Pilot Analysis of Maryland Phase I MS4 Permit Water Quality Data

Final Report Describing the Results of the Pilot Study and Providing Recommendations for Future Analysis of MS4 Data and the MS4 Monitoring Program

May 29, 2020

Prepared by

Rikke Jepsen

Interstate Commission on the Potomac River Basin

30 West Gude Drive, Suite 450

Rockville MD 20850

www.potomacriver.org

and

Deb Caraco, P.E.

Center for Watershed Protection

3290 North Ridge Road, Suite 290

Ellicott City, MD 21043

www.cwp.org

ICPRB Report 20-2

This final report will be available for download from the Publications tab of the Commission's website, www.potomacriver.org. To receive hard copies of the report, please write:

Interstate Commission on the Potomac River Basin

30 West Gude Drive, Suite 450

Rockville, MD 20850

or call 301-984-1908

Disclaimer

Suggested citation for this report:

Jepsen, R. and Caraco, D. 2020. Pilot Analysis of Maryland Phase I MS4 Permit Water Quality Data. ICPRB Report 20-2. Interstate Commission on the Potomac River Basin, Rockville, MD.

Funding for this pilot study was provided by Maryland Department of the Environment (MDE PCA #40320, Purchase Order #P9400523). The project team consisted of Rikke Jepsen, Claire Buchanan, and Andrea Nagel of the Interstate Commission on the Potomac River Basin and Deborah Caraco, William Stack, and Lisa Fraley-McNeal of the Center for Watershed Protection and was directed by Jeff White and Katherine Slater of the Water and Science Administration, MDE. The team would like to thank the MS4 program staff of Frederick County, Carroll County, and Baltimore City, and especially Donald Dorsey, Byron Madigan, and Prakash Mistry, respectively, for hosting field visits at their monitoring sites. County staff provided very useful reviews of this report.

Table of Contents

Executive Summary

The Interstate Commission on the Potomac River Basin (ICPRB) and the Center for Watershed Protection (CWP) conducted a pilot study of water quality data collected at Moores Run in Baltimore City, Airpark Business Center in Carroll County, and Urbana in Frederick County to characterize stormwater discharges and evaluate watershed restoration activities. The overarching objectives were to determine if there are trends in water quality over time and, if any trends are found, attempt to relate them to watershed restoration efforts or the implementation of Best Management Practices (BMPs). Another goal of the pilot study was to provide recommendations for future analysis of MS4 monitoring data and improving the monitoring requirements in Maryland's Phase I MS4 permits.

Trend analysis of water quality parameters, loading rates, and runoff coefficients was performed using a variety of statistical methods, including permutation methods, least squares and log regression, step trends, Mann-Kendall, Seasonal Kendall, and Seasonal Autoregressive Integrated Moving Average (SARIMA) models. Trends were identified in each watershed. Overall, the Carroll County watershed showed the most noticeable response to watershed restoration efforts, with decreasing metals and nitrogen species noted at the outfall monitoring site, where a large wet pond retrofit had taken place. The impact of restoration efforts was less noticeable in Frederick County, likely because BMP construction and increasing impervious cover occurred simultaneously; however, declining trends in water quality were not observed, so BMP implementation may have mitigated effects of development. Very little development or BMP implementation occurred in the Baltimore City watershed, but trends in loading rates and runoff coefficients suggest that repairs to its sewershed may have impacted both runoff volume and water quality, as increased runoff coefficients and loading rates were observed.

A Principal Components Analysis was conducted to determine if identified trends were associated with changes in land cover and BMP implementation. The results suggested that the impacts of the retrofit at the Carroll County outfall could be directly linked to reductions in both pollutant concentrations and pollutant loads; however, the effects of this restoration practice were generally not observed at the instream station of the same watershed. At the Frederick County location, it was not possible to assess the benefits of a similar retrofit at the outfall, since very little monitoring data were available before the practice was retrofit. Another confounding issue at the Frederick County instream and outfall sites was that development and BMP implementation occurred simultaneously with land development, making it difficult to determine the effect of restoration efforts.

In addition to evaluating whether changes in the watershed were associated with significant changes to water quality or runoff volumes, this study also focused on the quality of the data in the MS4 monitoring database. Some data quality issues included gaps in the data record, censoring of non-detected values, and inconsistent calculations/measurements of event concentrations, event runoff volumes, and precipitation intensities. Associated with these evaluations of data quality, the study included a power analysis to determine the data record needed to detect trends of various strengths, expressed as percent change in pollutant concentration per year. The results varied somewhat by location, but in general, the number of

samples currently required by the MDE were inadequate to detect very strong (10% per year) changes within one permit cycle. Within two five-year permit cycles, on the other hand, the standard of 12 samples per year was adequate to detect strong (10%) changes for most parameters. For moderate (5% per year) trends, a greater sampling rate of between 12 and 24 samples per year would be necessary for most parameters to detect a change within two permit cycles. Weak (2% per year) trends could generally not be detected within two cycles, even with very frequent sampling, with most parameters requiring greater than 48 samples per year.

Based on these analyses, the report made a series of recommendations both for analyzing the remainder of the data in the MDE database, as well as setting standards for future monitoring protocols.

Recommendations for analyzing the remaining data in the MDE database based on this study's findings include:

- Focus on watersheds where restoration impacts can be detected, e.g., watersheds with one or few large BMPs or several smaller BMPs implemented over a relatively short period of time, with data from before and after the watershed restoration practices are implemented, and sites that have limited development over time.
- Select appropriate statistical techniques. Permutation methods are recommended, especially for data that do not have an equal-interval time series. Logistic regression is recommended for parameters with outliers, such as bacteria.
- Incorporate changes to the landscape that may not be apparent from the database or readily available land cover data, such as buried utilities and changes to vegetation.
- Continue to incorporate seasonal variability and rainfall characteristics into the analysis.
- Conduct field visits, as they may provide information about the watersheds not apparent in the data.

The second set of recommendations focus on adapting the MS4 monitoring protocols and include the following:

- Develop a Quality Assurance Project Plan (QAPP) for MS4 monitoring.
- Provide more information regarding flow measurements, including how stormflow and runoff are measured, and a measure of discharge for stormflow and baseflow.
- Develop a specific protocol for reporting censored values.
- Maintain and clarify sampling frequency.
- Sample a range of storm depths.
- Use a flow-weighted EMC calculation.

List of Abbreviations

BACI	Baltimore City (used to refer to the Moores Run watershed)
BMP	Best Management Practice
BOD	Biochemical Oxygen Demand (mg/L)
CACO	Carroll County (used to refer to the Airpark Business Center watershed)
CCCIC	Chesapeake Conservancy Conservation Innovation Center
CWP	Center for Watershed Protection
E. coli	*Escherichia coli* (MPN/100)
EMC	Event Mean Concentration
FRCO	Frederick County (used to refer to the Urbana watershed)
GIS	Geographic Information Systems
ICPRB	Interstate Commission on the Potomac River Basin
LOWESS	Locally Weighted Scatter Plot Smoothing
MDE	Maryland Department of the Environment
MS4	Municipal Separate Storm Sewer System
NO_{23}	Nitrite plus Nitrate (mg/L)
NPDES	National Pollutant Discharge Elimination System
QAPP	Quality Assurance Project Plan
QA/QC	Quality Assurance/Quality Control
SARIMA	Seasonal Autoregressive Integrated Moving Average
SHA	State Highway Administration
TCU	Total Copper (µg/L)
TKN	Total Kjeldahl Nitrogen (mg/L)
TP	Total Phosphorus (mg/L)
TPB	Total Lead (µg/L)
TPH	Total Petroleum Hydrocarbons (mg/L)
TSQVOL	Total Storm Flow Volume (gallons)
TSS	Total Suspended Solids (mg/L)
TZN	Total Zinc (µg/L)

Introduction

The Maryland Department of the Environment (MDE) includes monitoring requirements in the Phase I Municipal Separate Storm Sewer System (MS4) permits. In the first round of permits, starting in the mid-1990s, the goal of the monitoring requirements was the characterization of storm sewer discharges, particularly by dominant land use type. Permittees were required to monitor for a variety of water quality constituents at as many as five outfalls in their system. Monitoring instream stations associated with the outfalls was also required. Around 2000, in the second round of permits, each permittee was required to monitor at only one outfall and one instream location downstream of the outfall, but in addition to water quality monitoring, biological monitoring, habitat assessment, and physical (geomorphic) monitoring were also required downstream of the outfall. The goal was still discharge characterization, with the variation in site characteristics occurring statewide rather than within each permittee's jurisdiction. Physical monitoring was also required in a second small watershed to assess the effectiveness of Maryland's stormwater control regulations. Starting around 2004, in the third round of permits, while the monitoring requirements remained roughly the same, the goal of the monitoring was redirected to determining the effects of stormwater BMPs and watershed restoration on water quality, habitat, and the health of biological communities. Permittees were directed to monitor watersheds where watershed restoration was anticipated, and pre- and post-implementation conditions could be monitored. Current monitoring permits specify roughly the same monitoring requirements for the same reason: determining the effectiveness of watershed restoration.

Monitoring data has been collected under the MS4 program for over twenty years, and MDE is interested in using the data to answer questions related to the water quality characterization of discharged stormwater and the effectiveness of BMPs and watershed restoration, including:

- Do concentrations of water quality constituents in discharged stormwater vary with the dominant land use type in a catchment?
- Does an increase in treated impervious cover (i.e., water quality volume = 1 inch of precipitation) lead to improved water quality?
- Have any improvements in water quality conditions been observed in MS4 monitoring watersheds, where impervious restoration has been implemented? If so, can these improvements in water quality be attributed to watershed restoration efforts?
- Has the overall quality of stormwater discharged by Maryland's MS4s been improving over time?

In this pilot study, the Interstate Commission on the Potomac River Basin (ICPRB) and Center for Watershed Protection (CWP) performed trend analyses using the water quality data collected at the outfall and instream stations of three watersheds monitored by the Phase I MS4 jurisdictions. The study had the following goals:

1. Determine if there have been any observed water quality trends over time at the selected MS4 monitoring locations.

2. If there are any observed trends in water quality, determine if these trends can be attributed to watershed restoration efforts.
3. Conduct a power analysis to determine if current monitoring protocols are sufficient to detect trends in water quality and relate them to watershed restoration or other changes in the watershed.
4. Make recommendations for 1) performing similar trend analyses at other MS4 monitoring locations, and 2) modifying the water quality component of Maryland's MS4 monitoring program.

This technical report for MDE integrates the findings of the statistical analyses with results from previous work done by ICPRB and CWP to QA/QC Maryland's MS4 monitoring data and generate data for potential explanatory variables (i.e., exploratory analysis and preparation of explanatory, confounding, and auxiliary variables). The report then provides recommendations for changes in the required monitoring programs for Maryland MS4 jurisdictions based on the analysis of the water quality data, as well as recommendations for the next steps in analyzing existing water quality data from the MS4 program.

Review of Previous Technical Memorandums

Explanatory, Confounding, and Auxiliary Variables

CWP produced a technical memorandum that investigated possible explanatory, confounding, and auxiliary variables for the water quality trend analyses. The memorandum characterized the land use and BMPs within the monitoring watersheds selected for the pilot study, and it also documented BMP and land use sources and methods used by the jurisdictions in their data collection efforts (Fraley-McNeal, 2019). This report addresses the need to identify variables that can have an impact on explaining or obscuring trends, but which may not be apparent in the water quality data over time. The main findings are given in the following paragraphs.

BMP data were obtained from the National Pollutant Discharge Elimination System (NPDES) geodatabase for Baltimore City, Carroll County, and Frederick County. It was apparent that the monitoring site drainage area boundaries in the NPDES geodatabase did not align with adjacent BMP drainage area boundaries. Possible explanations are that the drainage areas in the NPDES geodatabase were delineated prior to BMP construction or that development altered drainage patterns and delineation was not completed after this occurred. To correct this issue, the monitoring site drainage areas were adjusted to align with the BMP drainage area boundaries (Appendix A: Figures 1a through 1c).

There are five different categories of BMPs in the geodatabase: new, redevelopment, conversion, restoration, and alternative. See Appendix A's Table 1 for the current counts of the BMPs in each watershed, by category. There were some issues with the data, including that some BMPs were reported in incorrect categories, and there were incorrect drainage areas and impervious cover included in the BMP attributes. All BMP data within the monitoring drainage areas were reviewed to verify accuracy, and revisions were made where necessary. See Appendix A: Tables 2 through 6 for more information about the BMPs.

Impervious cover in the Baltimore City monitoring watershed was developed based on data from the Chesapeake Conservancy Conservation Innovation Center (CCCIC). The total amount of impervious surface was calculated as the sum of the impervious road, impervious non-road, tree canopy over impervious, and 30% of the fractional impervious land use class acres (fractional impervious cover is defined as areas that are 30% impervious surfaces and 70% mixed open land). For Carroll and Frederick Counties, land cover data used were digitized impervious cover layers provided by MDE for most years during the period from 2005 to 2018. Land cover summaries are included in Appendix A, Figure 2 and Table 7. The acres of impervious cover within the individual BMP drainage areas were compared to impervious cover in the monitoring watersheds they were located within to determine the portion of the monitoring watersheds treated by BMPs.

In addition to the variables discussed in CWP's technical memorandum, other variables were explored for this report. These include the effects of storm intensity throughout the monitoring time period, seasonality, and precipitation depth. In addition, the analyses investigated unit loads and runoff coefficients (i.e., unit per depth of rainfall) for both trend and land cover analyses.

Exploratory Analysis

ICPRB and CWP produced a technical memorandum documenting exploratory data analyses, the goal of which was to perform preliminary testing of the water quality data to form a bridge between assembling and reviewing the available data and formulating and testing statistical hypotheses (Jepsen and Caraco, 2019). Several statistical and graphical methods were used throughout the exploratory analysis, including tests of normality, histograms, matrix plots, scatterplots, locally-weighted scatterplot smoothing (LOWESS), nonparametric correlation testing, and GIS analyses. More specifically, ICPRB calculated descriptive statistics, explored correlations between water quality data and other variables, and qualitatively examined change in concentration over time. Graphical methods included LOWESS curves to detect raw trends and matrix plots to help identify correlations. CWP undertook preliminary testing of the water quality data, including tests of normality, as well as performed time series decomposition and explored metrics associated with runoff. CWP also investigated whether parametric methods could be used. The following paragraphs provide an overview of the exploratory analysis results.

To begin, some data quality issues were identified with the event mean concentration (EMC) calculations and with data entry (e.g., values being recorded as baseflow and stormflow on the same date). Also, the majority of the data had non-normal distributions, meaning that nonparametric tests would be emphasized for later statistical analysis.

LOWESS curves were generated in order to detect trends in the untransformed datasets with nonparametric regression. Based on visual inspection of these plots, as well as time series decompositions, few water quality parameters appeared to be strongly increasing or decreasing during the monitoring time period. However, there did appear to be some overall patterns in concentration over time, seasonal effects on concentrations of most pollutants, and correlations between instream and outfall concentrations.

Correlations between the water quality parameters and between storm event parameters (i.e., duration, intensity, total storm flow volume, depth) were also evaluated. Many parameters were found to be correlated, which helped to bolster confidence in the quality of the data because expected relationships were observed in many circumstances. For example, metals tend to adsorb onto the organic fraction of sediment particles (Leisenring et al., 2011), so one would expect metals and sediments to be correlated during stormflow conditions, when the majority of sediments are transported through a stream system. This was reflected in the water quality correlations, as total suspended solids were correlated with copper, lead and zinc during stormflow at all sites except for Frederick County's outfall.

Finally, the amount of runoff resulting from the monitored storm events was evaluated. The runoff coefficient appeared to have a stronger trend over time than any pollutant concentrations.

Methods

Site Descriptions

Three watersheds with outfall and instream monitoring performed in compliance with Phase I MS4 permits were selected for this pilot study. Watersheds with the following characteristics were chosen:

- Have a long (ten years or more) monitoring record at fixed stations with regular frequency and minimum data gaps.
- Have relatively complete data sets with consistent monitoring of water quality parameters.
- Be relatively free of data quality issues (e.g., data entry errors, missing data).
- Have a documented change in watershed restoration or BMP implementation over the monitoring period.
- Show preliminary evidence of trends, as reported in a previous ICPRB report (Nagel and Mandel 2018), consistent with the changes in watershed restoration.
    - This 2018 report (Analysis of Monitoring Data Collected under Maryland's Municipal Separate Storm Sewer System (MS4) Permits: Database Design and Preliminary Analysis of Water Chemistry) describes how the monitoring data were integrated into a database and associated challenges with this process. Preliminary trend analysis using simple linear regression was also done.

Based on these requirements, the following watersheds were selected: Moores Run in Baltimore City; Airpark Business Center in Carroll County; and Urbana in Frederick County. In each case, the outfall station is located upstream of the instream station, so the instream station reflects inputs from a larger area.

Moores Run, Baltimore City (BACI)

Moores Run has water quality monitoring data available from 1999 to 2016, except for a gap in 2013. This watershed is highly urbanized with a significant amount of impervious cover. The outfall station is located at Hamilton Avenue and drains 113.2 acres. The instream station is located at Radecke Avenue and drains 2,247.6 acres (Figure 1). Even though the catchment area

is highly urbanized, the reach between the outfall and instream station has a significant wooded buffer. There has been minimal implementation of structural BMPs in this watershed.

A site visit was conducted at the Moores Run watershed on August 20, 2019, during which MDE, ICPRB, and CWP staff were shown the monitoring locations and sampling equipment by Baltimore City staff Prakash Mistry and Nick Mitrus. See Figures 2a through 2f for pictures taken from the site visit.

Airpark Business Center, Carroll County (CACO)

Chemical monitoring data are available for the Airpark Business Center watershed from 2000 to 2016. The outfall, WPU01, drains 207.3 acres and is located in an industrial park. The instream site, WPU02, drains 555.2 acres, and the watershed area between WPU01 and WPU02 is largely agricultural (Figure 3). BMPs have been implemented in the drainage areas of both the instream and outfall stations, notably a large detention basin retrofit at the outfall.

On July 22, 2019, MDE, ICPRB, and CWP staff visited the Airpark Business Center monitoring stations. Carroll County staff Byron Madigan gave a tour of the monitoring locations and described the sampling procedure. See Figures 4a through 4d for pictures taken from the site visit.

Urbana, Frederick County (FRCO)

Water chemistry data from 1999 to 2016 were used in this pilot study from one of Urbana's stormwater ponds, which was built in 2004. The pond's outfall station, Pond-R, is located in a high-density residential area and drains 30.1 acres. The instream site, located on Peter Pan Run within the Bush Creek watershed, drains 1,584.6 acres (Figure 5). The watershed area between the outfall and instream stations is primarily agricultural. Few BMPs are located in the outfall watershed, but their drainage areas are extensive. The instream drainage area contains significantly more BMPs.

On August 1, 2019, MDE, ICPRB, and CWP staff were given a tour of the FRCO monitoring sites and sampling equipment by Don Dorsey of Frederick County and Nathan Drescher of KCI Technologies, Inc. KCI presently monitors the sites, but another company performed monitoring during the pilot study time period. See Figures 6a through 6d for pictures taken from the site visit.

Data Sources and Preparation

MS4 Database - Construction and Variables

Working with MDE staff, ICPRB designed a relational database to store the monitoring data submitted by Phase I MS4 jurisdictions and State Highway Administration (SHA) under the requirements of their permits (Nagel 2019). The database is designed to hold water chemistry, habitat, biological, and physical data in separate tables, linked to other tables identifying sampling activities and monitoring locations. ICPRB staff populated the database with the available water chemistry data submitted by the permittees. Almost 97,500 records of chemical and flow parameters were included in the database, taken from nearly 5,000 sampling events at 69 monitoring locations. The most recent year for which data are used in this pilot study is 2016,

but note that data have been collected since then and are still being collected. Future work could explore analyses extending past 2016.

The database contains a number of water quality parameters, and nine of these parameters were used in the pilot study to investigate water quality response to BMP implementation: Biochemical Oxygen Demand (BOD, mg/L); Total Suspended Solids (TSS, mg/L); *Escherichia coli* (*E. coli*, MPN/100); Nitrite plus Nitrate (NO_{23}, mg/L); Total Kjeldahl Nitrogen (TKN, mg/L); Total Phosphorus (TP, mg/L); Total Copper (TCU, µg/L); Total Zinc (TZN, µg/L), and Total Lead (TPB, µg/L).

Data Censoring

Censored data points are known only to be "less than" or "greater than" a limit of detection, meaning that their true value is unquantifiable (Helsel and Hirsch, 2002). In this study, censored values were measurements below the limit of detection of a given parameter. Censoring was most prevalent with BOD and metals (TZN, TPB, TCU), so before statistical analyses could be performed, various issues concerning the detection limits needed to be addressed. First, in the data provided by the MS4 jurisdictions, values below the detection limit had been assigned the value of the detection limit, rather than being left as the instrument-measured value. Second, detection limits of zero and NA were present in the database. Third, there were often several detection limits listed for a specific parameter, some of which appeared to be data entry errors, and others reflected either a change in the laboratory methods over time, different analysts measuring the samples, or variations in sample quality (Appendix B).

Most statistical methods used require a single detection limit in order to be performed. To correct the issue of multiple detection limits, a dataset with updated values (called “result values” in the dataset) was created and used in the analyses. Uncensored values (i.e., values greater than the detection limit) were unchanged. Also unchanged were the value of data points whose detection limits were either zero or NA, as these were deemed to be uncensored. Values that were less than the indicated detection limit were interpreted to be uncensored values matched to an incorrect detection limit, so these were also classified as uncensored. For censored values (i.e., values equal to the detection limit), the detection limits had to be used as the “result value,” but were standardized to the same detection limit among each parameter and flow type combination, then set to half the detection limit.

Substituting in a value (e.g., zero, the detection limit) for nondetects is not the preferred method for handling censoring (Helsel and Hirsch 2002); however, as stated above, the data received by MDE from the jurisdictions already had these substitutions made and also included multiple detection limits per parameter. Furthermore, assumptions had to be made about which data points were truly censored. Selecting one of the already substituted detection limits for parameters for each jurisdiction/flow type combination and then halving it was thought to be a more conservative representation of the detection limit that would also minimize the effects of censoring.

BMP Data and GIS Layers

The MDE provided ICPRB and CWP with the NPDES BMP geodatabase that included information about locations and types of BMPs in MS4 jurisdictions, as well as built year and drainage areas. As discussed above, it was found that the monitoring drainage area boundaries of individual BMPs in the geodatabase did not perfectly overlay with one another or the watershed drainage areas. In most cases, the discrepancies were due to the delineation of the catchment or watershed areas, and these boundaries were corrected to align with BMP drainage area delineations. When additional information was needed about a specific BMP or drainage area, CWP staff contacted the appropriate jurisdiction.

In Baltimore City, impervious cover could be characterized using Chesapeake Conservancy Conservation Innovation Center layers because there were no major changes in land cover over the monitoring period. In both Frederick County and Carroll County, at least a portion of the watershed experienced land development during the monitoring period. Consequently, MDE staff digitized aerial imagery to create a time series that reflected changes in impervious cover and tree canopy, and these data were combined with BMP treatment data to characterize land cover change and BMP implementation as a time series.

Statistical Approaches

Several procedures and statistical tests were performed to answer MDE's management questions. Specifically, outfall and instream concentrations were compared to determine if they behaved differently, which would indicate additional factors influencing stream conditions below the stormwater outfall. Trend analysis of the outfall and instream concentrations was performed after data had been appropriately prepared, as described above. Methods included traditional trend analysis (i.e., Mann-Kendall and Seasonal Kendall tests), step trends, permutation methods, and other nonparametric and log-transformed methods. Runoff coefficients and loading rates were calculated and their trends analyzed using permutation analysis, log-link least squares regression, and logistic regression. In addition to analyzing trends, the analyses include the use of Principal Components Analysis to develop a relationship between a restoration index, reflecting BMP implementation and land cover, and observed values including pollutant concentrations, runoff coefficients, and pollutant loads (pounds per inch of rainfall). In addition, a power analysis was conducted to determine the number of samples that would need to be collected annually to detect change in one or two permit cycles. Finally, recommendations are given for changes in the MS4 monitoring program and the next steps in analysis of the existing water quality data. Results of this statistical analysis and recommendations for future analysis were presented to MDE in person at MDE's offices on November 15, 2019 and subsequently on November 27, 2019 in writing.

Pairwise Comparisons

The goal of the first analysis was to determine whether concentrations of the water quality parameters were significantly different between the instream and outfall sites in each watershed. To accomplish this, data needed to be paired such that individual dates had a measurement from both the instream and outfall ~0.45sampling locations. Consequently, dates

that had data from only the instream or outfall site were excluded from the analysis. Fortunately, pairing the data resulted in only a small number of observations being excluded.

Three methods were used to compare the paired data points. The first was resampling methods (i.e., permutation methods) with the "paired.perm.test" function in the "broman" R package. The second was the Wilcoxon signed-rank test, using the paired option of the base R "wilcox.test" function, and the third was a paired t-test conducted with the base R "t.test" function. For a paired t-test, the data do not have to be normally distributed, as long as the differences between the pairs are normally or nearly normally distributed, as was the case for this dataset. For all methods, a p-value of less than or equal to 0.05 was used to identify significant differences.

Trends

Trends in Concentrations

To address the first objective of the pilot study, which was to determine if water quality trends occurred over time, trend analysis was performed. A variety of methods were used: permutation, traditional methods (i.e., Mann-Kendall, Seasonal Kendall), least-squares regression, logistic models, Seasonal Autoregressive Moving Average (SARIMA) time series, and step trends. A significance level of 5% was used to identify significant p-values for each method.

Permutation Methods

Permutation methods offer an alternative to classical statistical methods, which rely on an underlying distribution to estimate the confidence or range of a particular value or slope (Elliffe and Elliffe 2019). For example, if linear regression is used to model a relationship between variables, the p-value of the slope is estimated by assuming an underlying normal distribution. In permutation methods, the distribution is estimated by randomly resampling the original data to simulate this distribution.

For this study, the R package "perm" (Fay and Shaw 2010) was used to detect trends using the "permTREND" function. This package uses either exact methods, through Monte Carlo simulation, or asymptotic methods, based on the quality of the data. In order to account for the variability introduced by precipitation depths, stormflow concentrations were first fit to a linear least squares model of precipitation versus concentration. The residuals of this model were then correlated with time to represent a precipitation-adjusted change over time.

Mann-Kendall and Seasonal Kendall Methods

The traditional trend methods used were the nonparametric Mann-Kendall and Seasonal Kendall tests. With guidance from Helsel and Hirsch (2002) and Buchanan and Mandel (2015), data were prepared for trend analysis by first classifying their temporal coverage and degree of censoring, as these factors determine which tests to use.

To assess temporal coverage of the data, the number of samples in every month was counted and arranged in tables for each site's station and flow type combination. Sampling frequency did not allow for weekly, monthly, or bimonthly time series, but trimonthly seasonal time series were possible. Four seasons were defined for the purposes of this analysis: winter

(December, January, February), spring (March, April, May), summer (June, July, August), and autumn (September, October, November). The percentage of months with measured values was calculated for each season, as well as for the entire monitoring period (called "annual" hereafter) (Tables 1a though 1c). If the annual coverage was greater than or equal to 75%, it was considered acceptable for trend analysis, but if its coverage was less than 75%, seasonal coverage was evaluated. Only seasons with greater than or equal to 75% coverage were analyzed. When annual trend analysis was possible, trend analysis was not conducted on each season independently, as data from each season would be included in the annual trend.

Once the temporal coverage was analyzed, the detection limits for the censored values needed to be standardized. To accomplish this, the data were classified based on the percentage of nondetects present in each parameter, broken down by site and flow type. Data were assigned as minimally censored if there was less than 5% censoring, moderately censored if there was between 5% and 50% censoring, and heavily censored if censoring was greater than 50%. A single detection limit was selected for moderately censored parameters but not the other censoring categories because 1) if less than 5% of the data are censored, trend tests will most likely not be affected, and the original detection limits were kept, and 2) heavily censored data are not suitable for traditional trend analysis, so were not evaluated. For moderately-censored parameters, the chosen detection limit was usually the most common detection limit among the censored values, as well as the lowest value in that dataset and was set to half of the detection limit in the updated dataset.

In order to produce a set of uniform time series for trend analysis, the data were culled to establish an even distribution over time for each parameter. Culled datasets were assembled by selecting the value closest to the midpoint of each season to represent that season-year. For seasonal time series, only the midpoints of the given season were used. The percent of data below the detection limit was then recalculated for these datasets (Tables 2a through 2c). In the case of moderately-censored datasets, an extra step was taken to reduce the bias of censored values. These data were seasonally-adjusted (also referred to as median-adjusted) by subtracting the median value for each season from that season's observations.

Flow adjustment is used to remove flow as a source of variability in concentration and to facilitate trend detection. Since flow measurements are not available in the database, a modified flow adjustment procedure was used, in which intensity of rainfall (inches/hour) was substituted for flow (cfs). This adjustment was only performed on annual, culled datasets with minimal censoring. Additionally, as intensity data are only associated with storm events, this procedure could not be performed on baseflow data. Data that have been treated in this way will be referred to as MFC (modified flow-corrected).

Based on the degree of censoring and data coverage, either Mann-Kendall or Seasonal Kendall tests were used. As shown in Table 3, time series with minimal censoring were analyzed using the Seasonal Kendall test for annual time series or with the Mann-Kendall test for seasonal time series. Both annual and seasonal time series with moderate censoring were analyzed with Mann-Kendall tests. Datasets containing more than 50% nondetects are too heavily censored to produce meaningful trend results. Instead, plots of the detection limits and measured values were

generated (Appendix C: Figures 1-22). See Tables 4a through 4c for the specific tests conducted for each parameter.

To perform the modified flow adjustment procedure, residuals were calculated from LOWESS curves of log-transformed concentration against log-transformed intensity. Seasonal Kendall tests were then applied to these residuals. Please see Tables 4a through 4c for which parameters were modified flow-corrected.

Once the data were prepared, trend analysis was performed using R 3.5.1 (R Core Team, 2013). Before executing a trend test, the degree of autocorrelation, or internal, lagged correlation, was evaluated by generating autocorrelation function (ACF) and partial autocorrelation functions (PACF) plots. Additionally, "modifiedmk" (Patakamuri and O'Brien, 2019), an R package that performs Mann-Kendall tests corrected for autocorrelation, was utilized. For Seasonal Kendall tests, the "Kendall" package (McLeod, 2011) was used. As well as providing p-values, Kendall's Tau is also reported. Its sign describes a trend as positive (increasing) or negative (decreasing).

Log-Link Least Squares Regression

This method uses simple least squares regression to determine if a trend is present over time by performing a regression with concentration versus time (in years), using a "log-link" function. The "log-link" function preserves the mean of the original sample, while using a log-transformation to reduce the influence of outliers. The log-link function was used because it was determined in the exploratory analysis phase that the data were not normal. The model was generated using the "glm" function, which is available in base R. The value of the coefficients represents an exponential function of decay, or increase over time, in the following form:

$$C = C_0 e^{at+bP}$$

Where:

C = Concentration
C_0= Initial Concentration
a = Coefficient of exponential change
t = Time (days)
b= Coefficient associated with Precipitation (only for stormflow)
P = Precipitation Depth (only for stormflow)

Precipitation is included in the equation to account for the variability introduced by this parameter during storm events, but the results presented in this document do not summarize the precipitation depth in detail.

Logistic Regression

Logistic regression estimates to what degree a variable influences the likelihood of a binomial variable being in a specified category. The regression estimates the logit (or the log of odds) based on predictive variables. In this case, concentration values were classified as "high" or "low," with high values being above the median over the entire monitoring period. The results of the logistic regression represent a change in this value in response to independent variables.

For this study, the logit values were regressed against time for baseflow values and against time and precipitation for stormflow values.

SARIMA Models

SARIMA models are used in time series data to account for correlations between data points as well as seasonal effects. One component of these models is a "drift" constant, which identifies the average movement in each time period. These methods are widely used for time series data because they can adjust for errors that can be introduced when each value is related in time to previous values. Accounting for these relationships can help to identify significant influences or trends that might not otherwise be apparent.

Data were aggregated to a monthly time step. While there were relatively few months with multiple observations in the same year, the data were not uniformly spaced, and a potentially greater issue was that resulting datasets had multiple periods without any observations. This gap in observations makes it difficult to determine the relationship between each successive value (i.e., the autocorrelation), which is a critical component SARIMA models. The "auto.arima" code, which is a part of the "forecast v8.9" package, was used to develop model issues (Hyndman and Kanjdakar, 2008). This package automates the process of selecting a model, but it is important to note that the model selection needs to be confirmed. Without changing the settings, the models produced by default had issues with autocorrelation. Next, the "sarima" function, which is a part of the "asta v 1.9" package, was used to check the models and to evaluate if the results were statistically significant. Models were evaluated with some changes to the standard settings, which included searching all models rather than using a stepwise selection and using maximum likelihood estimation rather than an approximation. Both of these changes improve the model but result in very long computation times. Using these packages together, the projected drift, or change per month, was estimated for models where a seasonal drift was determined to be significant, but this was not the case for many of the models.

Step Trends

Step trend analysis was used to determine if there were significant differences in the water quality parameter concentrations between three different time periods: a drought period from 1999-2002, in which the Potomac region experienced record dry conditions, as evidenced by extremely low flows in the Potomac River mainstem; 2003-2009, characterized mostly by moderate conditions; and 2010-2016, a moderate to wet period. The classification systems developed by Olson (2005) were used in characterizing the three time periods, referred to as Period I, Period II, and Period III, respectively. Analysis was performed on the culled datasets of all data except those that were heavily censored.

To determine if there was a significant difference between any of the time periods, the nonparametric Kruskal-Wallis test was performed. In cases where the Kruskal-Wallis test had a significant p-value, individual Mann-Whitney tests were used to determine which group or groups were different from one another. Base R functions were used.

Trends in Runoff Coefficients

The runoff coefficient represents the amount of surface runoff generated by a given rainfall depth. For this study, the coefficient is represented as a volume of runoff (cf) per inch of rainfall depth. This parameter is important for three primary reasons. First, the runoff coefficient is highly dependent on land cover in the watershed, particularly the amount of impervious cover (Schueler 1999; Pitt 1999). Therefore, the coefficient is a robust measurement of the impacts of land development. Second, modern stormwater BMPs are designed to both "treat" runoff and to reduce the volume of runoff (Battiata et al. 2010). Therefore, successful BMP implementation in a watershed should be correlated with reduced runoff volumes. Third, the volume of runoff helps to explain water quality conditions that may not be represented by pollutant concentrations alone. For example, channel erosion is influenced by increased stream flows associated with urbanization. This energy cannot be understood or captured by concentrations alone. In addition, the total pollutant load is influenced by both the concentration and flow of runoff.

The MDE database reports the volume of runoff (Total Storm Flow Volume; TSQVOL) in gallons, and rainfall depth (DEPTH) in inches. For this report, the runoff coefficient was calculated as follows:

$$Rv = (TSQVOL) / (DEPTH \times 7.48)$$

Where:

Rv = Runoff coefficient (cf / inch)
TSQVOL = Runoff volume (gal)
DEPTH = Rainfall Depth (inches)
7.48 = Conversion factor (gal / cf)

Two statistical methods were used to evaluate trends in the runoff coefficient: a permutation method using the "permTREND" package and a log-transformed least squares regression, as previously described.

Trends in Unit Loading Rates

Loads expressed as expected loading rate (e.g., pounds) per inch of runoff are useful for understanding how the total pollutant load delivered to a waterbody changes over time. Trends in unit loading are a combination of the changes in concentrations and changes in runoff volume. For the pilot study, loads were calculated for nutrients (NO_{23}, TP, and TKN) and TSS. The pollutant load, in pounds, was calculated for each observation as a product of runoff volume (gallons) and pollutant concentration (mg/L) and a factor (8.33×10^{-6}).

Land Cover Effects

This study hypothesizes that changes in runoff volumes, pollutant loads, and concentrations are the direct result of changes in land cover and BMP implementation over time in a watershed. These land cover changes were quantified for each drainage area (Fraley-McNeal 2019) and are summarized in this section for watersheds where land cover changes were observed. These include both the outfall and instream stations in CACO and the instream station in FRCO. No appreciable changes in land cover or BMP implementation occurred in BACI; at

the FRCO outfall, a single BMP was implemented in 2004, but the earliest available data at the site is 2003. As a result, these stations are not included in all analyses. The variables investigated included stormflow concentrations, baseflow concentrations, the runoff coefficient, and unit area pollutant loading rates (per inch of runoff). The analyses focused on four parameters: TSS, NO_{23}, TKN and TP.

Predictor Variables

For each analysis, the following variables were included: rainfall depth, season (represented by annual quarter), impervious cover acreage, and impervious cover capture by stormwater BMPs (acres(. The more detailed land cover descriptions characterized in Fraley-McNeal (2019) were condensed to develop straightforward predictors for each watershed of concern. These predictors were slightly different in FRCO and CACO due to differences in the types and extent of BMP implementation at each station. In addition, digitized land cover was only available from 2005 onward, so for years prior, impervious cover and BMPs were estimated using the sign methods outlined in Fraley-McNeal (2019).

Land Cover and BMPs

In CACO, the entire drainage area in the outfall catchment was initially captured by an older wet pond, but the area was classified as "untreated" since the pond was not designed to modern standards. In 2008, the wet pond was converted to a modern wet pond, and, at this point, impervious cover in the drainage was converted to the "treated" category. In the area within the drainage area of the instream station, but outside of the outfall drainage, land draining to a shallow wetland were also put into the "treated" category. Finally, small BMPs within the drainage area of either the pond or shallow wetland are considered "nested." The "treated" category includes the drainage area of all BMPs, including the nested BMPs.

In FRCO, the outfall drainage area had no change in land cover or BMP coverage for the period in which land cover data had been digitized; therefore, the analysis focused on the instream station. In contrast to the CACO location, a large portion of the watershed was untreated, and the BMPs that were in the constructed watershed encompassed a much broader range of types and designs. Finally, no nested BMPs were present in the watershed. Consequently, the factors used in the analysis were somewhat different, including untreated impervious cover and impervious cover treated by dry ponds, wet or micropool ponds, wetlands, filters, or other practices.

Other Predictors

In addition to the land cover parameters, the effects of rainfall and seasonal effects, as estimated by the annual quarter, were included as predictor variables. These parameters are included in the model to account for their influence but are not reported as variables of concern.

Methodology

Since several components represent land development and treatment in this method, we developed a "Treatment Index'" for each drainage area using Principal Components Analysis. This method reduces the number of variables in the model and eliminates the potential for

correlations between parameters. The specific elements of the regression were slightly different for Carroll County and Frederick County.

The components were developed using the "princomp" function, with the following results:

$$I_{CACO-} = -0.395 \times Untreat + 0.650 \times Treat + 0.650 \times Nest$$

$$I_{CACO-I} = -0.534 \times Untreat + 0.815 \times Treat + 0.223 \times Nest$$

$$I_{FRCO-I} = -0.594 \times Untreat - 0.533 \times Dry - 0.603 \times Wet$$

Where:

I = Treatment Index

Untreat = Area of Impervious Cover Not Treated by BMPs (acres)

Treat = Area of Impervious Cover Treated by One or More BMPs (acres)

Nest = Area treated by a smaller BMP "Nested" within the drainage area of a larger BMP (acres)

Dry = Area treated by Dry Detention Ponds (acres)

Wet = Area treated by Wet Detention Ponds (acres)

In general, a higher treatment index number reflects a greater degree of restoration, but the index is not as clear in Frederick County, where land development is confounded by BMP implementation. The component vectors accounted for almost all of the variability in the land cover data over time, with the following results (see Tables 5a-5c for eigenvectors):

- CACO outfall: 99.3% of the variability (eigenvalue of 3,845)
- CACO instream: 98.7% of the variability (eigenvalue of 5,242)
- FRCO instream: 94.8% of the variability (eigenvalue of 2,346)

The effects of change in each drainage area were then evaluated by developing a regression equation that includes the index for each drainage area, along with the quarter (as an indication of season) and precipitation (for pollutant concentrations) to predict pollutant concentrations, runoff coefficients, and pollutant loads (per inch of rainfall). The permutation regression was conducted using the "lmp" function in the "lmperm" R package.

Power Analysis

A power analysis was used to determine how many samples per year would be needed to detect a change within one or two five-year permit cycles. The "power.trend" and "generate.trend" functions from the "emom" package in R were used for this analysis. The "power.trend" function determines the power, which is the percent chance of detecting a trend at a given significance level, by comparing mean values and associated distributions for each year of data and by completing simulations given the standard deviation of the data, as well as the underlying average annual concentrations, using the following assumptions:

1. The data are lognormally distributed, and trends occur exponentially (i.e., at a given % reduction per year).
2. The annual variability is characterized by the standard deviation of the residuals from the log-link models completed as a part of the trend analysis.
3. Trends are detected at significance of 5%.
4. The desired power is 0.80.

The trends tested include three scenarios, each representing a reduction of a given percentage each year, including 2%, 5% and 10% concentration reductions per year. The average value per year was generated using the "generate.trend" function. In this case, annual decline in the log-transformed values was represented. For each generated trend, the power was tested using the "power.trend" function.

The power analysis assumes that a complete data record is available, with no significant gaps. By contrast, the data analyzed as a part of this study had some time gaps, both within seasons and over longer periods of time, for some parameters.

Results

Pairwise Comparisons

The results from the three methods (permutation, Wilcoxon signed-rank, and paired t-test) that were used to compare measures of central tendency between outfall and instream stations for both stormflow and baseflow conditions were almost identical, with only few differences in which station (i.e., outfall or instream) had the higher values and which differences were significant. These small differences are likely explained by the Wilcoxon signed-rank test comparing *medians*, whereas the permutation and paired t-test methods compare *means*.

Overall, the tests suggest that there are significant differences between outfall and instream stations for many parameters in both storm- and baseflow conditions. Although there were some differences between stations, some patterns emerge. For example, every station has significantly higher TSS concentrations at the instream station during stormflow compared to baseflow, and instream baseflow NO_{23} concentrations were greater in instream baseflow compared to stormflow (Tables 6a-6c and 7a-7c).

Trends

Trends in Concentrations

Each analysis method and its results are described in the subsequent sections, and the combined results of all methods are presented in Tables 8a-8c and Appendix D.

Permutation Methods

Overall, less change was observed at the BACI location than the other two watersheds in this study. TP, TKN, TCU, and TPB displayed no trends at the outfall or instream stations, but *E. coli* showed a strong and statistically significant negative trend over time in both the instream and outfall stations. The instream station also showed a weak negative correlation for TZN and a

moderate negative correlation for NO_{23}. The outfall station also experienced moderate positive trends for BOD and TSS (Table 9a).

Trends at the CACO location were quite different between the instream and outfall stations. The only significant trend at the instream station was a moderate negative trend in TCU. At the outfall, on the other hand, strong negative trends were seen for all metals, as well as moderate significant negative trends for TP and TSS. A moderate positive trend was also seen for BOD at the outfall (Table 9a).

FRCO also experienced different results between the instream and outfall stations, with moderate negative trends for NO_{23} and TCU, a weak negative trend for TPB, and a moderate positive trend for TKN at the instream station. At the outfall station, moderate negative trends were observed for TSS, NO_{23}, TP, and TCU, as well as a weak negative trend for TKN (Table 9a).

Only a few statistically significant trends were found in baseflow concentrations, including TKN at the BACI outfall station, TP at the CACO outfall station, NO_{23} at all CACO and FRCO stations, and TSS at the FRCO instream station. Baseflow results were not calculated for BOD or the metals (TCU, TPB, TZN). For these parameters, numerous non-detectable events were monitored, and the detection methods changed over time. As a result, any trends detected were impossible to separate from the methods used (Table 9b).

Mann-Kendall and Seasonal Kendall Methods

In BACI's watershed, instream NO_{23}, modified flow-corrected (MFC) NO_{23}, TZN, and TPB significantly decreased in stormflow. Instream baseflow TCU and TZN significantly decreased. At the outfall during storm events, spring TSS increased, while summer TP, TCU, and TZN decreased. Outfall baseflow experienced a significant negative decrease in TCU and TZN (Table 10a).

Carroll County's instream site did not experience a significant reduction of any water quality parameter in autumn stormflow. Instream baseflow had a significant decrease in spring NO_{23} and summer TP. At the outfall, there was a significant increase in autumn stormflow BOD, with a significant reduction in autumn stormflow NO_{23} and TCU. Outfall baseflow experienced an increase in summer TSS, but a decrease in spring TP and TCU. Overall, trends were much stronger at the outfall site (Table 10b).

Frederick County's instream site had significant reductions in stormflow NO_{23}, MFC NO_{23}, and TCU, but a significant increase in TKN. For baseflow at the instream site, only TSS was significantly reduced and no parameter significantly increased. At the outfall, stormflow TP, MFC TP, TCU, and MFC TCU significantly decreased (Table 10c).

Log-Link Least Squares Regression

For the purposes of this analysis, the results were expressed as the log of the coefficient of change, so the data presented in Tables 11a through 11b are represented as a fraction over a year, meaning that values greater than 1.0 represent an increase (e.g., 1.2 is 20%/year).

Overall, the BACI watershed experienced less change than others in this study, but a statistically significant negative trend in stormflow at the instream station for *E. coli*, NO_{23}, TZN and TPB, and at the outfall for *E. coli* was observed. There was also a significant increase in TKN and TSS at the outfall (Table 11a).

The trends at the CACO location were quite different between the instream and outfall stations. The only significant result at the instream site was a negative trend for TCU during stormflow. At the outfall, on the other hand, negative trends were observed during stormflow for all metals, TP, and TSS, with a weak positive trend for BOD (Table 11a).

At FRCO, negative trends were observed for stormflow NO_{23} and TCU at both stations, as well as a negative trend for TPB at the instream station. Negative trends in TKN, TP, TSS and BOD were observed during stormflow at the outfall station. Finally, a positive trend in TKN was observed at the instream station (Table 11a).

Only a few statistically significant trends were found in baseflow concentrations, including NO_{23} at the CACO instream station, NO_{23} and TP at the CACO outfall station, and NO_{23} and TSS at the FRCO instream station. Baseflow results were not calculated for BOD or the metals. For these parameters, numerous non-detectable events were monitored, and the detection methods changed over time. As a result, any trends detected were impossible to separate from the methods used (Table 11b).

Logistic Regression

The logistic regression results transform the logit value to reflect an odds ratio. For example, a value of 1.2 suggests that over a one-year period the odds of a high concentration are 20% higher.

Overall, the BACI location experienced less change than others in this study, showing a statistically significant negative trend over time at the instream station for stormflow *E. coli*, NO_{23}, and TPB, and at the outfall for *E. coli*. There was also an observed increase in BOD at both stations and for TSS at the outfall (Table 12a).

The trends at the CACO location were quite different between the instream and outfall stations. The only significant trends at the instream station were a negative trend for stormflow TCU. At the outfall, on the other hand, negative trends were observed for all metals and TP (Table 12a).

At the FRCO watershed, negative trends were observed for stormflow NO_{23} and TCU at both stations, as well as a negative trend for TPB at the outfall station. A negative trend in TP and TPB was also observed at the outfall station. Finally, a positive trend for TKN was observed at the instream station (Table 12a).

Only a few statistically significant trends were observed in baseflow concentrations, including TSS at the BACI outfall, NO_{23} and TP at both CACO stations, TSS and TP at the FRCO instream station, and NO_{23} at the FRCO outfall station. Baseflow results were not calculated for BOD or the metals (TCU, TPB, TZN). For these parameters, numerous non-

detectable events were monitored, and the detection methods changed over time. As a result, any trends detected were impossible to separate from the methods used (Table 12b).

SARIMA

The results of this method did not find any significant trends.

Step Trends

In the BACI watershed, the only parameter at the instream site to show significant differences between time periods during stormflow was NO_{23}, specifically between Periods I and II, I and III. At the instream site during baseflow, there were differences among the periods of TKN and TCU. All time periods of TKN differed from one another, and Periods I and III, II and III differed for TCU. At BACI's outfall during baseflow, Periods I and II, II and III were significantly different from one another for TKN. Regarding TCU measurements, Period I and III, II and III were significantly different, and for TZN, I and III, II and III differed (Table 13a).

In CACO, comparisons using Period I were not possible due to lack of data. Differences between Periods II and III only occurred at the outfall. Specifically, Periods II and III differed significantly for stormflow TCU. In baseflow, summer TSS and spring TCU had significantly different values recorded between Periods II and III (Table 13b).

Comparisons between time periods at FRCO only revealed significant differences at the instream site. Stormflow TCU was significantly different between Period I and III. Time periods during baseflow for TSS were all different from one another, and for TP, Periods II and III differed significantly (Table 13c).

Trends in Runoff Coefficients

Both permutation and log-transformed least squares regression methods suggest a statistically significant negative trend (at the 5% significance level) over time at the CACO outfall station, and a significant positive trend at the BACI outfall station. While the log-transformed method also detected a negative trend at the CACO Instream station, this trend was not detected using the permutation method. No significant trends were found in FRCO at either the outfall or instream station (Table 14).

Trends in Loading Rates

Results from the permutation method indicate significant trends for all parameters the BACI and CACO outfall stations (Table 15a). The trends were positive at the BACI station and negative at the CACO station.

Regression using a log-link function resulted in significant positive trends for all parameters at the BACI outfall station and negative trends for all parameters, except for TSS, at the CACO outfall station (Table 15b).

Results from the logistic regression method indicate significant trends for all parameters the BACI and CACO outfall stations (Table 15c). The trends were positive at the BACI station and negative at the CACO station.

Land Cover Effects

The change in land cover and BMP implementation over time is summarized in Tables 16-18 and Figures 7a-7c.

In CACO, storm concentrations of TSS and TP decreased at the outfall with increasing treatment, with no statistically significant findings at the instream station. At the FRCO instream station, results were mixed. As the treatment index increased, there was an apparent increase in NO_{23} and a decrease in TKN (Table 19a).

Results for baseflow concentrations were different in each watershed. An increase in the treatment index was associated with decreased TP and NO_{23} and increased TKN at the CACO outfall, decreased NO_{23} at the CACO instream station, and decreased TKN and increased NO_{23} at FRCO (Table 19b).

An increase in the treatment index was associated with a significant decrease in the runoff coefficient at the CACO outfall but had no other significant results (Table 20).

The results of the unit area pollutant loading rate analyses are generally similar to the results of the runoff coefficient analyses. For all applicable pollutants/parameters, unit area loading rates decreased as the treatment index increased at the CACO outfall station. However, at the CACO instream station, the only observed decrease in unit loading rate as the treatment index increased was for NO_{23}. There were no significant results for the FRCO station (Table 21).

Power Analysis

Tables 22a-22c and 23a-23c summarize the number of samples required to detect a change over one or two, five-year permit cycles, respectively, for weak (2% per year), moderate (5% per year) and strong (10% per year) trends. The results varied somewhat by location but in general the number of samples currently required by MDE (8-12 samples per year) were inadequate to detect very strong (10% per year) changes within one permit cycle. Within two five-year permit cycles, on the other hand, the standard of 12 samples per year was adequate to detect strong (10%) changes for most parameters, and 8 samples per year was sufficient for many. For moderate (5% per year) trends, a greater sampling rate of between 12 and 24 samples per year would be necessary for most parameters to detect a change within two permit cycles. Weak (2% per year) trends could generally not be detected within two cycles, even with very frequent sampling, with most parameters requiring greater than 48 samples per year.

Discussion

Pairwise Comparisons

The results of this study are consistent with other findings (Christianson et al. 2014), which have found that watersheds with natural stream channels have higher pollutant concentrations instream than at outfalls during storm events, particularly for sediment-borne pollutants. Conversely, Baltimore City, the only watershed with an armored channel, tends to have higher concentrations at the outfall. The results point to the importance of stream channel erosion in developing watersheds. Furthermore, it supports the contention that detecting trends in

the stream channel is more difficult than detecting them at the outfall due to the processes that occur within the channel itself (Johnson et al. 2016; Peterson et al. 2001).

Trends in Concentrations

Methods

Several statistical analysis methods were used to detect trends in water quality parameters over time, including Mann-Kendall tests, log-link regression, permutation methods, logistic regression, and SARIMA models. Given the available data, the SARIMA models and Mann-Kendall method were difficult to apply due to gaps in the time series, and some of the nonparametric methods could only be applied seasonally due to limited data.

SARIMA models were the only method to find no significant trends in the water chemistry data. There were several challenges associated with this method that may have led to this outcome. First, developing these models can be time consuming since they rely on identifying coefficients that account for both the autocorrelation and moving average components of the data. Second, SARIMA models assume that the data have an even time step (e.g., monthly, daily or weekly data). While models can account for missing data, and periods with multiple observations can be aggregated, these changes can result in two problems: missing data can make it difficult to interpret correlations between data points, and aggregating data can result in a loss of information, such as when three data points become only one point. Furthermore, most models developed through this process were "stationary," with no component included for change over time. It is important to note that some nonsignificant trends were detected when standard settings were used on the "auto.arima" function, but the resulting models were problematic. In particular, the model should remove any correlations between the residuals over time, but the models developed with the standard settings showed significant correlations between the residual at a particular time point and "lag" values. The results suggest that this approach, although an excellent option for addressing some of the most common issues associated with time series data, was not very useful given the somewhat inconsistent time between samples for this data set.

Aside from the SARIMA models, the remaining methods were generally in agreement with one another. Table 24 summarizes how the methods differed in their ability to detect trends at 5% significance. In general, when one method did not find a significant trend, the others did not either. Additionally, trend directions (i.e., positive or negative) were frequently consistent, with only a few instances where one method detected a trend in the opposite direction of the majority. When this did occur, the p-values of the outlier trends were not significant.

Overall, in most cases where the methods disagreed, Mann-Kendall techniques and logistic regression methods were less likely to detect trends, whereas the log-link regression and permutation methods proved highly sensitive. In addition to being a sensitive method, permutations tended to agree with other methods, with only two occasions where the result using this method was an outlier. Conversely, the logistic and log-link regression methods had six and five instances, respectively, of being in the minority. Logistic regression is very well suited for some analyses; however, it was particularly useful for identifying trends in *E. coli* and other

parameters that are better represented by a percent exceedance. In summary, the permutation method, along with a logistic regression for extremely variable parameters such as *E. coli*, appear to be the most accurate and robust reflection of trends.

Results

BACI

In the Moores Run watershed, all applicable methods found the following parameters to significantly decrease over the monitoring time period: instream and outfall stormflow *E. coli*, instream stormflow NO_{23}, instream and outfall baseflow TCU and TZN. Additionally, all methods tested found a significant increase in TSS at the outfall site during stormflow.

Of the three watersheds, Baltimore City was the only one to experience a reduction in *E. coli*, which occurred during stormflow only. This was also the one watershed where an observed increase in the water quality parameters was common, with increasing trends identified for BOD, TSS, and TKN (Table 8a).

CACO

Significant trends of water quality parameters at the Airpark Business Center watershed were mostly negative and associated with the outfall. All available methods identified significant negative trends for outfall stormflow TP and all metals (TCU, TZN, TPB); outfall baseflow NO_{23} and TP; instream stormflow TCU; and instream baseflow NO_{23}. One out of three methods (log-link regression) found a significant decrease in instream baseflow TP. Only one parameter increased over time. Two out of three methods found that outfall stormflow BOD significantly increased (Table 8b).

The CACO watershed experienced fewer changes in water quality at the instream site compared to the outfall, which saw decreasing trends in TP and metals during stormflow. The instream station did, however, see decreases in NO_{23} and TKN during baseflow. The only positive trends were in outfall BOD and TSS during stormflow.

FRCO

Generally, analyses showed significant decreasing pollutant trends in the Urbana watershed. The only parameter where an increasing trend was observed was instream stormflow TKN, which all four applied methods determined. This increase in TKN may be due to a variety of factors, such as changes in "sanitary" factors (i.e., new sewer piping and septic systems from the development) and application of fertilizers both on the agricultural lands in the instream site's drainage area and lawns in the housing development.

At the instream site, all methods identified a significantly decreasing trend for stormflow NO_{23} and TCU, as well as baseflow TSS. At the outfall, all available methods identified stormflow TP and TCU to be significantly decreasing.

In the Frederick County watershed, more trends were identified in stormflow than in baseflow, including the only increasing parameter, TKN. Also, compared to BACI and CACO results, the trends identified at the FRCO sites were more variable as there was a relatively large number of trends identified as significant by only one or two of the applied methods (Table 8c).

Explanatory, Confounding, and Auxiliary Variables

In addition to variables discussed in the Explanatory, Confounding, and Auxiliary Variables Technical Memorandum (i.e., BMPs and land cover), other variables that may impact findings were identified during the statistical analysis. These variables include repairs to sewer pipes, changes in vegetation, erosion, and the potential impacts of seasonality.

Repairs to Sewershed

One of the findings of this study was that loading rates and runoff coefficients increased during the monitoring time period at the BACI watershed. A potential explanation may be repairs to sanitary sewer pipes in the area. Based on a report from the City of Baltimore Department of Public Works, Bureau of Water and Wastewater (2009), from June 2003 to June 2006, upwards of 7,000 linear feet of sewer piping was constructed to replace the existing middle section of the Moores Run sewer interceptor. In the lower section, more than 13,000 linear feet of piping was installed from June 2005 through June 2008. From June 2004 to June 2006, the upper section of the Moores Run sewer interceptor had almost 4,000 linear feet of piping replaced. Altogether, around 24,000 linear feet of pipes were replaced in the Moores Run sewershed in five years.

Before these repairs to the sanitary sewer took place, the sewage pipes may have been cracked or otherwise degraded such that baseflow and runoff from small amounts of precipitation could seep into the pipes as groundwater, thereby reducing the amount of surface runoff into streams. Conversely, during larger precipitation events, sewage and stormwater would be expelled from the sanitary pipes due to pressure. Fixing the degraded pipes could then have increased the amount of runoff into the stream system due to less water entering the sewershed and more flowing over ground, carrying with it the nutrients and sediments whose loading rates increased. During storm events, sewage-laden water would be retained in the pipes, which may explain the decrease in stormflow *E. coli*.

Vegetation

The BACI and FRCO watersheds experienced changes in vegetation during the monitoring time period. Beginning with BACI, based on aerial imagery (Figures 8a though 8b), the tree canopy in the stream corridor widened markedly between 1994 and 2017. Additionally, during the site visit, the tree canopy was noted to be essentially closed at the outfall and provided good coverage at the instream station (Figures 2b, 2c, 2d). Unfortunately, quantitative information associated with tree canopy cover was not available, so it could not be included in the land cover analysis.

In the FRCO watershed, the area where the pond and outfall site later were established was actively farmed prior to the start of development, and became grassy with no trees or shrubs for several years during and post-construction of the Villages of Urbana housing development, as seen in Figures 9a and 9b from 2002 and 2004 (Frederick County Division of Public Works 2003 and 2004). When the site visit was conducted in 2019, the area downstream of the pond's embankment and below the outfall was abundantly vegetated with tall grasses, shrubs, and small trees (Figure 9c). For this site, tree canopy cover was provided by MDE in the form of hand-

digitized land cover, but a GIS layer for non-tree vegetation was not available, so the increase in non-tree vegetation around the outfall was not documented and accounted for in the Principal Components Analysis.

Channel Erosion

At FRCO's instream site in Peter Pan Run, the stream bank is incised and actively eroding. This was evident in photographs taken by the Frederick County Division of Public Works (2004) in 2004 (Figure 10a), as well as from the August 2019 site visit (Figure 10b). During the site visit, it was also observed that a point or gravel bar had formed the week prior as a result of a storm event, which also caused sediments to cover the stilling well but not the intake tubing for the monitoring station (Figure 10c). The monitoring data collected supports the photographical evidence, as the instream site had particularly high TSS compared to the outfall during stormflow (Tables 6b and 7b). This addition of sediments and other constituents during stormflow can potentially mask the reductions observed at the outfall due to a variety of other tributary inputs.

Seasonality and Intensity

For the summaries in Appendix D, plots were generated of every water quality parameter against storm intensity. The data points in the plots were then colored by season (i.e., winter, spring, summer, autumn) and year. When the season-colored plots were interpreted, an interesting pattern was observed in the BACI plots. For all parameters except NO_{23} and *E. coli*, the most intense storms were associated with summertime values, but the measured values were relatively low. Springtime values were associated with low-to medium- intensity storms and had the highest measured values. Spring and winter values followed similar patterns, as did summer and autumn values. In the CACO and FRCO watersheds, this pattern was observed for a few parameters, but not nearly as strongly as in BACI, nor uniformly across the outfall and instream sites.

Another difference between the Baltimore City watershed and the other watersheds is that the City appeared to measure a greater range of storm intensities (Figure 11b). The most intense storm in BACI is more than 2 inches/hour, but in CACO and FRCO, the highest measured intensities are ~0.8 inches/hour and ~0.45 inches/hour, respectively. There is a documented weather phenomenon in the Baltimore-Washington metropolitan area called the Bay breeze, which can result in heavier rainfall events in the cities (Ryu et al. 2016). The Bay breeze may contribute to the elevated storm intensities measured in Baltimore, but another reason may stem from a difference in how and how often storms of varying intensities are sampled between jurisdictions. Baltimore City staff use stage activation to automatically trigger sample collection during storm events. This means that, unlike jurisdictions that rely on checking weather forecasts for storms to sample, Baltimore City may be recording fast-moving, high intensity storms that may not appear in a forecast with advance notice. Regardless of why intensities appear higher in BACI than in CACO and FRCO, if the total precipitation regime of a watershed is truncated due to sampling methods, estimates of runoff and loads may be skewed as they do not represent the entire distribution of rainfall events.

Detecting Effects of BMPs and Land Cover

This pilot study was not able to compare land cover between watersheds to determine if differing land cover conditions between stations resulted in variable loads or concentrations. With only three locations and apparent confounding influences, such as sewer repairs and changing land use over time, the land cover analysis instead focused on changing land cover and BMP implementation during the monitoring time period. Results suggest that the influence of stormwater BMPs and land cover can clearly be observed in pollutant concentrations, unit loads, and runoff coefficients during particular conditions. In this study, the clearest relationship was the observed benefit of the pond retrofit at the CACO outfall station. The station met several conditions that are worth considering in future studies:

1. The retrofit impacted a very large area in a short period of time.
2. There was little new development occurring in the outfall drainage area in concert with the retrofit.
3. Extensive monitoring data were available both before (1999-2007) and after (2008-2016) the retrofit.
4. The retrofit was implemented at an outfall so that instream effects did not affect the observed loads, flows, or concentrations.

For watersheds that experience a complicated land development pattern, as in FRCO, detecting the benefits of stormwater BMPs used to treat development is somewhat more challenging. In this watershed, for example, BMPs were implemented concurrently with increases in impervious cover. Consequently, separating the impacts of development and restoration can be challenging. However, it is encouraging that no trends in declining water quality conditions were not observed.

Benefits of restoration at CACO and, to a lesser extent at FRCO, were not as obvious at instream stations, possibly due to instream effects and changing land cover conditions. This result suggests that the methodology of pairing an instream station with an outfall station is beneficial, particularly if restoration occurs in the outfall drainage area.

Finally, inconsistent data collection periods can lead to challenges determining restoration impacts. It was not possible to evaluate the benefits of the single retrofit at the FRCO outfall because only one year of monitoring data was available from the period before the dry pond was retrofitted.

In general, the pollutant loading trends at the CACO outfall were predictable in that pollutant loads decrease as the treatment factor increases and the area of untreated impervious cover decreases. Although only one outfall location had both the necessary water quality data and BMP/land use change data to allow for this analysis, the results suggest that links between land use change/BMP implementation and water quality may be more easily observed at outfall stations than at instream stations.

It also appears that differences in pollutant loading are more highly driven by runoff volume than pollutant concentrations, with the unit load results for almost every parameter corresponding with the same findings for the runoff coefficients. This result may indicate that the

effects of both BMPs and land development are more predictive of runoff volume than pollutant concentrations. This effect is reasonable given that pollutant concentrations may be affected by other factors, such as stream processing, channel erosion, and practices that are unaccounted for, such as lawn care.

The inconclusive findings at the FRCO stations can partially be explained by the pattern of development in the watershed. The treatment indices for the CACO stations showed a sharp demarcation between development ("untreated" factor) versus BMPs ("treated" or "nested" factors), but coefficients for both BMP implementation and development were positive at the FRCO station. This discrepancy occurs because of a more complex development pattern at the FRCO station, wherein BMPs were primarily implemented as a part of new development. As a result, BMP implementation is correlated with an overall increase in impervious cover in the watershed. By contrast, the CACO watersheds were dominated by the effects of a single retrofit in 2008 that caused an immediate shift from the "treated" to "untreated" categories. It is possible that the BMPs implemented in the FRCO watershed as a part of development were highly effective since the results do not suggest an increase in pollutant loads or runoff volume, but it was difficult to separate the effects of BMPs from the effects of land cover change.

Another potential factor in the FRCO watershed is the land cover data included in this analysis. Aerial photographs (Appendix D, page 48) suggest that brush cover increased over time in the FRCO watershed, but this analysis focused on impervious cover. Tree canopy was also delineated, but it was not included in the analysis due to the small fraction represented in the watershed. Although the effects of land cover focused on urban land cover types, another potential influence in FRCO could relate to the agricultural land being converted to urban uses. In particular, the decrease in NO_{23} at the instream site could possibly result from a loss in agricultural land.

Impacts of BMP and Land Development on Flows

Although there is some evidence that BMPs may reduce the concentrations of pollutants, the BMPs in this study appear to have a more direct impact on flow volume, based on the results in Carroll County. Furthermore, it appears that unit loads (measured as lb/inch of rain) are generally driven by flow reductions or increases. This result is consistent with other findings and the goals of many modern stormwater programs, which focus primarily on runoff reduction as a treatment method. The results also highlight the importance of accurately measuring flows and recording rainfall characteristics.

Data Collection and Sampling Methods

Data collection and sampling methods impact the data quality in the MS4 database and can potentially affect conclusions regarding the data. As a part of this study, topics investigated included rainfall measurement, flow measurement, laboratory methods, and sample collection frequency.

Flow Measurement: Storm Events

Analyses suggested that flow is a crucial variable for tracking progress. The data suggested that reduced flow volumes were observed immediately after a retrofit in Carroll County, and flow increases were observed in Baltimore City as sewer repairs were implemented.

Laboratory Methods: Censored Values

For metals and BOD, the number of uncensored values contributed to variability in the data. In baseflow, the number of censored values prevented trend detection using certain methods. In addition to a high number of censored values, the laboratory methods changed over time, so that apparent trends were actually a relic of changing detection limits over time (Appendix B).

Sampling Method: Calculating Concentrations

Currently, the MDE sampling methodology relies on calculating the EMC using three samples: one on the rising limb, one at the peak, and one at the falling limb, called the Average Concentration Method. The project team reviewed data from Stony Run in Baltimore City to evaluate the differences between a flow-weighted EMC, versus MDE's current method. The results, completed for TN, TP, and TSS, suggest that the Average Concentration Method introduces a slight bias, with slightly higher median values for all three variables (Figures 12a through 12c; Figure 13), and increases the variability of the data. Finally, the method introduces error at each point.

Sample Collection Frequency

Currently, MDE's permit requires sampling 12 storms per year for large jurisdictions and 8 storms per year for medium jurisdictions, with no fewer than two storms in each quarter. The power analysis conducted as a part of this study suggests that this frequency, given the variability of the data at each station, is sufficient to detect very strong decreases in loads and concentrations within 10 years and moderate trends (e.g., 5% per year) within 20 years. In the current database, some stations did not meet this criterion, as there were long periods of time lacking sampling for some parameters.

Recommendations

The pilot study resulted in two sets of recommendations: 1) Next Steps In Analyzing Existing Water Quality Data and 2) Changes to the MS4 Monitoring Program.

Recommendations for Next Steps in Analysis of Existing Water Quality Data from the MS4 Program

Focus on watersheds where restoration impacts can be detected

For future work in evaluating existing water quality data from the MS4 program, the first suggestion is to be very selective when choosing which watersheds to use in the statistical analyses. If possible, prioritization for data analysis should be given to watersheds that have: 1) one or a few larger BMPs or several smaller BMPs implemented over a relatively short period of time; 2) data from before and after watershed restoration practices are implemented; and 3)

limited development over time. Additionally, data from each candidate watershed should be carefully examined for issues with large gaps or sample clustering during a portion of the year. The watersheds studied for this project did have long monitoring histories with limited gaps on the surface, but when individual parameters were assessed, issues were noted with sampling frequencies.

Select appropriate statistical techniques

Next, certain methods are recommended for future trend analysis. Permutation methods are suggested because they do not require assumptions about the data's distribution or homoscedasticity, making them very useful for testing environmental data, which are often skewed (Elliffe and Elliffe 2019). Moreover, when compared to other methods used for trend analysis, permutation appeared to be sensitive, not only agreeing well with other methods, but also tending to find significant trends when other methods did not. The log-link method also performed well, with similar results as permutation methods. Methods that require an equal-interval time series, such as the Mann-Kendall, Seasonal Kendall, and SARIMA, are robust, but they require either a very long time series or a fine and regular time step with few long gaps, and are consequently only recommended for data with those characteristics. Logistic regression methods tended to not find a trend when others did, but they performed well for highly variable parameters with outliers, such as *E. coli*.

Understand changes to the landscape that may not be apparent from the MDE database or readily available land cover data

In this study, changes in impervious cover, BMP implementation, and tree canopy were incorporated into the analysis, but other land cover changes such as growth in low-lying shrubs or other vegetation may also be considered. The next analysis should also consider the impacts of buried infrastructure (e.g., sewer, stormwater, and drinking water infrastructure) on results. In this study, extensive sewer work in the City of Baltimore seemed to result in greater storm runoff volumes. Without understanding these dynamics, some outcomes would be challenging to explain.

Continue to incorporate seasonal variability and rainfall characteristics into the analysis.

Further consideration and exploration of confounding, auxiliary, and explanatory variables should be incorporated in future work. Seasonal variation of rainfall depth, intensity, and other storm attributes can explain why certain patterns are observed in concentrations.

Conduct field visits

Finally, field visits to the watersheds of interest are recommended. Such visits are essential to understanding watershed characteristics that are not apparent from the database alone. In this study, field visits provided insights about areas of active channel erosion, sewer repairs, sampling methods, sampling constraints, and regrowth of vegetation.

Recommendations for Changes to the MS4 Monitoring Program

Both quantitative analyses and observed data gaps provided sufficient background to improve the MS4 monitoring program, including methods for flow measurement, EMC computation, reporting of censored values, and number of storms sampled.

Develop a QAPP for MS4 monitoring

Inconsistencies between MS4s and lack of information regarding how samples were collected, analyzed, and reported was a challenge in using data in the MDE database. In the future, it is recommended that MS4 permittees who do not currently have a QAPP filed with MDE do so, and that the QAPPs are distributed to the individuals performing the statistical analysis.

Provide more information regarding flow measurements

The MDE database currently reports a single value for stormflow (i.e., the volume of stormflow in gallons). QAPPs should provide detail regarding how stormflow is measured, including specific methods to calculate runoff volume, including when to start and discontinue flow measurements. In addition, the database should include some measure of discharge both instream and at the outfall for both storm and non-storm events. The lack of discharge data made it difficult to use any methods that employ flow-correction to detect changes or trends.

Develop a specific protocol for reporting non-detected values

In this study, censored values were problematic for some parameters because they were not treated uniformly across jurisdictions, and changing methods resulted in false trends for some parameters. The following specific measures can help to improve this issue in the future:

1. Record the instrument-measured value rather than the detection limit. This value helps to characterize the variability of censored values.
2. The detection limit should be reported even if the reported value is above the limit. The database has some entries of "0" or "NA" for the detection limit.
3. It appeared that some detection limits were incorrect in the database (in some cases a single value orders of magnitude higher than the others). When methods do change, the MS4 or laboratory should provide a note or other indication that the method has changed, or if an alternative method needs to be used due to sample characteristics such as salinity or turbidity.

Sampling frequency

The current sampling frequency of 12 events per year was inadequate to detect even strong (10% per year) change in one permit cycle for most parameters, but could detect this change within two permit cycles. Greater sampling frequency (up to 24 per year) would still be unable to detect even strong trends within one permit cycle, but could enable detection of moderate (5% per year) trends within two permit cycles for many parameters. Weak trends (2%) could generally not be detected even with very frequent monitoring. Given the number of storms that can reasonably be sampled annually, the current rate of sampling may be adequate, but only relatively strong trends can be detected using this frequency, except over longer periods of time. In addition, the expectation of detecting change within a single permit cycle may be unrealistic, even with very large sampling rates.

Sampling storm diversity

Although the sampling record suggests that a wide range of storms was sampled at these locations, some monitoring designs may exclude large and quickly-moving storms. This would

be the case both in jurisdictions where staff are prevented from collecting high-flow discharge measurements due to safety concerns and for jurisdictions that rely on weather forecasts to predict when to set up samplers for storm events, rather than having stage-activated pressure transducer samplers. The MS4 permit should clearly outline the range of storms that can be monitored at a given location.

EMC calculation

Currently, MDE's standards require calculation of an "Average EMC," which averages the rising limb, peak, and falling limb of a storm. An analysis completed for this study suggests that the method introduces a slight bias toward higher concentrations when compared with a flow-weighted method for the same storm events. This finding was consistent with other research (Ma et al., 2009) which found that approximately 30 grab samples per event were required to estimate a flow-weighted composite sample within 20%. A flow-weighted sample is a better estimate of the true concentration because, when multiplied by the event runoff volume, it results in an unbiased estimate of the pollutant load (Gulliver et al., 2012). In addition, the method of EMC collection (time-weighted, flow-weighted, or grab sample; and discrete vs. composite) should be reported along with the EMC.

References

Battiata, J., Collins, K., Hirschman, D., and Hoffmann, G. 2010. The Runoff Reduction Method. *Journal of Contemporary Water Research & Education*. *146(1).*11-21. Available online at https://doi.org/10.1111/j.1936-704X.2010.00388.x.

Buchanan, C. and Mandel, R. 2015. Water Quality Trend Analysis at Twenty-Six West Virginia Long-Term Monitoring Sites. ICPRB Report 14-6. Interstate Commission on the Potomac River Basin, Rockville, MD.

Christianson, R., Fraley-McNeal, L., Law, N., and Stack, B. 2014. Technical Memorandum: Analysis of Stream Sediment Monitoring in Support of Objective 1 of the Sediment Reduction and Stream Corridor Restoration Analysis, Evaluation and Implementation Support to the Chesapeake Bay Program Partnership. Prepared for: Chesapeake Bay Program Office. Annapolis, MD.

City of Baltimore Department of Public Works, Bureau of Water and Wastewater. 2009. City of Baltimore Sanitary Sewer Overflow Consent Decree. Retrieved online from: https://publicworks.baltimorecity.gov/sites/default/files/QT%20Report%2029.pdf.

Elliffe, D. and Elliffe, M. 2019. Rank-permutation tests for behavior analysis, and a test for trend allowing unequal data numbers for each subject. *Journal of the Experimental Analysis of Behavior. 111(2)*. 342-358. Available online at: https://doi-org.proxy-bc.researchport.umd.edu/10.1002/jeab.502

Fay, M. and Shaw, P. 2010. Exact and Asymptotic Weighted Logrank Tests for Interval Censored Data: The interval R package. Journal of Statistical Software. Retrieved from: http://www.jstatsoft.org/v36/i02/. 36 (2):1-34.

Fraley-McNeal, L. Draft submitted to MDE June 2019. Summary of Methods used for Explanatory Variables Data Analysis.

Frederick County Division of Public Works. 2003. National Pollutant Discharge Elimination System Annual Report.

Frederick County Division of Public Works. 2004. National Pollutant Discharge Elimination System Annual Report.

Gulliver, J.S., A.J. Erickson, and P.T. Weiss (editors). 2010. "Stormwater Treatment: Assessment and Maintenance." University of Minnesota, St. Anthony Falls Laboratory. Minneapolis, MN. http://stormwaterbook.safl.umn.edu/

Helsel, D. and Hirsch, R. 2002. Statistical Methods in Water Resources Techniques of Water Resources Investigations, Book 4, Chapter A3. U.S. Geological Survey. Retrieved from: https://www.practicalstats.com/resources/Helsel&Hirsch.PDF

Hyndman, R. and Khandakar, Y. 2008. "Automatic time series forecasting: The forecast package for R." *Journal of Statistical Software.* 26(3).

Jepsen, R. and Caraco, D. Revised draft submitted to MDE August 2019. A Technical Memorandum Describing the Exploratory Analysis of the MS4 Pilot Watersheds

Johnson, T., Newcomer, A., Kaushal, S., Mayer, P., Smith, R., and Sivirichi, G. 2016. Nutrient Retention in Restored Streams and Rivers: A Global Review and Synthesis. *Water, 8(4)*: 116. DOI:10.3390/w8040116

Leisenring, M., Clary, J., Lawler, K., and Hobson, P. 2011. International Stormwater Best Management Practices (BMP) Database Pollutant Category Summary: Solids (TSS, TDS and Turbidity). International Stormwater BMP Database. Available online at http://www.bmpdatabase.org/Docs/BMP%20Database%20Solids%20Paper%20May%202011%20FINAL.PDF

Ma, J.S., Kang, J.H., Kayhanian, M. and Stenstrom, M. Sampling Issues in Urban Runoff Monitoring Programs: Composite versus Grab. Journal of Environmental Engineering, 135, 3(118). https://doi.org/10.1061/(ASCE)0733-9372(2009)135:3(118)

McLeod, A. 2011. Kendall: Kendall rank correlation and Mann-Kendall trend test. R package version 2.2. Available online at: https://CRAN.R-project.org/package=Kendall

Nagel, A. 2019. Analysis of Water Chemistry Data Collected Under Maryland's Municipal Separate Storm Sewer System (MS4) Permits: Database, Trends, Challenges, And Recommendations. ICPRB report prepared for Maryland Department of the Environment.

Nagel, A. and Mandel, R. 2018. Analysis of Monitoring Data Collected under Maryland's Municipal Separate Storm Sewer System (MS4) Permits: Database Design and Preliminary Analysis of Water Chemistry. ICPRB report prepared for Maryland Department of the Environment.

Olson, M. 2005. Seasonal Flow Characterizations for the Principal Tributaries of Chesapeake Bay 1984-2004. Report to the Chesapeake Bay Program, Tidal Monitoring and Analysis Workgroup.

Patakamuri, S. and O'Brien, N. 2019. modifiedmk: Modified Versions of Mann Kendall and Spearman's Rho Trend Tests. R package version 1.4.0. Available online at: https://CRAN.R-project.org/package=modifiedmk.

Peterson, B., Wollheim, W., Mulholland, P., Webster, J., Meyer, J., Tank, J., Martí, E., Bowden, W., Valett, H., Hershey, A., McDowell, W., Dodds, W., Hamilton, S., Gregory, S., and Morrall, D. 2001. Control of Nitrogen Export from Watersheds by Headwater Streams, Science, 292: 86-90.

Pitt, R. 1999. Small Storm Hydrology and Why it is Important for the Design of Stormwater Control Practices. Advances in Modeling the Management of Stormwater Impacts, Volume 7. (Edited by W. James). Computational Hydraulics International, Guelph, Ontario and Lewis Publishers/CRC Press, 1999.R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available

online at: http://www.R-project.org/.Schueler, T. 1999. The Importance of Imperviousness. *Watershed Protection Techniques,* 1(3): 100-111. Available online at: http://scc.wa.gov/wp-content/uploads/2015/06/The-Importance-of-Imperviousness_Schueler_2000.pdf.

Ryu, Y.-H., Smith, J., Bou-Zeid, E., Baeck, M. 2016. The Influence of Land Surface Heterogeneities on Heavy Convective Rainfallin the Baltimore–Washington Metropolitan Area. *Monthly Weather Review. 144(2)*. 553-573. https://doi.org/10.1175/MWR-D-15-0192.1