Predicting biological conditions for small headwater streams in the Chesapeake Bay watershed

Kelly O. Maloney\textsuperscript{1,3}, Zachary M. Smith\textsuperscript{2,4}, Claire Buchanan\textsuperscript{2,5}, Andrea Nagel\textsuperscript{2,6}, and John A. Young\textsuperscript{1,7}

\textsuperscript{1}US Geological Survey, Leetown Science Center, 11649 Leetown Road, Kearneysville, West Virginia 25430 USA
\textsuperscript{2}Interstate Commission on the Potomac River Basin (ICPRB), 30 West Gude Drive, Suite 450, Rockville, Maryland 20850 USA

\textbf{Abstract:} A primary goal for Chesapeake Bay watershed restoration is to improve stream health and function in 10% of stream miles by 2025. Predictive spatial modeling of stream conditions, when accurate, is one method to fill gaps in monitoring coverage and estimate baseline conditions for restoration goals. Predictive modeling can also monitor progress as additional data become available. We developed a random forests model to predict biological condition of small streams (<200 km\textsuperscript{2} in drainage) in the Chesapeake Bay watershed. Biological condition was measured with the Chesapeake Bay Basin-wide Index of Biotic Integrity (Chessie BIBI), a stream macroinvertebrate index. Our goal was to predict biological condition in all unsurveyed small streams present in a 1:24,000 scale catchment layer as a 2004–2008 baseline. We reclassified the 5-category Chessie BIBI ratings into two categories, poor and fair/good, to align with management goals of the Chesapeake Bay Program. The model included 12 geospatial predictor variables including measures on spatial location, bioregion, land cover, soil, precipitation, and number of dams in local catchments. We trained the model with a random 75\% subset of Chessie BIBI data (n = 1449), and used the remaining 25\% of Chessie BIBI data (n = 484) as test data. The model performed well, correctly predicting 72\% of samples in training data and 73\% of samples in test data, but model accuracy varied among bioregions. We performed uncertainty analyses by adding bands of either ±0.05 or ±0.10 BIBI units to the cutoff between poor and fair/good. These uncertainty analyses resulted in 14.5\% (±0.05 band) and 24.8\% (±0.10 band) of samples in test data being classified as in uncertain condition. For 95,877 small stream reaches in the Chesapeake Bay watershed, the model predicted 64\% in fair/good condition, the ±0.05 uncertainty analyses predicted 57\% in fair/good condition, and the ±0.10 uncertainty analysis predicted 50\% in fair/good condition. These reported values have different implications for the number of improved stream miles required to meet the goal of improving 10\%. Incorporating uncertainty provides an assessment of model strength as well as confidence in predictions. We, therefore, suggest increased reporting of uncertainty in studies that spatially predict stream conditions.

\textbf{Key words:} Chesapeake Bay, Basin-wide Benthic Index of Biotic Integrity, Chessie BIBI, spatial modeling, stream integrity, stream health, benthic macroinvertebrates

Since Hynes’ 1975 seminal paper that posited “the valley rules the stream”, much research has focused on understanding the relationships between catchment (or landscape) conditions and streams. This research has resulted in a greater understanding of how land use affects streams (i.e. Allan 2004, Walsh et al. 2005). Further, this body of work has culminated in an appreciation of the complex interplay between natural and anthropogenic factors and their effects on streams—namely, that landscape context and history matter (Harding et al. 1998, Merriam et al. 2013). Over the same period, advances in biological assessments have furthered our understanding of the ecological condition of streams (i.e., integrity) and the effect of anthropogenic stressors on those conditions (Davies and Jackson 2006, Hawkins et al. 2010).

Recently, researchers have developed models to predict stream biological condition from landscape variables (Wang et al. 2001, Villeneuve et al. 2015, Waite and Metre 2017). For example, Carlisle et al. (2009) modeled stream conditions for the eastern portion of the United States based on US Geological Survey, National Water Quality Assessment (NAWQA) Program data and used this model to predict biological condition for 552 unsampled basins in southeastern Wisconsin, USA. Further, Maloney et al. (2009) used state collected data to predict stream conditions at unsurveyed sites throughout the Chesapeake Bay portion of Maryland, USA. More recently, Hill et al. (2017) used the US Environmental Protection Agency (US EPA) National Rivers and Streams Assessment (NRSA) data to predict conditions for

E-mail addresses: 3kmaloney@usgs.gov; 4zsmith@icprb.org; 5cbuchan@icprb.org; 6anagel@icprb.org; 7jyoung@usgs.gov

DOI: 10.1086/700701. Received 18 January 2018; Accepted 23 August 2018; Published online 17 October 2018.

Freshwater Science. 2018. 37(4):795–809. © 2018 by The Society for Freshwater Science. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), which permits non-commercial reuse of the work with attribution.

For commercial use, contact journalpermissions@press.uchicago.edu.
all nonintermittent streams in the contiguous USA. Such studies not only produce spatially explicit predictions of unsurveyed sites, but can also highlight areas where a model does not predict observed conditions effectively and, thus, where predictor and response data may need to be collected or reevaluated.

The Chesapeake Bay watershed in the USA has experienced extensive landscape modification by human activity. Since 1950, the human population within the Chesapeake Bay watershed has more than doubled. It is currently over 18 million and is expected to increase to 20 million by 2030 (Chesapeake Bay Program 2017a). This population increase has resulted in high levels of watershed development (10.5 and 11.0% developed land cover in 2006 and 2011, respectively; National Land Cover Database [NLCD], https://www.mrlc.gov/; Fry et al. 2011, Homer et al. 2015). Moreover, a large portion of the watershed is used for agriculture (24.9 and 24.5% cover of the watershed in 2006 and 2011, respectively; NLCD). These anthropogenic stressors have resulted in a high number of biologically degraded streams. The US EPA estimates the percentage of poor stream conditions for the ecoregions within the watershed as 44, 62, and 69% in the Northern Appalachian, Southern Appalachian, and Coastal Plains ecoregions, respectively (USEPA 2016). Maloney et al. (2009) predicted that for the subarea of the Chesapeake Bay watershed within Maryland, 36% of streams were in very poor or poor condition. The projected increase in the human population of 2 million by 2030 in the watershed could exert additional anthropogenic stress on streams and increase the number of streams in degraded condition. Moreover, some areas of the watershed with relatively low anthropogenic stress are currently being developed for shale gas (Maloney et al. 2018), which could degrade even more streams.

A 2025 restoration goal for the entire Chesapeake Bay watershed is to improve stream health and function for 10% of stream miles above a 2008 baseline for the watershed (Chesapeake Bay Program 2017b). Currently, however, no published baseline exists with which to gauge improvement. Results from the US EPA National Rivers and Streams Assessment program could be used to identify a baseline condition for streams in the Chesapeake Bay watershed, but that survey (USEPA 2016) focused on a national assessment and was not tailored to the Chesapeake Bay watershed. Maloney et al. (2009) focused on a small subarea of the entire Chesapeake Bay, so their results are not applicable to the entire watershed.

For over a decade, research has focused on developing and improving a stream macroinvertebrate index of biotic integrity for the entire Chesapeake Bay watershed (Chesapeake Basin-wide Index of Biotic Integrity [Chessie BIBI], Smith et al. 2017). This work has culminated in an extensive database that includes standardized measures of benthic macroinvertebrate data from 21,343 stream samples. A family-level, region-scale BIBI generated from these data found that 49.5% of samples were in very poor or poor condition (Smith et al. 2017). When sample ratings were weighted by the proportion of the local subwatershed area (hydrologic unit code 12 [HUC12]) they represent, the results indicated that stream health is likely very poor or poor in 39.5% of the Chesapeake watershed area, and fair, good, or excellent in 49.2% of the watershed area. Incomplete coverage prevented a rating determination for 11.3% of the watershed area.

The Chessie BIBI had higher sampling frequencies in urban counties, and therefore unweighted sums of the ratings do not equally represent the entire watershed. The index also did not provide insight into unsurveyed reaches. Here, our first objective was to build a predictive model of stream integrity based on relationships between Chessie BIBI values and both natural landscape attributes and landuse cover. Our second objective was to use this model to predict conditions for all small streams (<200 km² in upstream drainage area) in the Chesapeake Bay watershed. We also incorporated uncertainty into the predictions to better support management decisions. Predictive models are most useful when they are developed at spatial grains at which management occurs and at spatial extents that match observed variation in the ecological endpoints of interest. Here, we wanted to create management relevant models applicable to headwater streams, so we built a model at the 1:24,000 scale that covered the entire Chesapeake Bay watershed.

**METHODS**

**Study site**

The Chesapeake Bay watershed drains approximately 168,000 km² of land from the states of Maryland, Virginia, West Virginia, Pennsylvania, New York, and Delaware as well as the District of Columbia (Fig. 1). The James, Potomac, Rappahannock, Susquehanna, and York are the largest rivers draining the watershed, and empty into the Chesapeake Bay—the largest estuary in the US (Chesapeake Bay Program 2017a). The watershed contains the major cities of Baltimore, Maryland; Washington, DC; Harrisburg, Pennsylvania; and Richmond, Virginia.

**Chessie BIBI**

The Chessie BIBI was developed from a database that includes stream habitat scores, water-quality data, and stream macroinvertebrate raw counts from 28 state, county, regional, federal, and other monitoring programs throughout the Chesapeake Bay watershed (n = 21,343). Smith et al. (2017) describe development of the BIBI in detail, and we provide only a brief summary here. The Chessie BIBI database includes data collected from 1st- to 4th order streams, based on NHDPlusV2 1:100,000 scale data, between 1992 and 2015, but does not include data collected in winter months (January, February, and December). The database
contains consistently collected habitat and water-quality data that can be used to identify degraded and reference samples. Smith et al. (2017) evaluated over 200 commonly used benthic macroinvertebrate metrics (e.g., Barbour et al. 1999) for their ability to discriminate between degraded and reference samples. They then standardized macroinvertebrate samples among sampling programs by resampling macroinvertebrate data to counts of approximately 100 individuals per sample prior to calculating richness and diversity metrics. Metric values were standardized on a continuous gradient from 0 to 100 and the mean of these values produced the final index scores. Sample index scores were assigned narrative ratings of very poor, poor, fair, good, or excellent, based on thresholds derived from the 10th, 25th, and 50th percentiles of the index scores of reference samples.

Smith et al. (2017) developed 9 versions of the Chessie BIBI based on 3 taxonomic resolutions (order, family, genus) and 3 spatial extents (watershed, region, bioregion). The regional BIBI separated the watershed into 2 areas (coastal and inland), for which 2 independent indices were developed. The coastal extent was defined by the middle Atlantic coastal plain and the southeastern plains level III ecoregions (Woods et al. 1999), whereas the inland region extent represented the remaining level III ecoregions in the watershed. Chessie BIBI bioregions represent level III ecoregions (Woods et al. 1999) or modifications of level III ecoregions meant to better encompass natural (undisturbed) variability in the benthic macroinvertebrate assemblages (Smith et al. 2017). Modification to the level III ecoregions included dissecting the Northern Piedmont and the Ridge and Valley level III ecoregions by major drainages (e.g., Susquehanna) and physiographic province (e.g., Great Valley). Twelve bioregions defined the Chesapeake Bay watershed, and an independent index was developed for each bioregion.

**Predictor data**

We used several natural and landuse related predictors. We obtained sample coordinates (latitude and longitude) and the dominant bioregion upstream of a sample site from the Chessie BIBI database. We used these 3 predictors because they allowed us both to incorporate effects of naturally occurring spatial factors on index scores and assess interactions with different types of land uses. We also used...
predictor data from the Spatial Hydro-Ecological Decision System (SHEDS, www.ecosheds.org) database. The SHEDS database is at the 1:24,000 map scale and has local catchment, upstream catchment, and riparian zone summaries of data including land use (NLCD 2006), natural watershed characteristics (drainage area, elevation, slope), climate (precipitation, air temperature), dams, and soils (% sandy, hydrologic group). We used the following specific predictors from the SHEDS database: impervious surface, tree canopy cover, agriculture, basin elevation and slope, minimum and maximum air temperature, precipitation, number of dams, atmospheric deposition of NO$_3^-$ and SO$_4^{2-}$, % sandy soils, and soil hydrologic groups. To reduce redundancy of predictor variables and to simplify the model, we removed collinear variables from the SHEDS database with correlation coefficients of $r > 0.70$ (Dormann et al. 2013). For number of dams we preferentially kept the number in the local catchment over the number in upstream catchments because we assumed the effects of dams on habitat and, thus, benthic macroinvertebrates would lessen with distance (Arle 2005, Fenc1 et al. 2015). In all other cases where we removed variables, we preferentially kept upstream catchment variables rather than local or riparian variables because 1) local stream conditions are influenced by cumulative upstream catchment conditions (Scott et al. 2002) and 2) the coarse resolution used to characterize local predictors (e.g., NLCD, 30 m) can lead to inaccurate representation of conditions at riparian buffer scales. We, therefore, removed all riparian measures and all local measures, except the number of dams in the local catchment, because they were highly correlated with upstream analogs. Additionally, upstream minimum and maximum air temperature and slope were correlated with upstream elevation. We retained elevation in the model, because it is a surrogate for air temperature, slope, and stream temperature, the latter of which we did not have estimates. We also removed upstream agriculture cover because it was correlated with upstream tree canopy cover. We retained tree canopy cover over upstream agriculture cover, because it may allow tracking of potential improvements in stream condition associated with future reforestation. Of the 5 correlated hydrologic soil measures, we used group D plus subgroups of groups A, B, or C with high water tables, because these soil classes indicate high runoff potential and low infiltration and water-transmission rates.

The final list of variables used in the model included 8 naturally occurring predictors (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; and dominant upstream bioregion) and 4 landuse-related predictors (upstream SO$_4^{2-}$ deposition, mean % upstream land surface cover as impervious surface, mean % upstream land surface in tree canopy cover, and number of dams within the local catchment (Table 1).

Data set development

For this study, we used the family-level Chessie BIBI developed at the bioregion-scale to optimize the tradeoff between taxonomic resolution (finer resolution contains more ecological information) and sample size (fewest samples were available at the genus-level). We used the bioregion-scale index because we needed an index that adjusted for the natural effects that biogeography, topography, and geology can have on benthic macroinvertebrate assemblages and developed IBIs (Feminella 2000, Maxted et al. 2000, Pond et al. 2013). This version of the index had been previously shown to correctly identify reference and degraded conditions (classification efficiency) between 70.4 and 90.0% depending on bioregion (Smith et al. 2017).

The Chessie BIBI numeric scores are assigned to 1 of 5 categories (very poor, poor, fair, good, excellent), and the Chesapeake Bay Program considers samples classified as fair, good, or excellent as desirable outcomes for stream restoration efforts (Chesapeake Bay Program 2015). To make model results better match management and restoration goals, we, therefore, reclassified the original 5 categories into 2 broader categories of poor (very poor or poor) and fair/good (fair, good, or excellent).

We used a subset of Chessie BIBI data samples, because the time periods over which macroinvertebrate and predictor variable (SHEDS) data were collected differed. The Chessie BIBI was based on data collected between 1992 and 2015, but the SHEDS data were based on the 2006 NLCD. We, therefore, selected BIBI data that were sampled between 1 January 2004 and 31 December 2008 to make the data sets more comparable. This data selection process yielded 8118 samples, of which 5079 were concentrated in Maryland and Fairfax County, Virginia (see Fig. S1). We then spatially linked each Chessie BIBI sample to a SHEDS catchment. We removed Chessie BIBI samples with upstream drainage areas $\geq$200 km$^2$. When a SHEDS catchment had more than one Chessie BIBI sample, we randomly selected one of the samples. Of the 3039 samples located in the non-Maryland and Fairfax County area of the watershed, 1649 were unique to a SHEDS catchment with $<200$ km$^2$ upstream drainage. Of the 5079 Chessie BIBI samples in Maryland and Fairfax County, 2727 were unique to a SHEDS catchment with $<200$ km$^2$ upstream drainage. To remove the potential for the spatial clustering in Maryland and Fairfax County to affect the model, we subset these data to match the spatial density of unique samples in the rest of the Chesapeake Bay watershed (142,327 km$^2$ with 1649 samples, or a density of ~0.012 samples per km$^2$). We, therefore, randomly selected 285 samples from Maryland and Fairfax County (area = 24,604 km$^2$) and combined these data with the remainder of the Chesapeake Bay data set (total n = 1933; following removal of 1 point that was found to be a duplicate for a SHEDS catchment). The resultant data set had 907 samples in the poor category and 1026 in the fair/
good category (Fig. 1) all of which were linked to predictor data.

Modeling approach

We built a random forests (Breiman 2001) model (randomForest R package, Liaw and Wiener 2002) to predict stream condition, classified as either poor or fair/good, from both the naturally occurring and landuse variables. Random forests offer several characteristics desirable for predictive modeling in ecology—most notably, these models account for nonlinear relationships as well as complex interactions among predictors (Cutler et al. 2007). Random forests models can also use both continuous and categorical predictors. Moreover, random forests have high classification accuracy compared with other classifying techniques (Prasad et al. 2006, Fernández-Delgado 2014). For this model, we set the number of trees to 1000 and the number of variables randomly sampled as predictors at each split (mtry) to 4 (the optimal number identified during model tuning).

Model calibration and validation

We used both training data and independent validation (test) data to assess model performance. The training data consisted of 75% of the samples ($n = 1449$) that were randomly selected from the data set, and the remaining 25% ($n = 484$) of samples were used as test data. Random forests uses out-of-bag (i.e., withheld) observations of the training data to assess model performance and is, therefore, difficult to overfit. We characterized model accuracy with 5 metrics: overall proportion of samples correctly classified (PCC), proportion of poor samples correctly classified (poor PCC), proportion of fair/good samples correctly classified (fair/good PCC), area under the receiver operation curve (AUC), and the $\kappa$ statistic (SDMtools R package, VanDer-
Wal et al. 2014). We conducted all statistical tests and analyses with the R Statistical Software Package, version 3.4.0 (R Development Core Team 2017). We optimized the cutoff threshold between poor and fair/good classes with the Youden Index (also known as the True Skill Statistic). We used the OptimalCutpoints R package (López-Ratón et al. 2014) to calculate the Youden Index, which maximizes (sensitivity + specificity − 1) (Freeman and Moisen 2008). We calculated accuracy statistics for both the training and test data sets. For the test data we used the Youden Index cutoff ±0.05 and ±0.10 as estimates of uncertainty. Application of these thresholds allowed us to identify a 3rd condition class (uncertain), which included values between the poor and fair/good categories.

**Model interpretation**

We used partial dependence plots to examine the association between stream condition and each predictor variable after controlling for all other predictors (Hastie et al. 2001, Cutler et al. 2007). We also used these plots to determine if the relationship between stream condition and each predictor variable was interpretable and consistent with published research. We also tested if each predictor variable differed among poor PCC, fair/good PCC, misclassified poor samples, misclassified fair/good samples, and uncertain samples in the test data with the ±0.10 uncertainty band with a Kruskal–Wallis test ($p = 0.0045$, Bonferroni adjusted for 11 tests). A significant result may indicate that specific predictors are driving misclassified and uncertain classified predictions. We then used a Dunn’s test of multiple comparisons ($p = 0.005$, Bonferroni adjusted for 10 comparisons) to test for group differences for predictors with a significant Kruskal–Wallis test.

**Model predictions**

We used the model to predict the condition of 95,877 stream segments. The SHEDS database included information on 105,417 catchments within the Chesapeake Bay watershed, but we excluded data for 8634 (8.2%) catchments with upstream drainage areas ≥200 km$^2$. We also excluded data for 906 catchments that were missing one or more predictors.

**RESULTS**

**Spatial representation**

The range of candidate predictor values across samples in the Chessie BIBI database represented the values present in the Chesapeake Bay watershed as a whole (i.e., the SHEDS database, Table 1). For example, upstream catchment area in the Chessie BIBI data set ranged from 0.8 to 198.6 km$^2$ (median = 16.8 km$^2$) and across the Chesapeake Bay watershed ranged from 0.1 to 199.9 km$^2$ (median = 3.0 km$^2$). Further, median upstream impervious surface cover in the Chessie BIBI data set (0.6%) approximated the median value for the Chesapeake (0.3%).

**Model performance**

Overall, the model performed moderately well. The random forests model had an out-of-bag error rate of 28% on the training data set. The Youden Index optimized cutoff was 0.48. The 6 most important variables in order of importance were upstream impervious cover, bioregion, upstream tree canopy cover, latitude, longitude, and elevation. The probability of a catchment being in poor condition rapidly increased with upstream impervious cover up to 10% (Fig. 2A), and decreased with increasing upstream tree canopy cover almost linearly up to 80% (Fig. 2C). Blue Ridge (BLUE), Lower-Northern Piedmont (LNP), Piedmont (PIED) and Southern Great Valley (SGV) bioregions had a higher probability of poor condition than the other bioregions (Fig. 2B). The probability of poor condition was highest for mid-latitude (Fig. 2D), more western (Fig. 2E), lower elevation (Fig. 2F) catchments.

At the 0.48 cutoff, the random forests model had an overall PCC of 0.72, correctly predicted poor conditions in 73% of cases, and correctly predicted fair/good condition in 72% of cases in the training data set. The model had an AUC of 0.80 and a $\kappa$ statistic of 0.45 for the training data, which indicates moderate accuracy (Landis and Koch 1977).

The predictive ability of the model on the training data varied among bioregions. For the Blue Ridge (BLUE), Central Appalachians (CA), and Northern Ridge and Valley (NRV) bioregions, both poor and fair/good were predicted better than chance (both $>0.60$ PCC, Table 2). For the LNP, PIED, SGV, and Upper-Northern Piedmont (UNP) the model predicted poor more than twice as accurately as fair/good. For the remaining bioregions, the model predicted fair/good 1.3 (NAPU) to 2.8 (MAC) more accurately than poor conditions (Table 2).

The model had an overall PCC of 0.73 for the test data set and a 0.48 cutoff threshold, with a poor PCC of 0.72 and a fair/good PCC of 0.74 (Table 2), which resulted in an AUC = 0.81 and a $\kappa$ = 0.45 (moderate accuracy). The model predicted bioregion test data with similar accuracy as it predicted the training data, except for the LNP bioregion where the model predicted the fair/good categories more accurately (PCC > 0.60), and for the NRV bioregion where the model had a lower accuracy in predicting the poor category (PCC < 0.60). When we removed samples within the ±0.05 uncertainty band (70 Uncertain samples, or 14.5% of samples), poor PCC increased to 0.76 and fair/good PCC increased to 0.79, resulting in an overall PCC = 0.78, a $\kappa$ value = 0.55, and an AUC = 0.84. When we removed samples within the ±0.10 uncertainty band (120 uncertain samples)
ples, or 24.8% of samples), poor PCC increased to 0.77 and fair/good PCC increased to 0.82, resulting in an overall PCC $= 0.80$, an AUC $= 0.85$, and a $\kappa$ value $= 0.59$ (high end of moderate agreement, Landis and Koch 1977).

Among correctly classified, incorrectly classified, and uncertain samples, predictor variables were only found to differ for impervious cover, tree canopy cover, elevation, and sandy soils (Fig. 3A–K). Samples correctly predicted as fair/good had significantly less impervious cover than fair/good samples incorrectly classified as poor, samples correctly classified as poor, and uncertain samples (Fig. 3A). Correctly classified poor samples also had higher levels of impervious cover than poor samples misclassified as fair/good and uncertain samples, and misclassified fair/good samples also had higher impervious cover than misclassified poor samples. Correctly predicted fair/good samples also had a significantly higher levels of upstream tree cover than misclassified fair/good samples, correctly classified poor samples, and uncertain samples (Fig. 3B). Correctly classified poor samples had lower levels of tree cover than misclassified poor samples and uncertain samples (Figs. 3E, 3J), and were higher in elevation than misclassified fair/good samples.

For the Chesapeake Bay watershed as a whole, at a cutoff probability of 0.48, the model predicted 64% of the catchments to be in fair/good condition, and 36% of catchments to be in poor condition (Table 3). The bioregions with the highest proportion of fair/good catchments were the NAPU, NCA, SEP, and SRV, and the bioregions with the highest proportion of predicted poor catchments were the LNP, PIED, SGV, and UNP (Table 3, Fig. 4A). When the model included an uncertainty band of $\pm 0.05$ around the 0.48 cutoff, it predicted that 57% of catchments were in fair/good condition with 13% in uncertain condition (Table 3). When the model included a band of $\pm 0.10$, it predicted that 50% of catchments were in fair/good condition with 26% of catchments in uncertain condition. Patterns across bioregions with the $\pm 0.05$ and $\pm 0.10$ uncertainty band analyses were similar to those predicted without uncertainty estimates. Catchments of uncertain status occurred most frequently in the upper, middle, and far eastern portions of the watershed (Figs. 4B, 4C). For both uncertainty bands, the CA, LNP, MAC, and PIED bioregions had the highest number of catchments classified as being in uncertain condition, and those bioregions with
Table 2. Confusion matrices and proportion correctly classified (PCC) for the training, test, and test data sets with uncertainty (Unc). Unc is the number of samples predicted to be in uncertain condition defined as probabilities ±0.05 or ±0.10 around 0.48 (i.e., had probabilities 0.43–0.53 or 0.38–0.58, respectively).

| Bioregion                  | Observed category | Predicted category | Training PCC | Test PCC | Test 0.05 uncertainty band PCC | Test 0.10 uncertainty band PCC |
|----------------------------|-------------------|--------------------|--------------|----------|--------------------------------|--------------------------------|
| BLUE Blue Ridge            | Poor              | 40                 | 0.80         | 11       | 2                              | 0.85                           |
|                            | Fair/good         | 15                 | 0.68         | 5        | 8                              | 0.62                           |
|                            |                   |                    |              |          |                                |                                |
| CA Central Appalachians    | Poor              | 11                 | 0.65         | 4        | 0                              | 1.00                           |
|                            | Fair/good         | 3                  | 0.86         | 2        | 7                              | 0.78                           |
|                            |                   |                    |              |          |                                |                                |
| LNP Lower-Northern Piedmont| Poor              | 70                 | 0.85         | 20       | 4                              | 0.83                           |
|                            | Fair/good         | 25                 | 0.24         | 2        | 9                              | 0.82                           |
|                            |                   |                    |              |          |                                |                                |
| MAC Middle Atlantic Coastal Plain | Poor   | 7                  | 0.30         | 0        | 7                              | 0.00                           |
|                            | Fair/good         | 7                  | 0.85         | 0        | 18                             | 1.00                           |
|                            |                   |                    |              |          |                                |                                |
| NAPU Northern Appalachian Plateau and Uplands | Poor | 45                 | 0.54         | 8        | 11                             | 0.42                           |
|                            | Fair/good         | 37                 | 0.70         | 14       | 31                             | 0.69                           |
|                            |                   |                    |              |          |                                |                                |
| NCA North Central Appalachians | Poor  | 9                  | 0.45         | 1        | 1                              | 0.50                           |
|                            | Fair/good         | 8                  | 0.87         | 1        | 24                             | 0.96                           |
|                            |                   |                    |              |          |                                |                                |
| NRV Northern Ridge and Valley | Poor | 45                 | 0.61         | 12       | 13                             | 0.48                           |
|                            | Fair/good         | 19                 | 0.82         | 7        | 32                             | 0.82                           |
|                            |                   |                    |              |          |                                |                                |
| PIED Piedmont              | Poor              | 87                 | 0.93         | 21       | 0                              | 1.00                           |
|                            | Fair/good         | 23                 | 0.41         | 15       | 8                              | 0.35                           |
|                            |                   |                    |              |          |                                |                                |
| SEP Southeastern Plains    | Poor              | 20                 | 0.57         | 4        | 6                              | 0.40                           |
|                            | Fair/good         | 9                  | 0.86         | 6        | 16                             | 0.73                           |
|                            |                   |                    |              |          |                                |                                |
| SGV Southern Great Valley  | Poor              | 84                 | 0.95         | 32       | 1                              | 0.97                           |
|                            | Fair/good         | 15                 | 0.29         | 3        | 0                              | 0.00                           |
|                            |                   |                    |              |          |                                |                                |
| SRV Southern Ridge and Valley | Poor | 15                 | 0.36         | 6        | 8                              | 0.43                           |
|                            | Fair/good         | 12                 | 0.90         | 3        | 44                             | 0.94                           |
|                            |                   |                    |              |          |                                |                                |
| UNP Upper-Northern Piedmont | Poor   | 82                 | 0.85         | 25       | 4                              | 0.86                           |
|                            | Fair/good         | 37                 | 0.39         | 15       | 13                             | 0.46                           |
|                            |                   |                    |              |          |                                |                                |
| All                        | Poor              | 516                | 0.73         | 144      | 57                             | 0.72                           |
|                            | Fair/good         | 210                | 0.72         | 73       | 210                            | 0.74                           |
DISCUSSION

Until the creation of the Chessie BIBI, management of stream conditions for the entire Chesapeake Bay watershed was hampered by lack of a uniform data set to establish baseline conditions. We used a model to predict condition classes derived from the Chessie BIBI for all small, headwater stream reaches in the watershed at the 1:24,000 scale, a scale more relevant to management decisions and more reflective of headwater streams (Fig. S2) than used for previous modeling efforts. Previously identified predictor variables of stream condition were important in the model and model strength varied among bioregions. Uncertainty analyses revealed that including a ±0.10 point uncertainty band reduced the watershed-scale estimate of fair/good conditions by 13.7%. This uncertainty is larger than the 10% improvement goal of stream miles for the watershed, and such uncertainty will obscure detecting future biological improvement. Future research could be aimed at reducing this uncertainty, either from improvements in the Chessie BIBI, inclusion of additional predictors, or by using a multiparameter assessment of stream condition that includes other biological measures (e.g., periphyton or fish) or physio-chemical measures (e.g., nutrients and sediment).

Landcover variables are often strong predictors of stream condition (Allan 2004). Impervious surface, in particular, is frequently associated with stream impairment (Schueler et al. 2009) and was the most important variable in our model. The probability of a site being predicted in poor condition increased rapidly after an increase in more than 8–10% cover, which is consistent with previous studies that showed biological condition declines when impervious surface is >5 to 15% of total upstream cover (Wang et al. 2000, Paul and Meyer 2001, Walsh et al. 2005). The partial dependence plot also showed a rapid increase in the probability of a
poor condition prediction between 0 and ~2% impervious cover, which agrees with studies conducted within Maryland (King et al. 2011, Maloney et al. 2012). Tree canopy cover also was an important land-cover variable in our model and is associated with minimally disturbed landscapes. In the Chesapeake watershed, tree canopy cover largely indicates absence of agriculture and impervious cover. Thus, it was not surprising that the probability of predicting a site as being in poor condition decreased with increasing tree canopy cover (up to 80%). Differences in impervious and tree canopy covers between correctly classified and misclassified samples may indicate that the Chessie BIBI did not effec-

Table 3. Number of the 95,877 SHEDS catchments in each bioregion and the percent (%) of them predicted to be in good/fair condition without and with ±0.05 and ±0.10 uncertainty bands. Bioregion abbreviation descriptions are listed in Table 2. Unc = uncertain.

| Bioregion | Total | % fair/good | Poor | Fair/good | Unc | % fair/good | % Unc | Poor | Fair/good | Unc | % fair/good | % Unc |
|-----------|-------|-------------|------|-----------|-----|-------------|------|------|-----------|-----|-------------|------|
| BLUE      | 4026  | 57.4        | 1401 | 2000      | 625 | 49.7        | 15.5 | 1183 | 1688      | 1155 | 41.9        | 28.7 |
| CA        | 3258  | 62.4        | 908  | 1668      | 682 | 51.2        | 20.9 | 591  | 1325      | 1342 | 40.7        | 41.2 |
| LNP       | 7065  | 29.5        | 4270 | 1359      | 1436| 19.2        | 20.3 | 3539 | 722       | 2804 | 10.2        | 39.7 |
| MAC       | 6697  | 70.0        | 1200 | 3693      | 1804| 55.1        | 26.9 | 607  | 2738      | 3352 | 40.9        | 50.1 |
| NAPU      | 13,414| 80.2        | 1733 | 9634      | 2047| 71.8        | 15.3 | 1064 | 8377      | 3973 | 62.4        | 29.6 |
| NCA       | 6420  | 94.5        | 258  | 5937      | 225 | 92.5        | 3.5  | 196  | 5697      | 527  | 88.7        | 8.2  |
| NRV       | 11,143| 68.0        | 2872 | 6910      | 1361| 62.0        | 12.2 | 2264 | 6222      | 2657 | 55.8        | 23.8 |
| PIED      | 10,007| 23.5        | 6645 | 1445      | 1917| 14.4        | 19.2 | 5616 | 748       | 3643 | 7.5         | 36.4 |
| SEP       | 9906  | 87.6        | 966  | 8326      | 614 | 84.1        | 6.2  | 734  | 7870      | 1302 | 79.4        | 13.1 |
| SGV       | 4197  | 13.4        | 3384 | 401       | 412 | 9.6         | 9.8  | 2986 | 311       | 900  | 7.4         | 21.4 |
| SRV       | 13,502| 92.1        | 779  | 12,044    | 679 | 89.2        | 5.0  | 538  | 11,505    | 1459 | 85.2        | 10.8 |
| UNP       | 6242  | 22.2        | 4365 | 990       | 887 | 15.9        | 14.2 | 3821 | 683       | 1378 | 10.9        | 27.8 |
| All       | 95,877| 63.6        | 28781| 54,407    | 12,689| 56.7      | 13.2 | 23,139| 47,886    | 24,852| 49.9        | 25.9 |

Figure 4. Spatial representation of predicted integrity scores for catchments without uncertainty bands (A), with a ±0.05 uncertainty band (B), and a ±0.10 uncertainty band (C). Numbers in parentheses indicate the number of catchments in each condition category. White areas represent catchments not modeled because they had upstream drainage area ≥200 km².
tively distinguish reference from degraded condition, and calibrating the Chessie BIBI with these landuse metrics may improve the classification efficiency of the Chessie BIBI, as well as our model performance.

Bioregion was the 2nd most important predictor of biological condition in our model, and its importance as a predictor of stream condition is supported by previous research (Maloney et al. 2009). Chessie BIBI bioregions represent level III ecoregions or modifications of level III ecoregions. The importance of bioregions in our model may, therefore, account for missing predictors that encompass natural or anthropogenic factors at the landscape-, reach- (colonization sources), local- (sample site), or population-level (interactions) (Abell et al. 2008, Domisch et al. 2015). Several of these factors may be hydrology, topography, or climate (Smith et al. 2017). Identifying the specific component (e.g., mean average flow) and scale (e.g., local habitat) of each main factor would improve our understanding of why biological conditions differ across streams in the Chesapeake Bay watershed and inform possible scale-relevant management actions.

Other important predictors in our model included latitude, longitude, and elevation, which, like bioregion, act as surrogates for other natural factors or anthropogenic factors that may be more management relevant. Latitude and longitude may indicate some missing factor that covaries with stream condition across the watershed, such as winter salting of developed roads and nutrient and sediment runoff from agricultural lands. Although we did not have data to effectively test surrogate variable explanations, incorporating them will likely improve model performance. For example, Merriam et al. (2018) found that stream water temperature, was a highly important variable in a Brook Trout occupancy model for the northern portion of the Chesapeake Bay watershed. In our model, we used elevation as a temperature surrogate, which may explain its relevance for biological condition.

Dams were not important in our model, even though they are known to have a strong influence on the biota that inhabit streams and rivers (Poff and Hart 2002). The lack of apparent influence of dams in our study may be a result of our focus on small, headwater streams, where there were relatively few dams—out of 91,788 catchments, only 4.3% had local dams. Additionally, in small, headwater streams the effects of dams may be more localized, with minimal influence on colonization (Tonkin et al. 2014). Alternatively, dams may not influence macroinvertebrate assemblages in the Chesapeake Bay watershed as observed elsewhere. Others have reported minimal effects of dams on flow patterns in portions of the Chesapeake Bay watershed (USACE-TNC-ICPRB 2014).

Predicted stream conditions varied widely among bioregions. For example, NCA, SRV, and NAPU all had greater than 80% of catchments classified as fair/good, likely a result of these 3 bioregions having relatively low median values (0.02–0.21%) for upstream impervious surface cover (Table S1). In contrast, LNP, SGV, and UNP had the lowest proportion of samples classified as fair/good, and had the highest median values for upstream impervious surface (0.91–1.86%). However, such patterns between more impervious surfaces and lower biological condition were not consistent across bioregions. SEP had >80% and MAC had 70% of catchments classified as fair/good, but both had intermediate levels of upstream impervious surfaces (0.61 and 0.69%, respectively), suggesting other landscape variables were also important. We urge caution when assessing bioregion patterns because in initial Chessie BIBI development, performance varied by bioregion (Smith et al. 2017). For example, the Chessie BIBI scores in the NAPU were least able to distinguish reference and degraded samples (classification efficiency = 70.4), whereas the Chessie BIBI had a classification efficiency of 90.0 in the UNP. These differences in performance most likely resulted from differences among bioregions in the number of reference and degraded samples. Some bioregions had a large imbalance of reference and degraded samples (e.g., MAC reference n = 17 and degraded n = 257, BLUE reference n = 133 and degraded n = 7), and Smith et al. (2017) concluded that additional data are required to further refine and validate these indices. Improvements in the performance of each bioregion index will likely, in turn, improve the performance of our predictive model.

Here, we opted to develop a single model for the entire Chesapeake Bay watershed rather than develop individual models for each bioregion. At a contiguous USA scale, separate regional models have outperformed a single national model for predicting biological condition (Hill et al. 2017). Thus, separate bioregion models may improve our results. With our current data set, however, small sample sizes limited our ability to build robust models for several bioregions (e.g., CA, Table 2). Expanding the data set to include all of the Chessie BIBI data instead of just 4 years may have enabled separate bioregion models, but the data set would have spanned a long time period. Such a long record of data could result in issues with temporal consistency of predictors, such as land use data from 2006 not being relevant to all Chessie BIBI data. Moreover, data limitations would still occur for some bioregions (e.g., BLUE, CA, and MAC).

Our model was moderately accurate based on PCC (0.72 for training data and 0.73 for test data), sensitivity (poor PCC of 0.73 and 0.72) and specificity (fair/good PCC of 0.72 and 0.74). These accuracy statistics align closely with those from a national predictive model based on the 2008–2009 US EPA NRSA data (final model PCC = 0.75, sensitivity = 0.73, specificity = 0.77; Hill et al. 2017). Our model accuracy is also similar to models used to predict biological conditions in Californian catchments (75 and 78% correct classification of impaired and unimpaired samples, respectively; Brown et al. 2012), eastern highlands and lowlands...
of the USA (highlands: PCC = 0.87, specificity = 0.51, sensitivity = 0.95; lowlands: PCC = 0.76, specificity = 0.57, sensitivity = 0.87; Carlisle et al. 2009), and for streams in France (test data: PCC Good = 0.76–0.86, PCC Bad = 0.16–0.61; Villeneuve et al. 2015).

In this study we opted for parsimony by incorporating 12 uncorrelated predictor variables, so it is encouraging to see that our model was as accurate as these other efforts. Additional predictor data for the Chesapeake Bay watershed, including data on pesticides, herbicides, and septic systems, is becoming available (Gordon et al. 2017) but are currently aggregated at the 1:100,000 scale. Aggregating these variables at the 1:24,000 scale (as in the SHEDS database predictors) and incorporating them into future finer-scale modeling efforts should improve model performance.

To meet conservation goals for the Chesapeake Bay watershed, managers need an estimate of baseline stream conditions for the watershed. Across the Chesapeake Bay watershed, our model (without accounting for uncertainty) predicted 64% of catchments to be in fair/good condition, which is 15% higher than the area-weighted value reported for the full region, family-level index from the Chessie BIBI data set (49%, Smith et al. 2017). However, 11.3% of the area in the Chesapeake Bay watershed was not assessed in the Smith et al. (2017) analyses because of insufficient sample size. If we proportionally include this unassessed area into the Smith et al. (2017) results, fair, good, or excellent conditions account for 56% of the watershed, which is 8% below our modeled estimate. Further, Smith et al. (2017) indicate many of the unsampled HUC12 watersheds were in predominantly agricultural or forested areas, which suggests that 49% and the adjusted 56% estimates may underestimate fair, good, or excellent conditions. Our higher estimate may also have occurred because of the high numbers of fair/good samples predicted in the northern, western, and southeastern portions of the Chesapeake Bay watershed (Fig. 4A–C), areas that were underrepresented in the raw Chessie BIBI data. Additionally, because the Chessie BIBI data set was compiled state-by-state from individually designed (although randomly stratified) sampling programs, it does not consistently represent conditions across the entire Chesapeake Bay watershed. We addressed potential sampling bias by standardizing the spatial density across the watershed. However, areas with higher uncertainty (Fig. 4B, C) could point to areas that require improved or additional data on stream conditions. Future research should evaluate if a more balanced data set would improve the consistency of predictions across the watershed.

Our predictions of biological condition are high relative to those derived from the Hill et al. (2017) national predictive model, which predicted that 36% of catchments with upstream drainages <200 km² were in good condition (based on the NHDplusV2 1:100,000-scale data set; Hill et al. 2017). We predicted a much higher prevalence of fair/good conditions in the northern and southeastern portions of the study area than did the Hill et al. (2017) model (compare Figs. 4 and 5). Several reasons could account for the difference between our 64% fair/good estimate and the Hill et al. (2017) 36% estimate. First, difference in scales between our studies (1:24,000 vs 1:100,000) gave us >20,000 more catchments. We also had nearly 20× more samples (n = 1933 vs n = 100 US EPA NRSA samples with benthic data) for the Chesapeake watershed. More catchments and samples for a smaller area provided us greater ability to predict conditions across the heterogeneous landscape of the Chesapeake Bay watershed. Second, our study used a coarser, family-level taxonomic resolution of macroinvertebrates. However, Smith et al. (2017) found little difference between genus- and family-level BIBI performance, so taxonomic resolution was likely not the reason for model prediction differences either. Third, subsample count size (100 for Chessie BIBI vs 300 for US EPA NRSA) could affect metric calculation, particularly for richness-based metrics, because increasing sample size results in increased richness (Gotelli and Colwell 2001). Unfortunately, given the variety of sampling programs used to compile the Chessie BIBI a 100 rarified subsample was required and, therefore, we cannot evaluate how increasing sample size would affect predictive

![Figure 5. Stream classification predictions for the Chesapeake Bay watershed from Hill et al. (2017) based on the NHDplusV2 1:100,000 scale catchment file and catchments with upstream drainage area <200 km².](image-url)
results. Fourth, our study included intermittent streams, but Hill et al. (2017) did not. Our model predicted intermittent streams to be in fair/good condition nearly as often as other streams (64 vs 65%). Thus, our inclusion of intermittent streams was not a main factor driving the difference. Finally, Hill et al. (2017) removed fair samples from their model and only predicted poor and good conditions, whereas we combined fair with good and excellent condition designations. Future research should examine how including or removing fair conditions affects both models; however, the original 5-class categorization of the Chessie BIBI may complicate such an analysis.

All models are simplifications of complex systems and an important component of any modeling exercise is documenting model error or uncertainty. Mapping such uncertainty, especially in studies with a goal of spatial prediction, aids interpretation of spatial patterns and can highlight spatial patterns or clustering (Elith et al. 2002). Here, we used ±0.05 and ±0.10 bands to identify those catchments where the confidence in prediction between poor and fair/good was low at 2 levels of uncertainty. Incorporating uncertainty helped identify samples or regions that had high numbers of samples in uncertain condition, and that may, therefore, need more data or additional predictors to improve model accuracy. When an end goal is to restore large-scale conditions, such as 10% improvement for the Chesapeake Bay watershed, incorporating uncertainty provides insight into model strength as well as confidence in inferences, such as overall predicted percentages in fair/good condition. Here, the uncertainty analysis suggests that the percentage of catchments in fair/good condition could be as low as 57% or 50% (because 13% or 26% of catchments were in the ±0.05 and ±0.10 uncertainty bands, respectively). These values have very different management implications than the initial conclusion that 64% of catchments are in fair/good condition. Future research that incorporates additional methods that quantify uncertainty arising from different steps in the modeling process (e.g., data quality, collinearity, model type, model tuning, and variable selection; Dormann et al. 2008, Freeman et al. 2016) may further help to understand the limitations of our model predictions.

ACKNOWLEDGEMENTS

Author contributions: KOM, ZMS, CB, and JAY developed the concept for the paper, interpreted the results, and wrote the manuscript. ZMS, CB, and AN developed the Chessie BIBI. KOM performed the random forests analysis.

We thank the many monitoring programs, which provided their data for the Chessie BIBI. We also thank members of the EcoSheds team for publicly providing their database. We thank Daren Carlisle, Charles Hawkins, Katie Sirianni, and two anonymous reviewers whose comments greatly improved this manuscript. Support for the project was provided by the Fisheries Program of the US Geological Survey’s Ecosystems Mission Area. Use of trade, product, or firm names does not imply endorsement by the US Government.

LITERATURE CITED

Abell, R., M. L. Thieme, C. Revenga, M. Bryer, M. Kottelat, N. Bogutskaya, B. Coad, N. Mandrak, S. C. Balderas, W. Busing, M. L. J. Stiassny, P. Skelton, G. R. Allen, P. Unmack, A. Naseka, R. Ng, N. Sindorf, J. Robertson, E. Armijo, J. V. Higgins, T. J. Heibel, E. Wikramanayake, D. Olson, H. L. López, R. E. Reis, J. G. Lundberg, M. H. S. Pérez, and P. Petry. 2008. Freshwater ecoregions of the world: A new map of biogeographic units for freshwater biodiversity conservation. BioScience 58:403–414.

Allan, J. D. 2004. Landscapes and riverscapes: the influence of land use on stream ecosystems. Annual Review of Ecology, Evolution, and Systematics 35:257–284.

Arle, J. 2005. The effects of a small low-head dam on benthic invertebrate communities and particulate organic matter storage in the Ilm stream (Thuringia/Germany). Friedrich-Schiller-Universität Jena, Thuringia, Germany. (Available from: http://d-nb.info/975469258/34)

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 edition. EPA 841-B-99-002. Office of Water, US Environmental Protection Agency, Washington, DC.

Breiman, L. 2001. Random forests. Machine Learning 45:5–32.

Brown, L. R., J. T. May, A. C. Rehn, P. R. Ode, I. R. Waite, and J. G. Kennen. 2012. Predicting biological condition in southern California streams. Landscape and Urban Planning 108:17–27.

Carlisle, D. M., J. Falcone, and M. R. Meador. 2009. Predicting the biological condition of streams: use of geospatial indicators of natural and anthropogenic characteristics of watersheds. Environmental Monitoring and Assessment 151:143–160.

Chesapeake Bay Program. 2015. Stream Health SHWP: management strategy 2015–2025. Volume 1. (Available from: https://www.chesapeakebay.net/managementstrategies/strategy/stream_health)

Chesapeake Bay Program. 2017a. Facts and figures. (Available from: https://www.chesapeakebay.net/discover/facts)

Chesapeake Bay Program. 2017b. Vital habitats. (Available from: https://www.chesapeakebay.net/what/goals/vital_habitats)

Cutler, D. R., T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. 2007. Random forests for classification in ecology. Ecology 88:2783–2792.

Davies, S. P., and S. K. Jackson. 2006. The biological condition gradient: a descriptive model for interpreting change in aquatic ecosystems. Ecological Applications 16:1251–1266.

Domsch, S., S. C. Jähnig, J. P. Siimaika, M. Kuemmerlen, and S. Stoll. 2015. Application of species distribution models in stream ecosystems: the challenges of spatial and temporal scale, environmental predictors and species occurrence data. Fundamental and Applied Limnology/Archiv für Hydrobiologie 186:45–61.

Dormann, C. F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. García Márquez, B. Gruber, B. Lafourcade, P. J. Leitão,
Hawkins, C. P., J. R. Olson, and R. A. Hill. 2010. The reference condition: predicting benchmarks for ecological and water-quality assessments. Journal of the North American Benthological Society 29:312–343.

Hill, R. A., E. W. Fox, S. G. Leibowitz, A. R. Olsen, D. J. Thornburgh, and M. H. Weber. 2017. Predictive mapping of the biotic condition of conterminous U.S. rivers and streams. Ecological Applications 27:2397–2415.

Homer, C. G., J. Dewitz, L. Yang, S. Jin, P. Danielson, G. Z. Xian, J. Coulston, N. Herold, J. Wickham, and K. Megown. 2015. Completion of the 2011 National Land Cover Database for the conterminous United States—representing a decade of land cover change information. Photogrammetric Engineering and Remote Sensing 81:345–354.

Hynes, H. B. N. 1975. Edgardo Baldi memorial lecture: the stream and its valley. Verhandlungen der Internationalen Vereinigung für theoretische und angewandte Limnologie 19:1–15.

King, R. S., M. E. Baker, P. F. Kazyak, and D. E. Weller. 2011. How novel is too novel? Stream community thresholds at exceptionally low levels of catchment urbanization. Ecological Applications 21:1659–1678.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33:159–174.

Liaw, A., and W. Wiener. 2002. Classification and regression by randomForest. R News 2:18–22.

López-Ratón, M., M. X. Rodríguez-Álvarez, C. Cardarso-Suárez, F. Gude-Sampedro. 2014. OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. Journal of Statistical Software 61:1–36.

Maloney, K. O., M. Schmid, and D. E. Weller. 2012. Applying additive modelling and gradient boosting to assess the effects of watershed and reach characteristics on riverine assemblages. Methods in Ecology and Evolution 3:116–128.

Maloney, K. O., D. E. Weller, M. J. Russell, and T. Hothorn. 2009. Classifying the biological condition of small streams: an example using benthic macroinvertebrates. Journal of the North American Benthological Society 28:869–884.

Maloney, K. O., J. A. Young, S. P. Faulkner, A. Hailegiorgis, E. T. Slonecker, and L. E. Milheim. 2018. A detailed risk assessment of shale gas development on headwater streams in the Pennsylvania portion of the Upper Susquehanna River Basin, U.S.A. Science of The Total Environment 610–611:154–166.

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.

Merriam, E. R., J. T. Petty, M. P. Strager, A. E. Maxwell, and P. F. Ziemkiewicz. 2013. Scenario analysis predicts context-dependent stream response to landuse change in a heavily mined central Appalachian watershed. Freshwater Science 32:1246–1259.

Merriam, E. R., J. T. Petty, K. O. Maloney, J. A. Young, S. P. Faulkner, E. T. Slonecker, L. E. Milheim, A. Hailegiorgis, and J. Niles. 2018. Brook trout distributional response to unconventional oil and gas development: landscape context matters. Science of The Total Environment 628-629:338–349.

Paul, M. J., and J. L. Meyer. 2001. Streams in the urban landscape. Annual Review of Ecology and Systematics 32:333–365.

Poff, N. L., and D. D. Hart. 2002. How dams vary and why it matters for the emerging science of dam removal: an ecological classification of dams is needed to characterize how the tremendous
variation in the size, operational mode, age, and number of dams in a river basin influences the potential for restoring regulated rivers via dam removal. BioScience 52:659–668.

Prasad, A. M., L. R. Iverson, and A. Liaw. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9:181–199.

Pond, G. J., J. E. Bailey, B. M. Lowman, and M. J. Whitman. 2013. Calibration and validation of a regionally and seasonally stratified macroinvertebrate index for West Virginia wadeable streams. Environmental Monitoring and Assessment 185:1515–1540.

R Development Core Team. 2017. R: a language and environment for statistical computing. Version 3.4.0. R Foundation for Statistical Computing, Vienna, Austria. (Available from: https://www.r-project.org)

Schueler, T. R., L. Fraley-McNeal, and K. Cappiella. 2009. Is impervious cover still important? Review of recent research. Journal of Hydrologic Engineering 14:309–315.

Scott, M. C., G. S. Helfman, M. E. McTammany, E. F. Benfield, and P. V. Bolstad. 2002. Multiscale influences on physical and chemical stream conditions across Blue Ridge landscapes. Journal of the American Water Resources Association 38:1379–1392.

Smith, Z. M., C. Buchanan, and A. Nagel. 2017. Refinedment of the basin-wide 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, Maryland. (Available from: https://www.potomacriver.org/wp-content/uploads/2015/09/MPRWA_FinalReport_April2014.pdf)

USEPA (US Environmental Protection Agency). 2016. National rivers and streams assessment 2008–2009: a collaborative survey. EPA/841/R-16/007. US Environmental Protection Agency, Office of Water and Office of Research and Development, Washington, DC. (Available from: https://19january2017snapshot.epa.gov/sites/production/files/2016-03/documents/nrsa_0809_march_2_final.pdf)

VanDerWal J., L. Falconi, S. Januchowski, L. Shoo, and C. Storlie. 2014. SDMTools: species distribution modelling tools: tools for processing data associated with species distribution modelling exercises. R package version 1.1-221. (Available from: https://cran.r-project.org/web/packages/SDMTools/index.html)

Villeneuve, B., Y. Souchon, P. Usseglio-Polatera, M. Ferréol, and L. Valette. 2015. Can we predict biological condition of stream ecosystems? A multi-stressors approach linking three biological indices to physico-chemistry, hydromorphology and land use. Ecological Indicators 48:88–98.

Waite, I. R., and P. C. Van Metre. 2017. Multistressor predictive models of invertebrate condition in the Corn Belt, USA. Freshwater Science 36:901–914.

Walsh, C. J., A. H. Roy, J. W. Feminella, P. D. Cottingham, P. M. Groffman, and R. P. Morgan II. 2005. The urban stream syndrome: current knowledge and the search for a cure. Journal of the North American Benthological Society 24:706–723.

Wang, L., J. Lyons, P. Kanehl, and R. Bannerman. 2001. Impacts of urbanization on stream habitat and fish across multiple spatial scales. Environmental Management 28:255–266.

Wang, L. J., Lyons, P. Kanehl, R. Bannerman, and E. Emmons. 2000. Watershed urbanization and changes in fish communities in southeastern Wisconsin streams. Journal of the American Water Resources Association 36:1173–1189.

Woods, A. J., M. Omernik, and D. D. Brown. 1999. Level III and IV ecoregions of Delaware, Maryland, Pennsylvania, Virginia, and West Virginia. National Health and Environmental Effects Research Laboratory, US Environmental Protection Agency, Corvallis, Oregon.