Workshop Background Materials: Developing a 2008 Baseline for the CBP Stream Health Outcome Indicator

Cacapon Resort State Park, WV

5th – 6th April 2018

Higher quality streams are a desired outcome of the Chesapeake Bay's nutrient and sediment TMDL and many restoration and preservation efforts underway in the watershed. A goal in the 2014 Chesapeake Bay Agreement is:

"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 watershed."

The Chesapeake Bay Program (CBP) selected the "Chessie BIBI" as its biological indicator of stream health, and will use the index to track and report progress towards accomplishing this goal (Stream Health Management Strategy 2015). A 2008 baseline for the Chessie BIBI needs to be established.

Before a baseline can be established, technical decisions need to be made on how to express results from individual monitoring sites in terms of stream miles and how to minimize geographic bias across the Bay watershed. Six possible methods for doing this are illustrated in this document. Each has its strengths and weaknesses. When the technical issues are resolved, consensus can be built on a 2008 baseline and how to measure improvement in the stream health outcome indicator—and possibly other indicators.

Chessie BIBI Refinement

The Chessie BIBI, or Chesapeake Basin-wide Index of Biotic Integrity, is a reference-based, multi-metric index. It was developed from an integrated database of benthic macroinvertebrate data sets originally collected by federal, state, and county agencies, and volunteers in the Chesapeake Bay watershed. The database was last updated in 2016 and holds macroinvertebrate taxonomic counts, habitat scores, and water quality results for 25,067 sampling events between 1992 and 2015. The Interstate Commission on the Potomac River Basin (ICPRB) used the updated database to refine the original Chessie BIBI index. The final report describing the refinement can be downloaded from ICPRB's website (Smith et al. 2017). Development of the original Chessie BIBI (Buchanan et al. 2011) and its refinement in 2017 were guided by Technical Advisory Groups (TAG) comprised of the data providers and area experts. In the future, CBP would like to routinely update the database and report on stream health using the refined Chessie BIBI index. Results will be incorporated into the Program's CEDR database.

Taxonomic Resolution

The Chessie BIBI refinement explored three taxonomic resolutions (i.e., order, family, and genus) and three spatial resolutions (i.e., Bay watershed, region, and bioregion). Order-level indices provided promising results but variability in these indices was often greater than the family and genus resolutions. Genus-level indices generally had classification efficiencies comparable to the family-level indices, and therefore, following the concept of parsimony, the family-level indices were favored. Using season as an additional blocking factor for generating indices may result in a significant improvement in genus-level performance, but was beyond the scope of the current refinement.

Spatial Resolution

The coarsest spatial resolution uses just one suite of macroinvertebrate metrics to evaluate all streams across the basin. Morphological differences between inland and coastal streams, and the resulting natural differences in stream populations, strongly bias the results and a single index for the entire Bay watershed is not recommended. The region and bioregion indices account for natural geographic differences that affect biological communities and produce relatively unbiased results. The two regional indices (Coastal, Inland) are favored for CBP reporting purposes because of parsimony and are also supported by Waite et al. (2014). By aggregating large swaths of the basin, the regional indices are derived from large sample sizes and more reference sites. They are constructed for each region with metrics that clearly discriminate between reference and degraded conditions. The twelve bioregion indices reflect macroinvertebrate community composition in smaller, more homogeneous areas. The bioregion indices are also constructed with the metrics most sensitive to each bioregion's natural characteristics. The number and quality of reference conditions varies by bioregion and affects the distributions of Reference index scores.

Narrative Ratings

Even though the region and bioregion indices are developed in the same manner, the numeric values of one index's scores should not be directly compared to another index's scores. This is because each index reflects the biases and limitations of the underlying data. The same habitat and water quality thresholds are used to identify Reference sites everywhere, but the count of Reference sites and the inherent quality of the Reference sites in each region or bioregion are different. The differences influence how the index's component metrics are scored, which in turn affects the final index score. These differences are overcome when a common approach is used to assign narrative ratings to an index's numeric scores. For each region and bioregion, the 50th, 25th, and 10th percentiles of the index scores in the reference environmental conditions are used to define Excellent, Good, Fair, and Poor macroinvertebrate status. A fifth rating, Very Poor, is defined by half the value of the 10th percentile. The narrative ratings of the various indices indicate stream

health in each region or bioregion relative to consistently identified reference conditions, and as such are directly comparable.

Application

The twelve bioregions may be slightly more specific and sensitive to geographic differences but summarizing and comparing results across the basin is easier and equally sensitive with the regional indices. Smith et al. (2017) recommend the regional family-level indices for CBP reporting purposes; the family-level bioregion indices represent a valid alternative and could be used in combination with the family-level regional indices to evaluate stream status and trends locally in the Chesapeake Bay basin.

2008 Baseline Workshop

A morning workshop scheduled for the 2018 Association of Mid-Atlantic Aquatic Biologists (AMAAB) conference at Cacapon Resort State Park on April 5th will summarize the recent BIBI index refinement. Speakers will explain why CBP selected the BIBI as its stream health indicator, present some results illustrating BIBI responses to nutrient enrichment, flow alteration, and land use, and lay out the technical challenges of reporting monitoring results in terms of stream miles and developing a 2008 baseline. The workshop will be open to AMAAB participants. A second, smaller workshop will convene immediately following the AMAAB workshop and extend into the next day. Its aim is to bring together biologists familiar with stream macroinvertebrate monitoring data and managers who use and apply the results in Chesapeake Bay region policy-making, to resolve technical challenges and build consensus on a 2008 baseline. A final report on the workshop findings and recommendations will be drafted and submitted to the CBP Stream Health Workgroup.

Six Potential Methods

The remainder of this document presents materials to be discussed at the workshop. Six potential methods for reporting Chessie BIBI results in the Chesapeake Bay watershed in terms of stream miles are outlined below. For simplicity, the five index ratings presented in the 2017 Chessie BIBI refinement report have been consolidated into three categories. “Acceptable” represents the Excellent and Good categories from the report, “Fair” represents the Fair category, and “Degraded” represents Poor and Very Poor categories. Areas that contain no Chessie BIBI ratings, are classified as “Insufficient.” The goal is to accurately represent the number of stream miles in the Chesapeake Bay basin that are Acceptable, Fair, Degraded, and Insufficient.

The workshop will explore preliminary results of the six potential methods for two potential baseline periods (2000-2008 and 2004-2008), two BIBI indices (family-level region indices and family-level bioregion indices), and several spatial resolutions where applicable (HUC8, HUC10, and HUC12 watersheds and catchments). This document

provides examples of each method using the family-level bioregion indices and the 2004-2008 period to represent the 2008 baseline.

Spatial bias is the largest issue to resolve when reporting stream health in the Chesapeake Bay watershed. Federal, state, county, and volunteer stream monitoring programs in the watershed tend to work independently of one another and can overlap spatially. Thus, samples are not randomly distributed across the basin. For example, sampling locations are more numerous and more frequently sampled in urban areas with both state and county monitoring programs compared to rural or forested areas. Five of the six methods below attempt to reduce spatial bias by aggregating the data into spatial units defined by HUC8, HUC10, and HUC12 watersheds and catchments. Methods 1-5 use spatial units obtained from the NHDPlus Version 2 medium resolution data set. Method 6, Random Forest, uses ecosheds 1:24,000 high resolution scale catchments (www.ecosheds.org).

1. No Spatial Aggregation

This is the simplest approach for reporting stream health and contains the most spatial bias. Chessie BIBI ratings are not aggregated by any spatial feature and the method does not account for differences in sample density (Figure 1). The proportions of sample locations classifying as Acceptable, Fair, and Degraded are used to represent overall stream condition in the Chesapeake Bay watershed. Due to the greater density of samples in urban areas, and the fact that urban streams are generally more degraded, the overall assessment of Chesapeake streams is biased towards degraded. When results are shown in a low-resolution map (e.g., Figure 1), a viewer's perception of stream condition is also biased by the fact that, in densely sampled areas, dots indicating station reults of one category can overlay and mask those of other categories. Additionally, this method provides no direct way to relate the ratings to stream miles and provides no measure of areas (or stream miles) that are insufficiently sampled. The method is useful for other purposes: it indicates where sampling gaps occur, and illustrates station spatial distributions and fine-scale differences between catchments when the results are shown in high-resolution maps.

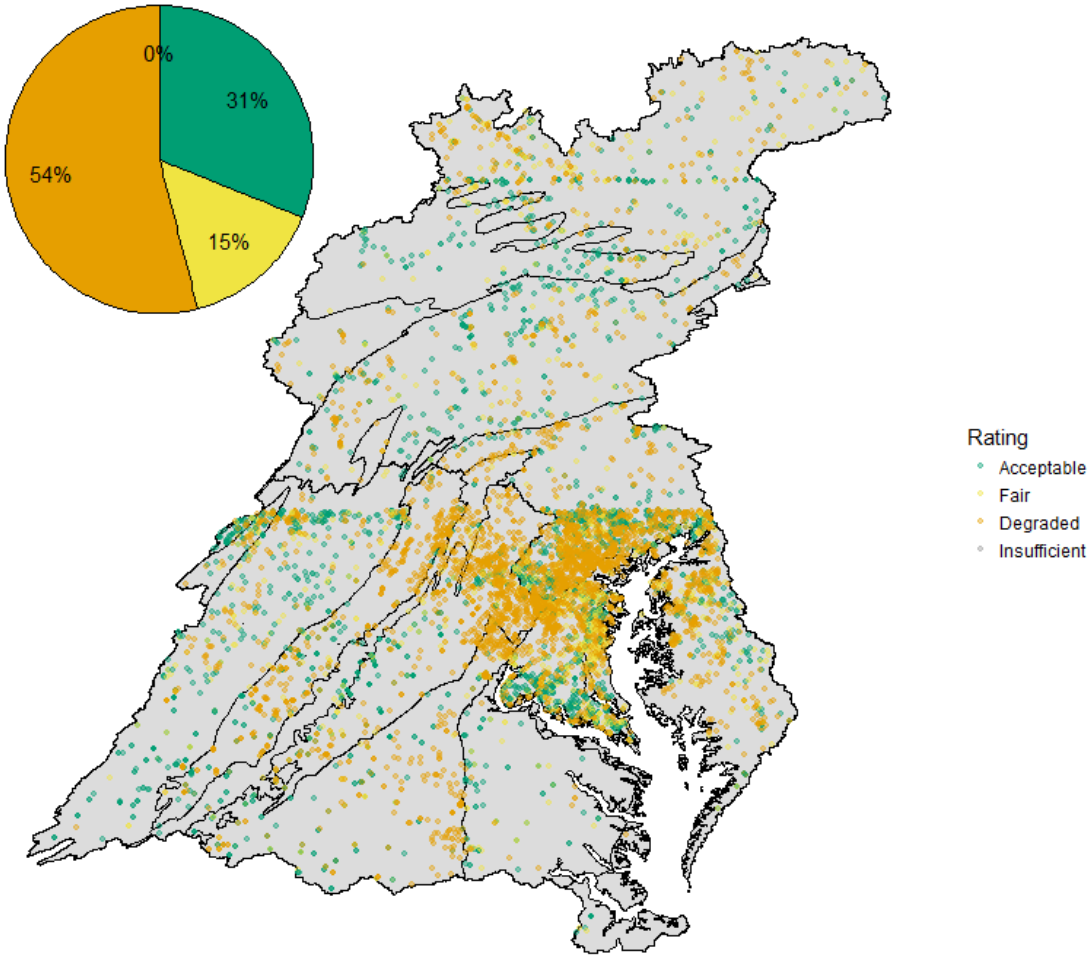

Figure 1. Each sampling location in the Chesapeake Bay basin were plotted as a point and colored based on the assigned Chessie BIBI rating. Note: stations with Degraded Chessie BIBI scores are overlaying and masking those with Fair and Acceptable scores in this example.

2. Watershed Mean Score

The Chessie BIBI scores and ratings are aggregated by a spatial unit (HUC8, HUC10, HUC12, or catchment), the mean Chessie BIBI score is calculated for each unit cell, and each unit cell is assigned a rating based on the mean score (Figure 2). Each ratings is then weighted by the number of stream miles in its unit cell and the weighted ratings are summed to represent the number of stream miles classified as Acceptable, Fair, Degraded, and Insufficient in the basin.

This method provides a more holistic view of the Chesapeake Bay basin condition by incorporating the area of the basin that cannot be classified accurately (i.e., Insufficient). Bias caused by data from densely sampled areas is also reduced. The mean is a strong

indicator when it is derived from large enough sample sizes and individual sampling events are not given undue weight. The method may misrepresent areas where just one sample represents an entire unit cell. Requiring a minimum number of samples per unit cell could resolve the issue; however, it would exclude unit cells with few sample locations and increase the proportion of spatial units classified as Insufficient. Finally, the rating classification scheme was developed for individual sampling events and assigning a rating to the mean of multiple sampling event scores may have unintended consequences.

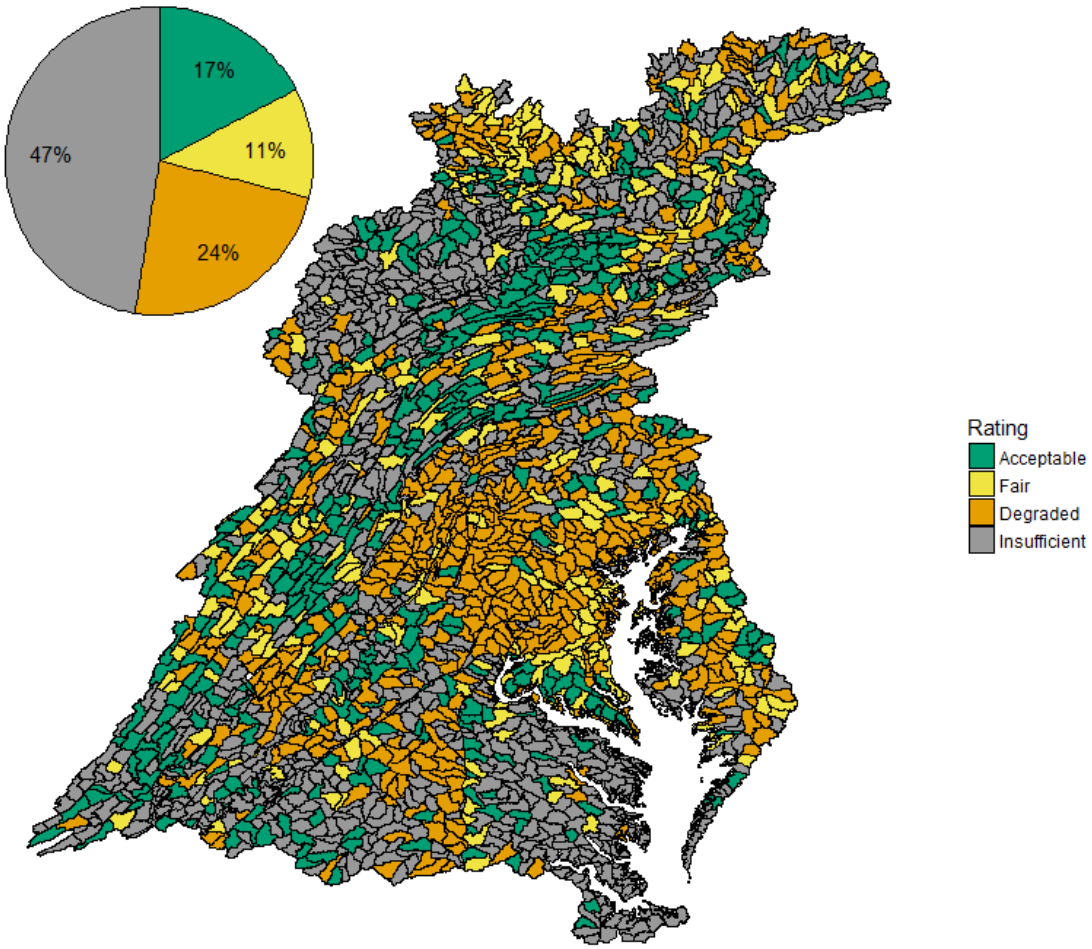

Figure 2. The mean Chessie BIBI score is found for each HUC 12 in the Chesapeake Bay basin and assigned a rating. HUC 12 means are calculated on 1 or more data points. HUC 12's with fewer than 1 are deemed to have insufficient data.

3. Proportional Watershed Rating

Each Chessie BIBI rating in a spatial unit (e.g., HUC 8, HUC 10, HUC 12, catchment) is weighted by a proportion of the spatial unit's total stream miles. For a HUC 12 spatial unit,

the total number of stream miles found in a given HUC 12 is divided by the number of samples it contains. This creates weights of equal sizes for that HUC 12 (Figure 3A). These weights are then applied to each sample rating in the unit. The sum of the stream miles associated with "Acceptable" ratings is divided by the total number of stream miles in the HUC 12 to obtain the proportion streams that are assumed to be Acceptable in the HUC 12; the sum of the stream miles with "Fair" ratings is divided by the total number of stream miles to obtain the proportion of streams that are assumed to be Fair in the HUC 12; and so forth. Stream miles associated with each rating in the HUC 12s also can be summed up to the Chesapeake watershed scale and divided by the total number of stream miles in the watershed to obtain estimates of % Acceptable, % Fair, and % Degraded for that scale (Figure 3B).

Weighting unit cells by stream miles provides a more holistic view of the Chesapeake Bay basin condition by incorporating the stream miles in the basin that cannot be accurately classified (i.e., Insufficient). The method may flip the spatial bias from densely sampled areas to poorly sampled areas because the weight of a few sampling events in poorly sampled areas is magnified. If a single sampling event is considered an accurate representation of the unit cell it is located within, then this spatial bias would be minimal or negligible. It may be beneficial to require a minimum number of sampling events to be present for a unit cell to be included in this calculation. One benefit of this method is it can express stream health as a percent of stream miles.

The method does not provide an accurate spatial representation of the ratings. When there is more than one sample in a cell, then the cell's stream miles are evenly, but arbitrarily divided. The division does not reflect any actual spatial distribution. This does not have large impact when viewing the basin as a whole to identify general areas with a particular rating but it would be erroneous to interpret the results on a cell by cell basis because the position of the color within cell is arbitrary. Figure 3A provides an example of this issue. The left side of the image depicts the sample locations and their Chessie BIBI ratings within a HUC 12. The right side of the image shows an even division of the HUC 12 but the ratings are arbitrarily assigned to the subdivisions and do not correspond with the sampling points.

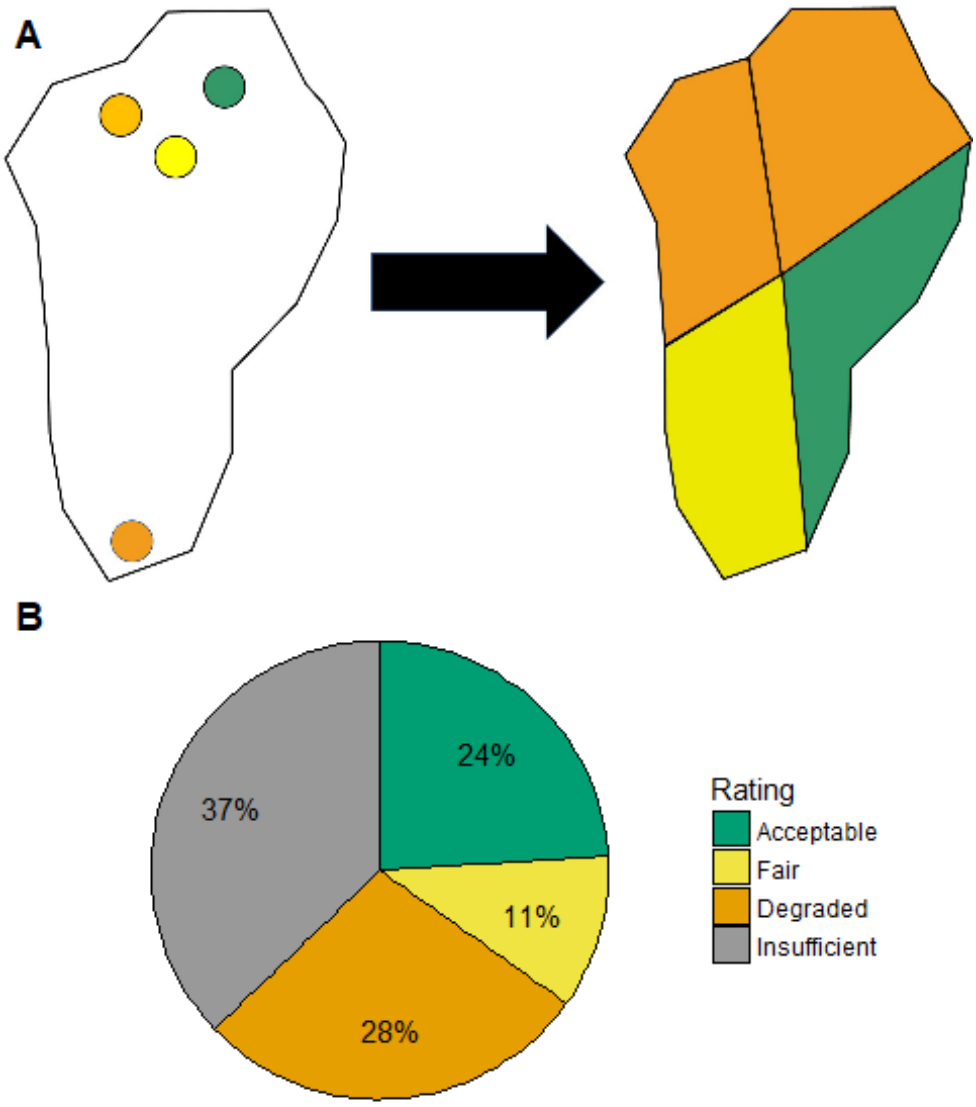

Figure 3. A) The rating proportions calculated for the Chesapeake Bay basin using the area weighted methodology. B) An example of how a HUC 12 is arbitrarily divided into equal parts during the area weighting process.

4. Random Sample

Chessie BIBI ratings are aggregated by a spatial unit, the number of stream miles in each unit cell is determined, and a random selection of one sampling location in each unit cell is used to represent all stream miles in that cell. Figure 4 provides an example of one random sample of HUC12s in the Chesapeake Bay basin. The number of stream miles representing each rating is then summed to the desired spatial scale. The initial unit must cover the entire Chesapeake basin, including areas of the basin without Chessie BIBI ratings. Additionally, the cells in the spatial unit should be approximately the same size, giving each cell an equal probability of being selected during sampling.

This method reduces spatial bias but may result in a large loss of sample size. Small sample sizes may not accurately represent the basin. Also, a single random sample may be a poor representation of its unit cell. For example, 9 points in a unit cell may represent "Acceptable" and 1 point may represent "Degraded." Arguably the cell should probably be rated "Acceptable" but during the random selection it is possible to misrepresent this unit cell by selecting "Degraded." The bootstrapping method in the section aims to reduce several of the issues associated with one random sample of unit cells to represent the Chesapeake Bay basin.

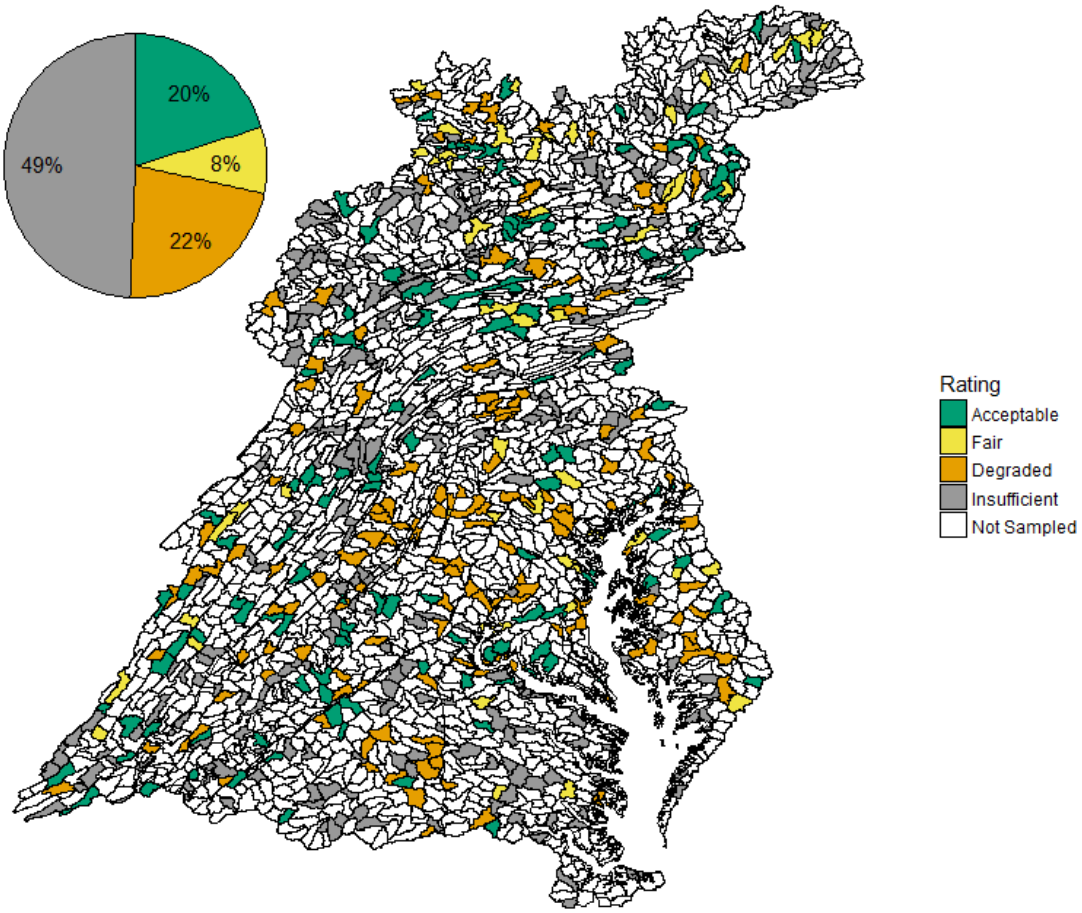

Figure 4. A random selection of HUC 12s in the Chesapeake Bay basin meant to reduce spatial bias and provide an accurate prediction of rating proportions in the basin.

5. Bootstrap

The Chessie BIBI ratings are aggregated by a spatial unit, the total number of stream miles in each unit cell is determined, and bootstrap sampling is applied to each unit cell's

samples. Bootstrap sampling iteratively samples the spatial unit with replacement, choosing one sample from a set number of unit cells in each iteration. The unit must cover the entire basin, including areas of the basin without Chessie BIBI ratings. Additionally, cells in the spatial unit should be approximately the same size, giving each cell an equal probability of being selected during sampling. After many iterations, the bootstrap samples can be summarized to represent the mean number of stream miles associated with each Chessie BIBI rating and provide a standard deviation around each mean.

This method reduces spatial bias and provides a measure of variability around the estimated condition of the basin, which should enable parametric or non-parametric comparisons to be made across designated time periods to detect trends (e.g., 2008 baseline period vs. 2010 period). Preferably each of the cells in the selected spatial unit would be the exact same size to provide equal probability of selecting each cell during random sampling. Additionally, it would be preferable if each cell represented an equal number of stream miles to provide equal weight to each cell. However, the cell size and stream miles represented within each HUC or catchment can vary quite a bit. The variability of cell size and/or stream miles represented by each cell most likely increases the variability of the stream condition estimates. Furthermore, the results from this method cannot be presented as a map of the ratings because this method summarizes many sampling iterations; each iteration is effectively an individual map (See Random Sample section above). The results represent the mean proportions and standard deviations of the ratings for the entire basin, not a mean proportion per unit cell. Figure 5 presents the basic concept that each bootstrap iteration can be plotted as a map but the power of the bootstrap comes from summarizing the many iterations to obtain the mean and standard deviation of each rating.

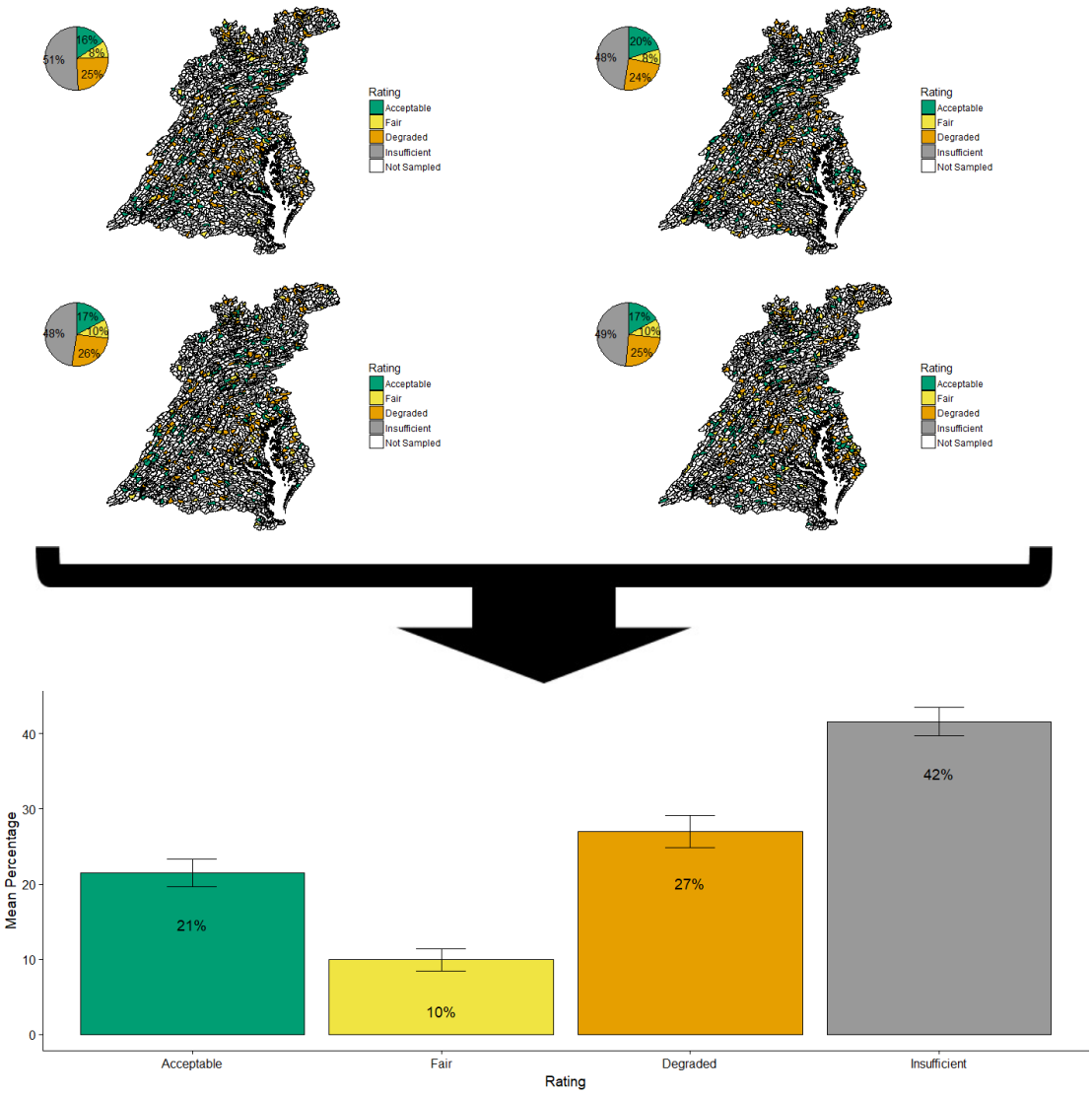

Figure 5. A visualization of the bootstrap process. Each bootstrap iteration represents different HUC 12s throughout the basin resulting in different rating proportions. After sampling many times the average rating proportion and associated variability can be calculated for the basin (depicted as the bar chart with standard deviation bars around the mean).

6. Random Forest

Geospatial predictor variables (e.g., land cover, soil, precipitation, and number of dams) were acquired for the basin. A random forest model was developed to predict Chessie BIBI ratings based on the geospatial predictor variables (Maloney et al. in review). Random forests are an ensemble machine learning technique that incorporates many decision trees to make predictions. The random forest model predicts the Chessie BIBI rating using the geospatial predictor variables for the entire Chesapeake Bay basin, including areas that do not currently contain Chessie BIBI ratings. The underlying data set used for this analysis

was the 1:24,000 high resolution Spatial Hydro-Ecological Decision System, SHEDS, database (www.ecosheds.org, Figure 6). For our initial random forest model, we combined Fair and Acceptable into a single category – FairGood, and built the model to predict either Poor or FairGood conditions.

This method can be used to fill in spatial units (e.g., catchments, HUC12, HUC10) in the basin that do not have Chessie BIBI ratings. It can also estimate Chessie BIBI ratings for the entire Chesapeake Bay basin. The model does not contain an "Insufficient" rating but one can independently rate units as "Uncertain" based on confidence in the model predictions. In the example below, we defined sites as "Uncertain" if their modeled predicted probability of Poor fell ±0.10 an identified optimized cutoff. These units represent a "gray-zone" in the random forest prediction where the Chessie BIBI rating cannot be strongly classified as Acceptable, or Degraded. Areas of uncertainty may be reduced through the exploration of additional geospatial predictor variables but most likely required more sampling or further refinement of the Chessie BIBI indices. This method may be susceptible to compounding error. Each Chessie BIBI index has a classification error. The random forest model is then developed using the Chessie BIBI indices and has its own classification error.

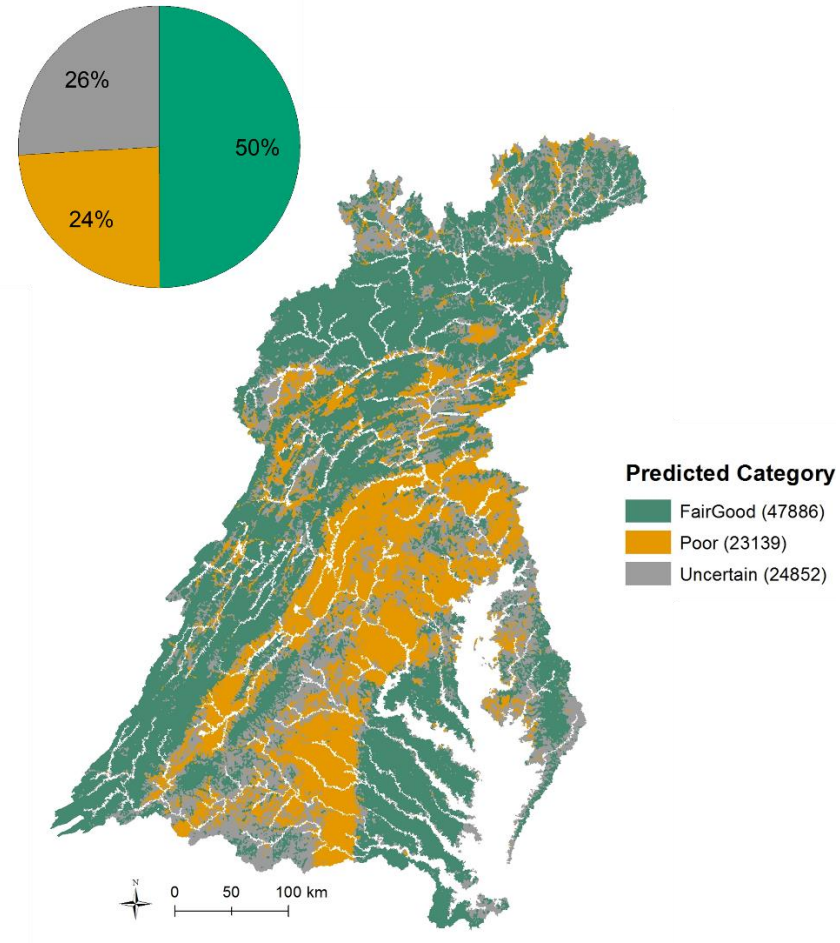

Figure 6. Random forest model predicted ratings in the Chesapeake Bay basin using the 1:24,000 high resolution SHEDS database.