/waterquality/

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

Upper Susquehanna Coalition (USC). 2017. Stream Corridor Assessment Guide. http://www.u-s-c.org/html/Stream%20Corridor%20Assessment%20Guide%204-21-17.pdf

Table 1. Random Forest model results for 95,877 catchments in the Chesapeake basin, 2006 – 2011. Different model versions predict the index response as raw scores, five ratings, three ratings, and two ratings. Raw scores are grouped into the five-, three- and two-rating categories for comparison. Overall percentages of Excellent, Good, and Fair (%EGF) and of Poor and Very Poor (%PVP) for each model version are shown at the bottom. Results are not area-weighted by catchment size. See text for further details.

Predicted response	Raw Scores			Five-Ratings	Three-Ratings	Two-Ratings
Excellent	5.4%	36.5%	64.9%	22.9%	45.1%	60.6%
Good	31.0%			21.8%		
Fair	28.4%	28.4%		15.3%	17.4%	
Poor	31.4%	35.1%	35.1%	19.7%	37.5%	39.4%
Very Poor	3.7%			20.4%		
%EGF		64.9%		60.0%	62.5%	60.6%
%PVP		35.1%		40.0%	37.5%	39.4%

Table 2. Direct comparison of model- and monitoring-based ratings in 4,888 Chesapeake catchments. For catchments with two or more monitoring stations, station scores are averaged, and the averages rated with the relevant bioregion-specific thresholds. The five-rating results from the monitoring data are regrouped to form the equivalent ratings produced by the four model versions. Modeled raw scores are also grouped into five-, three-, and two-rating categories for comparison purposes. "Match" indicates rating categories match exactly (e.g. Fair versus Fair); "Near Match" indicates rating categories are adjacent (e.g., Good versus Fair); "Disagree" indicates rating categories are separated by at least one rating categories (e.g., Good versus Poor).

	Monitoring			Model					
Rating				Raw Scores			Five-Ratings	Three-Ratings	Two-Ratings
Excellent	17.2%	33.6%	49.3%	7.1%	27.0%	50.1%	16.6%	31.5%	45.4%
Good	16.3%			19.8%			15.9%		
Fair	15.7%	15.7%		23.1%	23.1%		13.5%	14.8%	
Poor	27.5%	50.7%	50.7%	37.5%	49.9%	49.9%	25.4%	53.7%	54.6%
Very Poor	23.2%			12.4%			28.6%		
	Match			51.7%	71.7%	80.4%	63.5%	75.7%	83.5%
	Near Match			37.7%	22.0%	NA	22.9%	15.2%	NA
	Disagree			10.6%	6.3%	19.6%	13.6%	9.1%	16.5%

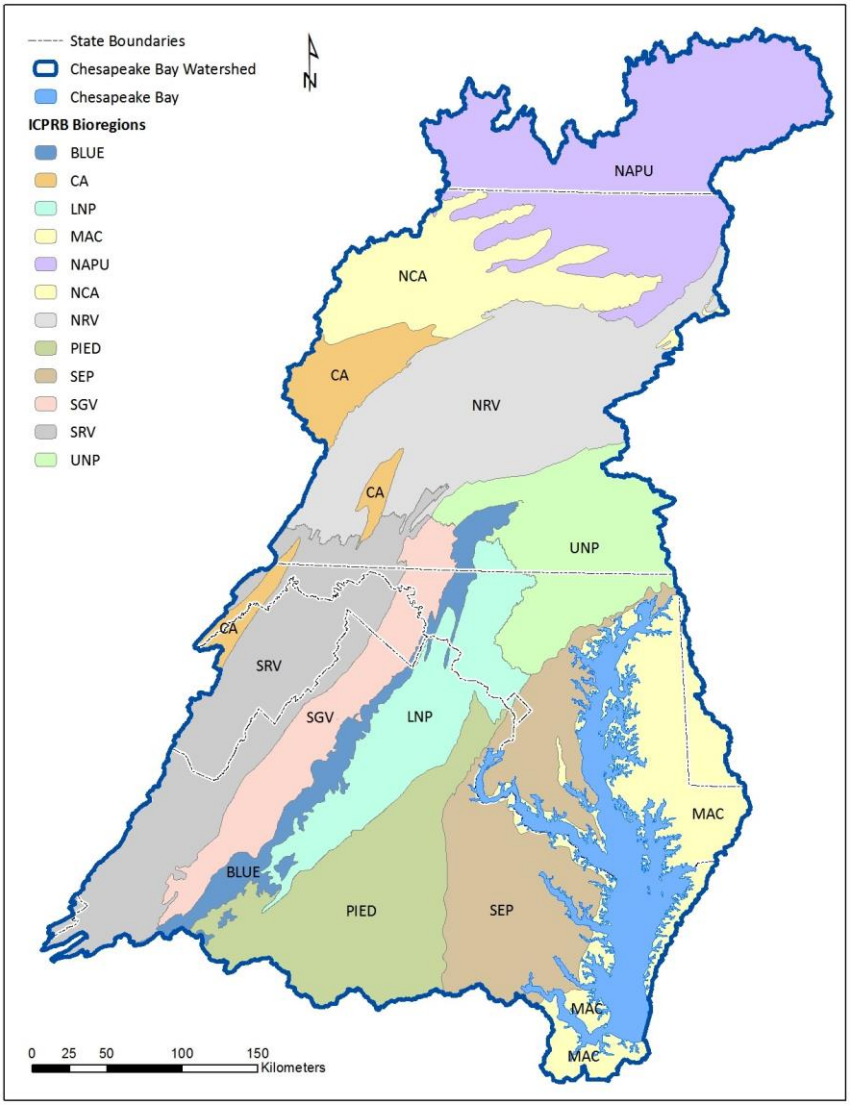

Figure 1. The twelve Chesapeake bioregions. BLUE (Blue Ridge), CA (Central Appalachians), LNP (Lower Northern Piedmont), MAC (Middle Atlantic Coastal Plain), NAPU (Northern Appalachian Plateau and Uplands), NCA (North Central Appalachians), NRV (Northern Ridge and Valley), PIED (Piedmont), SEP (Southeastern Plains), SGV (Southern Great Valley), SRV (Southern Ridge and Valley), UNP (Upper-Northern Piedmont).

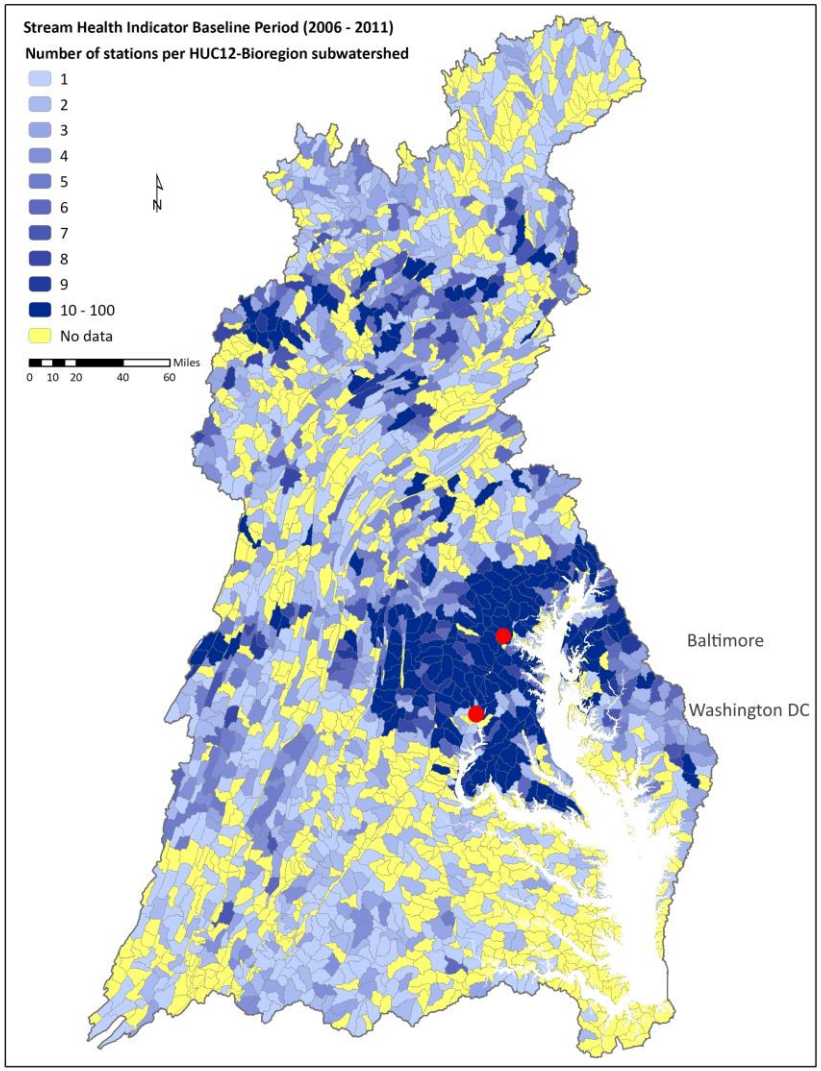

Figure 2. Stream macroinvertebrate sampling intensity in HUC12-Bioregion subwatersheds of the Chesapeake Bay basin, during baseline period (2006 – 2011). Red dots indicate Baltimore and Washington, DC.

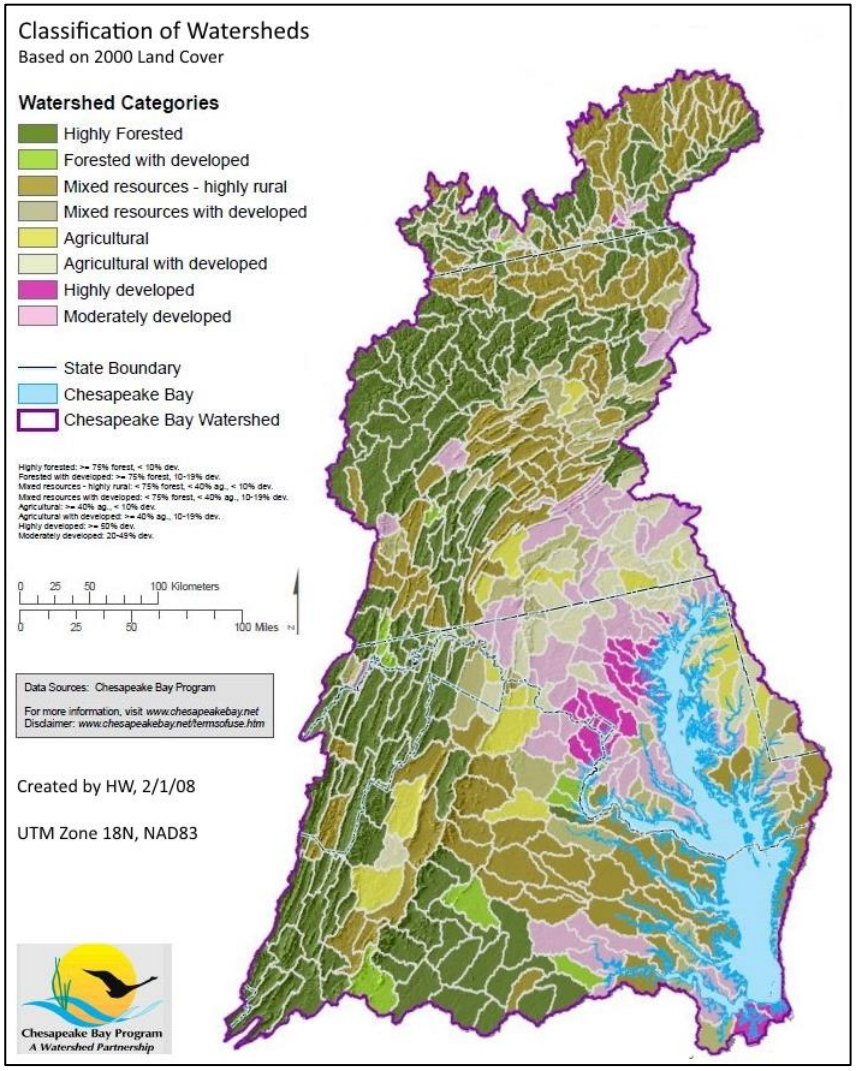

Figure 3. Land cover (2000) in the Chesapeake Bay basin (adapted from CBP).

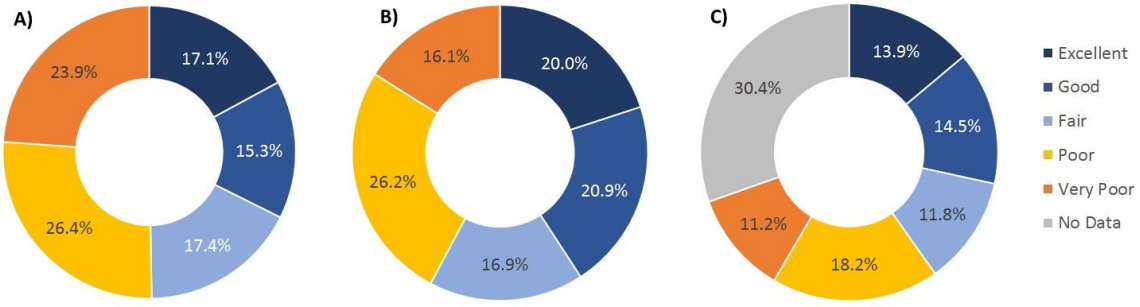

Figure 4. Basinwide percentages of the five Chessie BIBI ratings calculated three ways: A) straightforward summation of the number of stations with each rating; B) summation of the ratings after each has been area-weighted by a weight equal to the total area of its HUC12-bioregion subwatershed divided by the number of sampled stations in the subwatershed (includes subwatersheds with just one station); and C) same as B but includes the areas of unsampled HUC12-bioregion subwatersheds ("No Data").

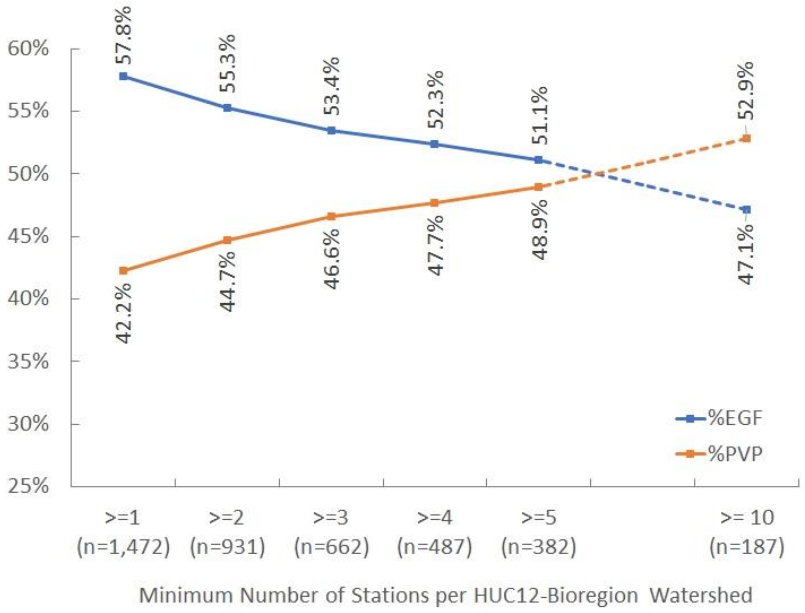

Figure 5. Area-weighted, normalized* percentages of Chessie BIBI ratings during the baseline period calculated with different requirements for the *minimum* number of stations per subwatersheds (n). For example, all watersheds are included in ">=1" case; only subwatersheds with two or more stations are included in the ">=2" case, etc. %EGF is the percentage of sampled subwatersheds with Excellent, Good or Fair ratings; %PVP is the percentage of sampled subwatersheds with Poor or Very Poor ratings. *Unsampled subwatershed areas are excluded and percentages are calculated from the total area of the sampled subwatershed areas.

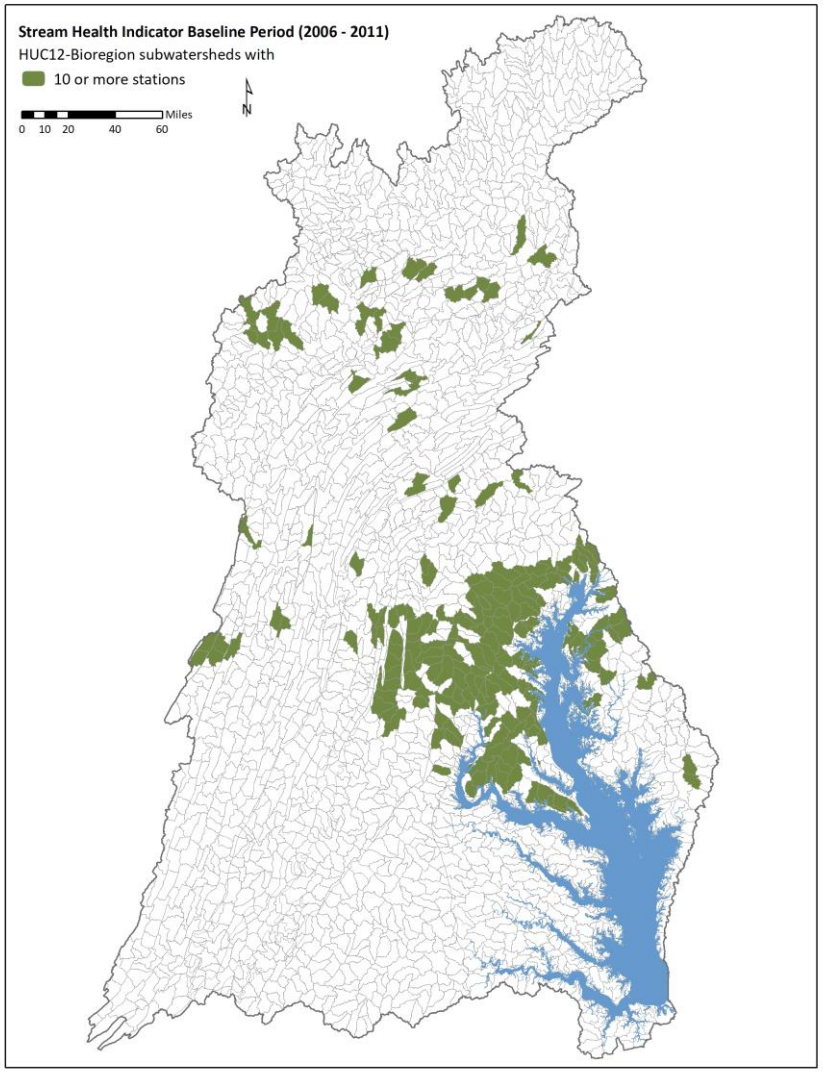

Figure 6. HUC12-Bioregion subwatersheds with ten or more stations in the 2006 – 2011 baseline period.

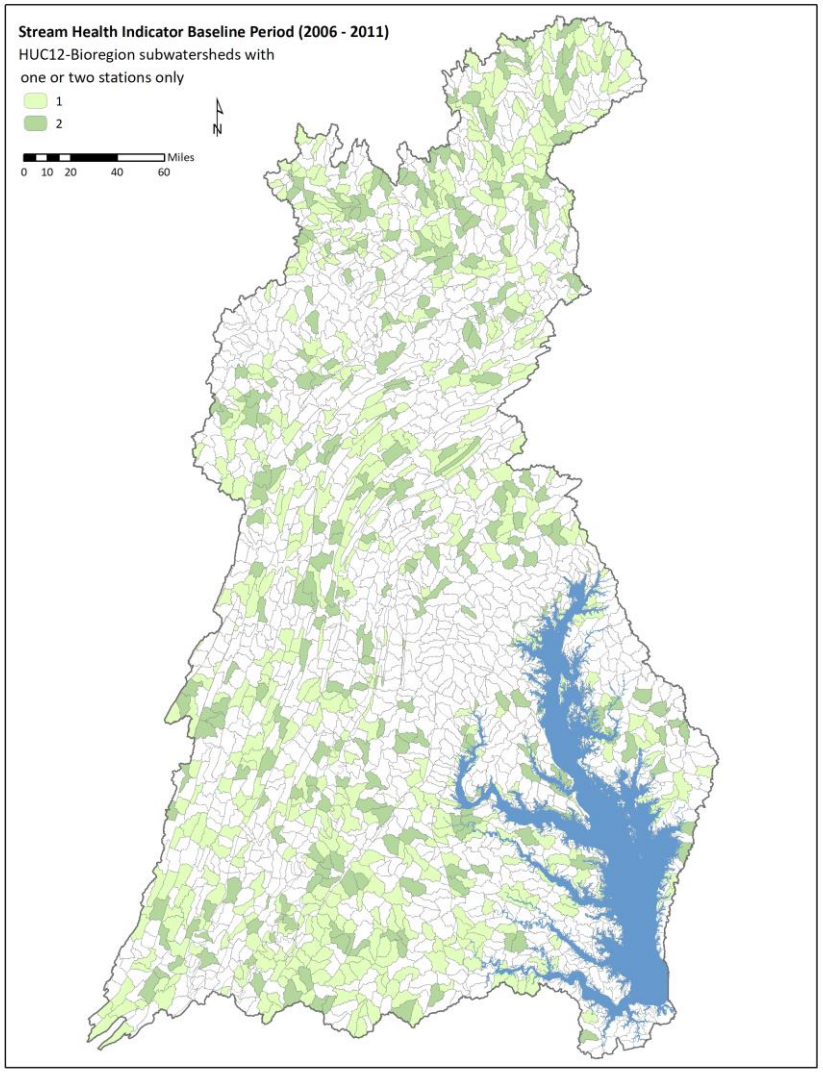

Figure 7. HUC12-Bioregion watersheds with only one or two stations in the 2006 – 2011 baseline period.

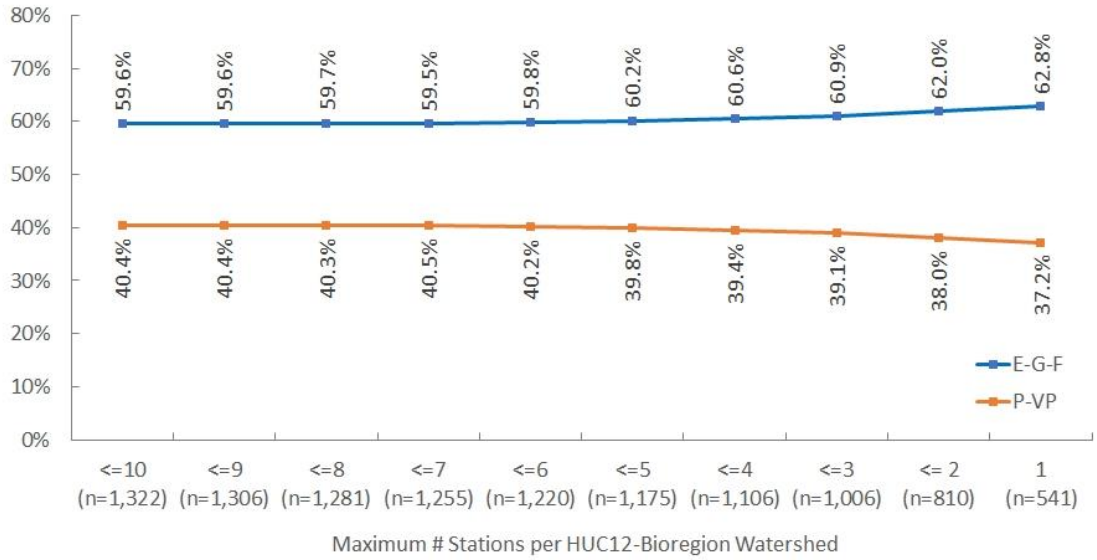

Figure 8. Area-weighted, normalized* percentages of Chessie BIBI ratings during the baseline period calculated with different requirements for the *maximum* number of stations per subwatersheds (n). For example, only subwatersheds with 10 or fewer stations are included in the case of "<=10", … , only subwatersheds with a single station are included in the case of "1." %EGF is the percentage of sampled subwatersheds with Excellent, Good or Fair ratings; %PVP is the percentage of sampled subwatersheds with Poor or Very Poor ratings. *Unsampled subwatershed areas are excluded and percentages are calculated from the total area of the sampled subwatershed areas.

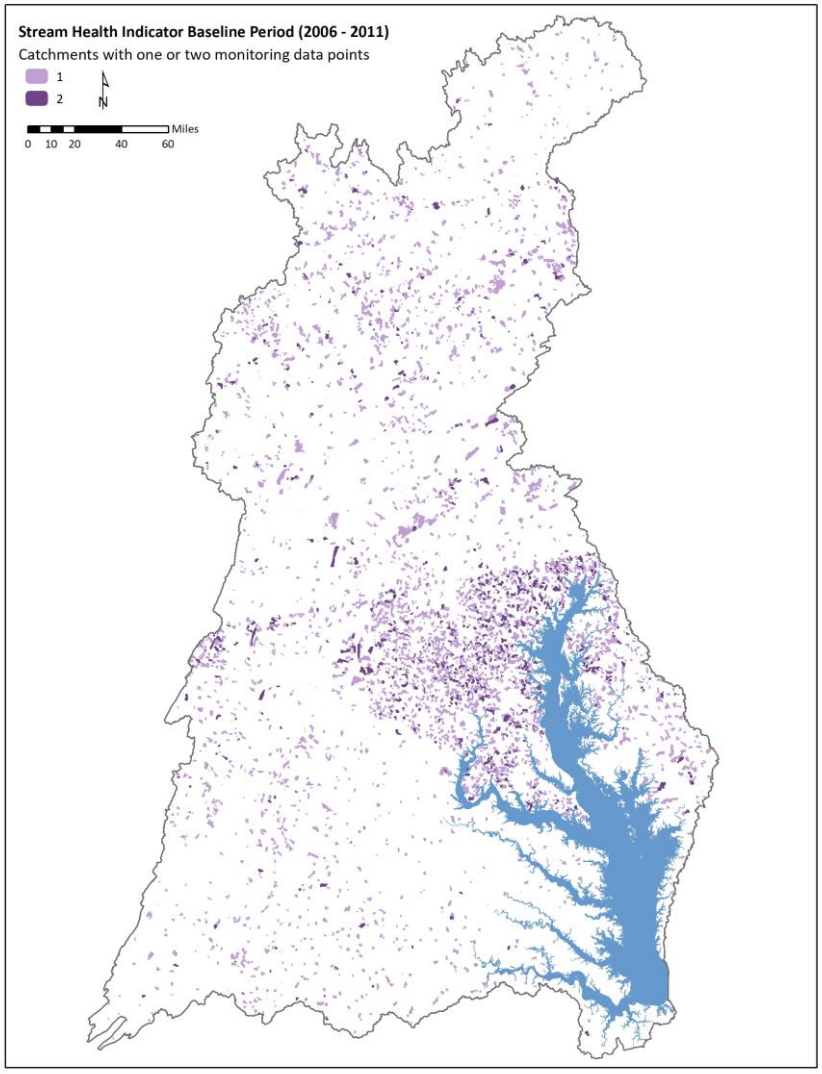

Figure 9. Catchments with only one or two monitoring stations in the 2006 – 2011 baseline period.

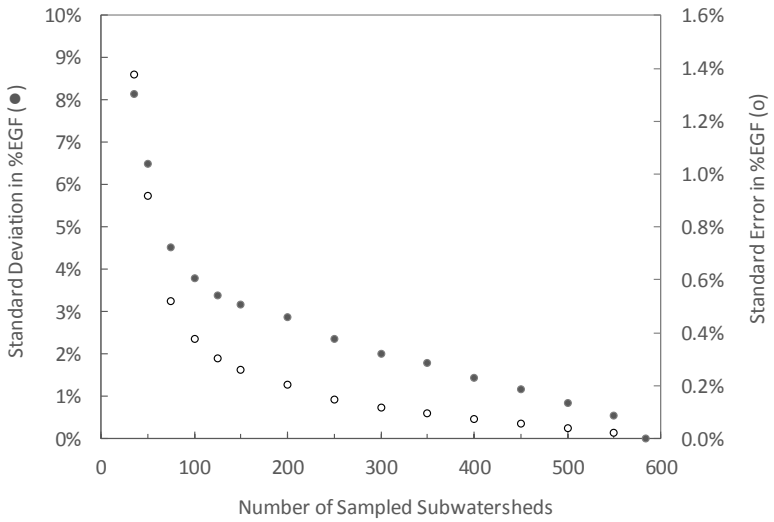

Figure 10. Standard deviation (●) and standard error (o) in the percent of Excellent, Good, and Fair (%EGF) Chessie BIBI ratings *versus* number of HUC12-Bioregion subwatersheds sampled. See text for further details.

Appendix A. Area-Weighting Calculation

Example: find the overall, area-weighted percentages of EGF and PVP in thirteen HUC12 subwatersheds.

1. Stations are first grouped by HUC12-Bioregion subwatershed units. Average each station's Chessie BIBI index scores and rate the average score according to bioregion-specific rating thresholds (Table A-1). Ratings are: Excellent (E), Good (G), Fair (F), Poor (P), and Very Poor (VP).

Table A-1. Bioregion-specific thresholds used to rate numeric Chessie BIBI index scores. Values less than the VP|P thresholds are rated "Very Poor", values less than the P|F threshold are rated "Poor", values less than the F|G threshold are rated "Fair", values less than the G|E threshold are rated "Good", and values equal to or greater than the G|E threshold are rated "Excellent." Ratings are for the family-level, bioregion version of the Chessie BIBI index (from Smith et al. 2017).

Bioregion Name	Code	VP|P	P|F	F|G	G|E
Blue Ridge	BLUE	30.87	61.74	82.21	92.07
Central Appalachians	CA	27.88	55.77	69.55	78.56
Lower Northern Piedmont	LNP	35.16	70.32	80.15	91.30
Mid-Atlantic Coast	MAC	22.72	45.43	63.18	76.85
Northern Appalachian Plateau & Upland	NAPU	17.93	35.85	47.42	61.95
North Central Appalachians	NCA	15.73	31.46	56.25	78.66
Northern Ridge & Valley	NRV	19.10	38.21	50.81	75.00
Piedmont	PIED	30.31	60.62	73.18	81.83
Southeastern Plains	SEP	16.28	32.55	56.66	83.90
Southern Great Valley	SGV	29.93	59.86	66.02	76.48
Southern Ridge & Valley	SRV	21.58	43.16	57.97	71.93
Upper Northern Piedmont	UNP	31.30	62.59	69.85	80.52

2. Re-group monitoring station ratings into the three rating categories and area-weight. Group stations by HUC12 subwatershed and area-weight each station's rating. Weights are the area of the HUC12 divided by the number of monitoring stations in the HUC12. Weights are applied by multiplying the number of stations with a given rating by the weight (blue text) as shown below.

HUC12	E	G	F	P	VP	Station Count	HUC12 km²	weight	EG	F	PVP
020501010102				1		1	106.71	106.71	0	0	106.71
020501010103		1	2	1		4	100.55	25.14	25.14	50.28	25.14
020501010202			2	1		3	66.28	22.09	0	44.19	22.09
020501010203		1		1		2	53.02	26.51	26.51	0	26.51
020501010204			1	1		2	59.51	29.75	0	29.75	29.75
020501010301				1		1	102.20	102.20	0	0	102.20
020501010304		1	1			2	60.86	30.43	30.43	30.43	0
020501010405	1	1				2	113.43	56.72	113.43	0	0
020501010501				1		1	50.96	50.96	0	0	50.96
020501010502		1				1	76.96	76.96	76.96	0	0
020501010603			1			1	108.45	108.45	0	108.45	0

020501010604	1	1		2	122.54	61.27	61.27	61.27	0
020501010701			1	1	54.82	54.82	0	0	54.82

3. Sum the HUC12 areas and each of the weighted ratings as shown below. %EGF is equal to (333.74 + 324.36) divided by 1,076.29, or 61.1%. %PVP is equal to 418.18 divided by 1,076.29, or 38.9%.

Sums:	1,076.29		333.74	324.36	418.18
HUC12	HUC12 km^2	weight	EG	F	PVP
020501010102	106.71	106.71	0	0	106.71
020501010103	100.55	25.14	25.14	50.28	25.14
020501010202	66.28	22.09	0	44.19	22.09
020501010203	53.02	26.51	26.51	0	26.51
020501010204	59.51	29.75	0	29.75	29.75
020501010301	102.20	102.20	0	0	102.20
020501010304	60.86	30.43	30.43	30.43	0
020501010405	113.43	56.72	113.43	0	0
020501010501	50.96	50.96	0	0	50.96
020501010502	76.96	76.96	76.96	0	0
020501010603	108.45	108.45	0	108.45	0
020501010604	122.54	61.27	61.27	61.27	0
020501010701	54.82	54.82	0	0	54.82