ANL/EVS-20/6

Chicago Area Waterway System Microbiome Research — Phase III Final Report

Principal Investigators: M. Cristina Negri, Jack A. Gilbert and Mark Grippo
Research Team: Anukriti Sharma, Jules F Cacho, Melissa Dsouza and Patty Campbell
Environmental Science Division, Argonne National Laboratory

Prepared for
Metropolitan Water Reclamation District of Greater Chicago

July 2020

CONTENTS

CONTENTS (Cont.)

FIGURES

FIGURES (Cont.)

FIGURES (Cont.)

FIGURES (Cont.)

FIGURES (Cont.)

FIGURES (Cont.)

FIGURES (Cont.)

FIGURES (Cont.)

FIGURES (Cont.)

TABLES

TABLES (Cont.)

TABLES (Cont.)

TABLES (Cont.)

This page intentionally blank

OVERALL SUMMARY

This report summarizes a seven year investigation of the microbial communities within the Chicago Area Waterway System (CAWS). This study coincides with the Metropolitan Water Reclamation District of Greater Chicago's (MWRD) efforts to implement disinfection and storm water reservoir control management. MWRD initiated construction disinfection facilities at the Calumet and O'Brien Water Reclamation Plants (WRPs), which were completed by the end of 2015. Ultraviolet radiation based disinfection technology was implemented at O'Brien WRP, which was expected to kill harmful bacteria and pathogens released into the North Shore Channel and North Branch of the Chicago River. Chlorination and dechlorination based disinfection technology was implemented at the Calumet WRP. Furthermore, the Thornton Composite Reservoir (TCR), which is part of the Tunnel and Reservoir Plan (TARP) within the service area of the Calumet WRP, became operational by December in 2015. This tunnel system provides 7.9 billion gallons of storage and since its completion has captured more than 11.0 billion gallons of combined stormwater and sewage from Calumet WRP that would otherwise have overflowed into the CAWS in rainy weather and high flow condition.

Using molecular DNA sequencing techniques we characterized the bacteria found in water and sediment at 16 different sites across the CAWS for a seven years period (2013–2019). This analysis provides the most detailed investigation of microbial dynamics in an urban waterway ever attempted. In addition to helping us to understand how both the native and bacterial communities from various sources varied with weather and location, we were also able to quantify the impact of the two water quality improvement efforts taken up by the MWRD, providing an unparalleled opportunity to determine if these efforts reduced microbial contamination, but also whether they impacted native river water and sediment bacterial community structure and function.

OS.1 GOALS AND OBJECTIVES

The overall goal of this study was to both understand the native microbial composition of the CAWS and determine how the implementation of disinfection at the two WRPs and the TCR, which is associated the Calumet TARP system, affected microbial communities across the CAWS, including directly downstream of the two WRPs. We also examined the impact of dry and wet weather events and Combined Sewer Overflows (CSOs) on the microbial community. Other objectives included:

- Identifying probable sources of bacteria found in the CAWS;
- Characterizing the distribution of virulence associated genes across both space and time in the CAWS;
- Characterizing the relationship between microbial communities and fecal coliform bacteria in dry and wet weather events, with and without CSOs;

- Characterizing the relationship between water chemistry and the community diversity and composition in the CAWS river ecosystem,

The report also describes the CAWS-Fecal Indicator Bacteria (FIB) model. Using chemical data, climate data, and hydrology data model as inputs, we developed a model to 1) predict fecal indicator bacteria concentrations at any point along the CAWS, and 2) estimate the probability that fecal coliform density will exceed the regulatory standard.

OS.2 METHODS

In this study, we have employed molecular genetic analysis methods to track microbial diversity at 16 different sites covering the Calumet River System, North Chicago River, Main stem, South Branch Chicago river, and South Fork River System from 2013–2019. To better understand the microbial sources impacting the CAWS quality, we collected multiple sample types—river water, river sediment, beach water, effluent discharged from the two WRPs, raw sewage inflowing into the two WRPs, fish mucus and fish guts and water from Lake Michigan beaches. Samples were divided into two groups: 2013–2015 (pre-disinfection/pre-TARP period) and 2016–2018 (post-disinfection/TARP implementation period).

In developing the CAWS-FIB model, we used both classical statistical methods and machine learning approaches. Machine learning is the subfield of computer science that allows computers to learn without being explicitly programmed. Each modeling approach was compared for accuracy in predicting fecal indicator bacteria and the probability of exceeding the regulatory standard. In addition, we attempted to identify the most important variables for predicting fecal coliform levels.

OS.3 RESULTS

Microbiome Analysis. Molecular characterization of bacteria from samples collected between 2013 and 2019 indicate that the CAWS maintains a diverse bacterial community, with more than thirty thousand species of bacteria in the water and sediment alone. The composition and diversity of this community is significantly regulated by variables such as weather (dry and wet events), the presence of CSOs, physicochemical water properties, and the presence of absence of a CSO. Sample type (sediment, water, etc.) and the location of a sampled site also produce distinct microbiome signatures, and importantly, disinfection applied to the WRPs had a significant impact on the microbiome. Key findings include:

- Compared to the pre-disinfection period (2013–2015), in the four years post-disinfection (2016–2019) we observed a sequential decrease of sewage and fecal indicator bacteria in both the final outgoing treated effluent and river water samples. These results suggest the ongoing disinfecting activity was effective in reducing sewage and fecal indicators bacteria discharge at both O'Brien and Calumet WRPs.

- Across the CAWS there was an increase in sewage indicator bacteria during the wet events (at sites both with and without the CSOs). However, with the implementation of Calumet TARP system, we observed a decrease in the sewage indicators in the Calumet region. These results are in agreement with the observations that CSO events in the Calumet Region were significantly lower following implementation of disinfection/TARP. This highlights the efficiency of Calumet TARP system which captured the CSO discharges in the reservoir. In the CAWS north (O'Brien WRP) also, we observed a significant reduction in the sewage indicators at the downstream sites in the years 2016–2019, which can be attributed to the disinfection process.

- Metagenome sequencing data revealed a very low abundance (0–0.25%) of virulence associated genes in the CAWS sites across all the years of collection (2013–2019), which suggests that pathogenic bacteria are also low in abundance. However, this study cannot determine whether or not this could be associated with a reduced risk of disease in people who have surface water contact.

Furthermore, the reduction in sewage associated bacteria following the disinfection/TARP implementation, demonstrated by the molecular analysis, was confirmed by the results of MWRD's monthly fecal coliform plate counts. Fecal coliform concentrations in the treated effluent water were significantly lower in the post-disinfection/TARP implementation period (2016–2019) compared to the pre-disinfection/TARP implementation period (2013 to 2015) at the O'Brien WRP and Calumet WRP. Fecal coliform concentrations were significantly lower in the 2016–2019 period compared to 2013–2015 in river water samples from the North Shore Channel and North Branch of the Chicago River sites downstream of the O'Brien WRP, Little Calumet River and Cal-Sag Channel mainstem sites below the Calumet WRP. Fecal coliform concentrations in the mainstem Chicago River and South Branch of the Chicago River were not significantly different between the pre and post disinfection/TARP implementation periods. This finding may be due to the distance of these sites from the WRPs, strongly suggesting fecal coliform concentrations at these sites are influenced by sources other than WRP discharge.

CAWS-FIB Model. We found that none of the modeling approaches had satisfactory accuracy levels during model testing. The performance is likely related to the low, monthly, sampling frequency for fecal coliform, which limits the amount of data available for model training. However, we did produce an operational model with multiple functionalities and provided information that is critical for the success of future CAWS-FIB model development. For example, the results of our study suggest an artificial neural network model performed the best of all the algorithms examined. The model results also provided list of the most relevant explanatory variables influencing fecal coliform concentrations varied with site.

Improving the current CAWS-FIB model should start with collecting fecal coliform samples 50 times per water quality sampling site per month from March to October. The increased sampling frequency would also benefit the model by taking FIB samples during CSO (gravity or pumped) discharges. However, this monitoring frequency will be difficult to achieve in reality given the effort required. In addition, the CAWS-FIB model should be "re-trained" and "re-tested" on an annual-basis, thereby taking into account important, incremental changes in the system such as land use/land cover (LULC) and structural changes in the hydraulic system (e.g.,TARP).

TECHNICAL SUMMARY

This report summarizes a seven year investigation of the microbiome associated with the Chicago Area Waterway System (CAWS). This includes characterizing the changes in the microbial community structure and function in water and sediment that associate with dynamics such as weather and the Metropolitan WRP District of Greater Chicago's (MWRD) efforts to implement disinfection and storm water reservoir control management. MWRD initiated construction of two large disinfection facilities at the Calumet and O'Brien Water Reclamation Plants (WRPs), which were completed by the end of 2015 ahead of recreational season of 2016. Ultraviolet radiation-based disinfection technology was implemented at O'Brien WRP, which was expected to kill harmful bacteria and pathogens released into the North Shore Channel and North Branch of the Chicago River. The O'Brien WRP treats an average of 230 million gallons of wastewater per day (mgd), has the capacity to treat 450 mgd, and serves more than 1.3 million people within 143 square miles, making it the world's largest ultraviolet disinfection facility. At the Calumet WRP, chlorination and dechlorination based disinfection technology was used. The Calumet WRP has a capacity to treat 400 mgd and serves a population of more than 1 million people in an area of about 300 square miles. Furthermore, as part of the Chicago area Tunnel and Reservoir Plan (TARP) that is designed to intercept the combined sewer overflow (CSO) discharges, the Thornton Composite Reservoir (TCR) in the Calumet WRP service area became operational in December 2015. This tunnel system provides 7.9 billion gallons of storage and since its completion has captured more than 11 billion gallons of combined stormwater and sewage from the Calumet WRP that would otherwise have overflowed into the CAWS during rainy weather and high flow conditions.

Using metagenomic and amplicon DNA sequencing techniques we characterized the taxonomic and functional dynamics of the bacteria community from both water and sediment across 16 different sites (Calumet River System, North Chicago River, Main stem, South Branch Chicago River, and South Fork River) over a seven years period (2013–2019). This analysis provides the most detailed investigation of microbial dynamics in an urban waterway ever attempted. As well as helping us to understand how both the native and wastewater-associated bacterial communities from various sources varied with weather and location, we were also able to quantify the impact of the two water quality improvement efforts employed by the MWRD. This provided an unparalleled opportunity to determine if these efforts reduced wastewater associated microbial contamination, as well as determining the impact on the native bacterial community structure and function. As well as water and sediment from the rivers, we also collected water from Lake Michigan beaches, final disinfected effluent discharged from the two WRPs, raw sewage inflowing into the two WRPs, and the mucus and gut contents of fish living in the rivers, so as to understand the microbial sources impacting the composition of the CAWS microbial community.

TS.1 GOALS AND OBJECTIVES

The overall goal of this study was to (i) characterize the temporal and biogeographic native microbial composition of the CAWS, (ii) determine the potential sources of bacteria into the CAWS, and (iii) determine the impact on these characteristics by the MWRD's two major improvement efforts, i.e., disinfection at the two WRPs and the TCR capturing Combine Sewer Overflows (CSOs) in the Calumet WRP service area. In addition, this study aimed to explore the distribution of resistance and virulence associated genes across both space and time in the CAWS and build correlative models to predict bacterial burden (especially fecal coliform bacteria) across the CAWS in association with weather events.

Our specific objectives include:

1. To characterize the microbiota in CAWS water and sediment samples using 16S rRNA amplicon sequencing, so as to characterize the community composition and alpha/beta diversity across 16 locations over seven years (including pre- and post-disinfection years). Please note that the sediment sampling was conducted from 2013–2018 and not for 2019.

2. To determine the impact of dry and wet weather events on the microbial dynamics of the CAWS specifically upstream and downstream of the WRPs. We further attempted to catalogue the impact of the disinfection and TARP interventions during the dry and wet events across different CAWS sites, including the impact of stormflow events.

3. To determine the correlation between fecal coliform data (CFU/100 ml) and 16S rRNA amplicon and metagenomic species proportions. This was performed across wet, dry and CSO events at sites associated with the Calumet and O'Brien WRPs, main CAWS stem and the south branch.

4. To characterize the probable sources of specific bacterial strains throughout the CAWS data, using the SourceTracker algorithm to assign Bayesian probability against both local environmental sources and those collected in the Earth Microbiome Project database.

5. To identify functional gene dynamics (e.g., antibiotic resistance and virulence genes) and genotype-traits that associate with environmental and weather events, as well as those impacted by the MWRD interventions.

6. To determine how microbial community dynamics correlated with changes in the physicochemical properties of the river water (e.g., pH, dissolved oxygen, nitrate, temperature, specific conductance, flow, ammonia, total organic carbon, turbidity, chlorophyll, and volatile suspended solids). Further, these physicochemical characteristics were used to define water quality indices that could be associated with changes in the microbiome.

7. To develop and test a CAWS-Fecal Indicator Bacteria (FIB) model to explore the applicability of using a data-driven modeling platform for predicting FIB concentrations in the CAWS. Using chemical data, climate data, and hydrology data generated by Duflow modeling as inputs, we explored the use of both classical statistical methods and machine learning (ML) algorithms to 1) predict FIB concentrations at any point along the CAWS, and 2) estimate the probability that a predicted FIB density will exceed the regulatory standard for fecal coliform for a given set of environmental conditions.

TS.2 METHODS

In this study, we have employed genetic analysis methods to track microbial diversity at different sites covering the Calumet River System, North Chicago River, Main stem, South Branch Chicago river, and South Fork River System across the seven years period (2013–2019). For this we collected seven different sample types i.e., river water, river sediment, beach water, effluent discharged from the two WRPs, raw sewage inflowing into the two WRPs, fish mucus and fish guts, across different sites throughout the CAWS. All microbiome analyses were performed on sites upstream and downstream from the WRP and tributaries sites (86 – Grand Calumet River; 56 – Little Calumet River Upstream of WRP; 76 – Little Calumet River Downstream of WRP; 57 – Little Calumet River; 59 – Cal-Sag Channel; 43 – Cal-Sag Channel; 96 – North Branch Chicago River; 112 – North Shore Channel Upstream of WRP; 36 – North Shore Channel Downstream of WRP; 73 – North Branch Chicago River; 100 from the main stem; 108 from South Branch Chicago river; and 99 from South Fork River System).

In developing the CAWS-FIB model, we examined multiple machine learning approaches. Machine learning is the subfield of computer science that allows computers to learn without being explicitly programmed. Testing data for each of these sites were divided into two groups: 2013–2015 (pre-disinfection/pre-TARP period) and 2016–2018 (post-disinfection/TARP implementation period). We also compared the accuracy of the machine learning models to classical statistical methods.

TS.3 OVERVIEW OF KEY RESULTS

Microbiome Analysis. The CAWS maintains more than thirty thousand bacterial taxa from the sampled water and sediment alone. The differences in microbial diversity are significantly associated with variables such as disinfection status, weather conditions i.e., dry and wet events with and without CSOs, sample type, site location (i.e., distance from the WRPs). The differences in microbial community structure are significant correlated with changes in the physicochemical properties of the river water.

The microbial diversity (measured by alpha diversity indices) of the river water (at sites downstream of the WRPs) and treated effluent (post-disinfection) samples was significantly lower in the year 2016 compared with 2013–2015. This reflected a significant decrease in fecal indicator bacteria and microbial taxa associated with sewage. The microbial diversity of river water, sediment, sewage and treated effluent samples increased and remained stable from 2017–2019 compared with 2016.

The four years post-disinfection phase (2016–2019) demonstrated a sequential decrease in the proportion of sewage and fecal indicator bacteria in samples downstream of WRPs; a continued increase in the proportion of native river water associated bacteria was also observed, highlighting the ongoing impact of disinfection.

Between 2014–2015, wet weather events, both with and without CSO discharge, were associated with a significant increase in the proportion of sewage associated and fecal indicator bacteria, including Acinetobacter, Arcobacter, and Bacteroides. Between 2016–2019, post disinfection/TARP implementation at Calumet, the proportion of sewage-associated and fecal indicator bacteria significantly decreased, while the proportion of Flavobacterium (a river water indicator organism) significantly increased, during wet weather events compared to 2014–2015. CSO events were also significantly decreased following TARP implementation. In the CAWS north (O'Brien WRP region) we observed a significant reduction in the proportion of sewage-associated and fecal indicator bacteria downstream between 2016–2019.

Fecal coliform plate count abundances significantly positively correlated with sewage-associated and taxa traditionally identified as fecal indicator organisms. Taxa both positively and negatively correlated with the plate count abundances had significant different biogeographic distributions, suggesting that the presence of sewage-associated and fecal-indicator bacteria results in a decreased proportion of native river water-associated bacteria.

Functional gene proportions from metagenomic data were significantly different between wet and dry weather events. Microbial communities within the CAWS river exploit unique anabolic and catabolic pathways to derive and store energy from the legacy organic compounds associated with historic pollution present in the water. The proportion of 30 virulence genes from potentially pathogenic bacteria was small compared to non-virulence associated genes (0–0.25% of all gene sequences). This suggests that pathogenic bacteria are also at low proportions (which validates the amplicon sequence data).

Fecal Coliform Plate Count Analysis. Applicable primary contact recreation designated CAWS waters are subject to general use water quality standards (Illinois Administrative Code Title 35, section 302.209) for fecal coliform, however the recreational season is considered March through November according to Effluent Disinfection standards outlined in Illinois Administrative Code Title 35, Section 304.224) which specify that 1) the geometric mean of five samples should not exceed 200 CFU/100 ml within 30 days and 2) no more than 10% of samples should exceed 400 CFU/100 ml during any 30-day period. These standards do not apply to incidental contact or non-contact recreational water. Because samples were collected monthly for assessment purposes, overall annual geomean to be less than 200 CFU/100 mL and the limit of 400 CFU/100 mL was used to evaluate the compliance.

The geometric mean fecal coliform concentration in the river water and treated effluent water was significantly lower in the post-disinfection/TARP implementation period (2016–2019) compared to the pre-disinfection/TARP implementation period (2013 to 2015) at the North Shore Channel, North Branch of the Chicago River sites downstream of the O'Brien WRP and Little Calumet River and Cal-Sag Channel mainstem downstream of the Calumet WRP; this validates the 16S rRNA and shotgun metagenomic analysis results. Fecal coliform exceedances of 400 CFU/100 mL occurred in almost all samples collected between 2013 and 2015, but were only found in a small fraction of samples collected in 2016–2019.

CAWS-FIB Model. Model training and testing indicated that overfitting was a problem for the CAWS-FIB models. This meant that regardless of the approach used (ML or classical statistics-based), the models performed well when predicting fecal coliform density during training, but performed poorly during model testing. The overfitting is likely related to the low, monthly, sampling frequency for fecal coliform, which limits the amount of data available for model training.

While the modeling approaches employed in this study fell short of consistently demonstrating acceptable predictive capabilities, we did produce an operational model with multiple functionalities and provided information that is critical for the success of future CAWS-FIB model development. Results of model tests showed that the ANN algorithm was the only model that produced reasonable fecal coliform concentration values. The ANN algorithm was also relatively robust in predicting the fecal coliform concentration in the CAWS despite of the limited sampling data available for model training and testing. The model results also provided a short list of the most relevant explanatory variables influencing fecal coliform concentrations. Thus the results of our study suggest the ANN model is the future algorithm of choice for the CAWS-FIB model as it outperformed the other three algorithms across most of the performance metrics.

Improving the current CAWS-FIB model should start with collecting fecal coliform samples 50 times per water quality sampling site per month from March to October. The increased sampling frequency would also benefit the model by taking FIB samples during CSO (gravity or pumped) discharges. However, this monitoring frequency will be difficult to achieve in reality given the current monthly sampling regime. In addition, the CAWS-FIB model should be "re-trained" and "re-tested" on an annual-basis, thereby taking into account important, incremental changes in the system such as land use/land cover (LULC) and structural changes in the hydraulic system (e.g.,TARP).

This page intentionally blank

1 MICROBIOME STUDY OF THE CHICAGO AREA WATERWAYS SYSTEM (CAWS)

1.1 INTRODUCTION

High quality fresh water is a fundamental natural asset which is being consistently undermined by human activities (Vörösmarty et al. 2010). With over half of the total population living in urban areas, urban waters act as a liaison between human activity and natural environment and therefore are mostly influenced by human interference (Sandra L. McLellan, Fisher, and Newton 2015). Microorganisms establish the quality of fresh water and for a long time have been utilized as markers of poor water quality due to their traceability to potential sources of contamination. For instance, *Escherichia coli* or coliform bacteria are broadly used to identify fecal contamination in drinking and recreational water (McLain et al. 2011). However, these tests heavily depend on the optimization of culture conditions in the laboratory and can generate both false negatives as well as false positives (McLain et al. 2011). Select pathogen Polymerase Chain Reaction (PCR)-based methods have also been used to characterize the CAWS microbial quality; however, these methods are limited in their ability to resolve the source of fecal and/or sewage contamination (Dorevitch et al, 2012; Rijal et al., 2003, 2009, and 2011). These methods do not completely describe the diversity of microbial communities present in the CAWS. Molecular gene and genomic sequencing can augment, for qualitative analyses, typical culture-based methods that currently only detect approximately 8% of known microbes.

Microbial communities can be characterized in terms of levels of diversity (e.g., richness, evenness), composition (which taxa are present), and functional potential (which protein encoding genes they have). Microbial diversity fluctuates with environmental factors that shape which bacteria can survive and thrive, such as temperature, nutrient availability, hydrology, metal contamination, and the myriad factors that control these properties (Zeglin 2015). Using molecular approaches such as 16S rRNA gene amplicon sequencing and shotgun metagenomic sequencing it is now possible to get a detailed profile of the taxonomic and functional composition of a microbial community. These approaches can help us to determine exactly which bacterial taxa, and which functional traits, are important in enabling a microbial community to respond to changes in the environment. These approaches have been used to explore riverine microbial communities in the past, however, many of these studies have employed a limited timeframe of investigation (e.g., only months or one year) and with no replicates to allow for appropriate statistical analysis (Eraqi et al. 2018; Van Rossum et al. 2015; Hu et al. 2014; Jackson et al. 2014; Read et al. 2015; Savio et al. 2015; Yongming Wang et al. 2015).

The microbial community in a river system, also referred to as a microbiome, can also act as a sentinel of change in the environment, exactly because it is so sensitive a biomarker of fluctuations in environmental parameters. Therefore, the microbiome can be used as a biomarker of pollution and other anthropogenic impacts, and potentially could be used to predict different risk variants for human health outcomes associated with the river system. An integral part of the urban water cycle is sewer infrastructure. Thousands of miles of pipes line cities as part of wastewater and stormwater systems. As stormwater and sewage are released into natural

waterways, traces of human and animal waste-associated microbes are also released, and by measuring these components scientists have been able to use such traces to reflect the sources and magnitude of fecal and sewage pollution in a system. The same paradigm can also be used to indicate the presence of pollutants, such as nutrients and chemicals. However, microbial traces associated with other components of a urban environment are also released when stormwater and sewage are allowed to re-enter the river system. Runoff from impervious surfaces delivers microbes from soils, plants and the built environment to stormwater systems. Further, urban sewer infrastructure contains its own unique microbial community seemingly adapted to this artificial habitat.

Here we present the results of a 7 year investigation (2013–2019) to employ molecular microbial characterisation approaches to augment existing data on the presence of microbes in the Chicago Area Waterway System (CAWS). By employing 16S rRNA amplicon and shotgun metagenomic sequencing approaches, this study aims to determine the temporal and biogeographic patterns of the microbial community across the different areas of the CAWS over time. These approaches capture the whole microbial community and can be used to elucidate it's functional potential (what nutrients the bacteria need to survive, how they process energy, what kind of xeno-chemicals they can degrade, etc.); they can also be used to infer the different sources of bacteria that find their way into the CAWS. Potential sources include effluent from WRPs, direct stormwater runoff, sediment resuspension, wildlife and combined sewer overflows (CSOs). In addition to investigating the microbial dynamics between different sites over the years, we have further focused on the impact of the MWRD's efforts to provide a cleaner water system.

MWRD implemented two disinfection systems in 2016 at the Calumet and O'Brien WRPs. At O'Brien, a UV based disinfection system was introduced; whereas at the Calumet plant a chlorination/dechlorination approach was implemented; both strategies have a common aim to disinfect the treated wastewater prior to its reintroduction into the CAWS.

Another major initiative taken up by MWRD is the Tunnel and Reservoir Plan (TARP) System's Thornton Composite Reservoir (TCR) that became operational by end of November in 2015. The TCR in the Calumet WRP service area was constructed to capture the CSO discharge that otherwise flows into the Calumet river system. Our sampling strategy started in 2013 and continued through the end of 2019, as such we have two phases, pre-disinfection/TARP (2013–2015) and post-disinfection/TARP (2016–2019); the multiple years of investigation provide an opportunity for robust statistical interpretation of the impact of disinfection on the microbial community dynamics of both the native and introduced microbiome in the CAWS. The combination of microbiome sequencing along with the available fecal coliform quantitative data, provides us with a formidable dataset with which to characterize the water quality (in terms of fecal and sewage indicators and native microbial biomarkers) which could further guide the management decision of recreation use.

To date, this is the first study to investigate the longitudinal and spatial impact of disinfection on the microbial ecology of an urban river (Drury, Rosi-Marshall, and Kelly 2013; Lu and Lu 2014; Wakelin, Colloff, and Kookana 2008). As the investigation was 7 years long, the project was divided into 3 phases:

Phase I aimed at understanding the microbial dynamics and source apportionment of 16 sites in the CAWS during the baseline, pre-disinfection period (2013–2015). Approximately 16.8 million high-quality 16S rRNA amplicon reads were generated from 891 CAWS samples collected between 2013 - 2015. No significant differences in overall microbial community dynamics, composition and structure were observed between the 3 different years, suggesting a stable riverine ecosystem. No significant difference were observed in species richness within each sampling location over time (either monthly or seasonally). However, sediment samples had significantly greater species richness than water samples. Microbial source tracking analysis suggested that while present, the probability that bacteria in the water and sediment of the CAWS originated from human stool was minimal, however, the percentage of potential apportionment did vary between sites and over time. As expected, wildlife-associated bacteria were a variable, but significant source of bacterial 'contamination' in the CAWS.

Phase II comprised years 2016–2017, and also an increase in the different sampling types explored. Water and sediment sampling continued, but we also sampled treated effluent discharged from the O'Brien and Calumet WRPs, including the prior 3 years, so that a snap-shot of the impact of disinfection and TARP implementation could be observed. We used microbial data collected between 2013–2015 as the baseline (pre-treatment) for comparative analyses. By comparing all sample types across this transition, we observed that post-treatment years comprised significant differences in microbial composition for water, sediment, effluent, and sewage samples when compared to the pre-treatment years. The water, sediment and effluent samples demonstrated a significant decrease in microbial diversity in 2016, i.e., immediately post-disinfection; but diversity rebounded in 2017. A significant reduction of known sewage and human fecal indicators such as *Acinetobacter*, *Cloacibacterium*, *Bifidobacterium,* and Clostridiales was observed post treatment in both water and sediment samples, but only at sites immediately impacted by municipal wastewater discharge.

Phase III includes an extension of the post-disinfection analysis through years 2018 and 2019, as well as finalizing the analysis of all data across the 7 years of observation. This report further focuses on the role of rainfall on microbial diversity by investigating the impact of dry weather events, wet weather events without CSOs and wet weather events with CSOs on sites which are upstream and downstream of the two WRPs i.e., Calumet and O'Brien, as well as sites which are further downstream and the main stem of the Chicago river system. We further investigated whether the disinfection treatment and TARP modulate the microbial community structure during the wet events. The combined sewage overflow and stormwater events resulted in a significant increase in sewage microbes. However, in the post-TARP phase, we observed a limited increase in sewage indicators when compared to the pre-TARP phase as well as an increase in native river water bacterial indicators in the Calumet region. This is suggestive of less impacted/contaminated river water system after the implementation of the TCR in the Calumet WRP service area. This is also potentially a result of the significant reduction in CSO events as a result of the implementation of the TCR (Gallagher and Wasik, 2019). Molecular microbiome data were also correlated with fecal coliform data, providing validation of the molecular approaches against quantitative traditional analyses. We further investigated the abundance patterns of this fecal coliform correlated dynamic community structure across different sites of CAWS including 6 sites from Calumet River System (Calumet WRP), 4 sites from North (O'Brien WRP) River System and three sites (one each from main stem, South branch Chicago

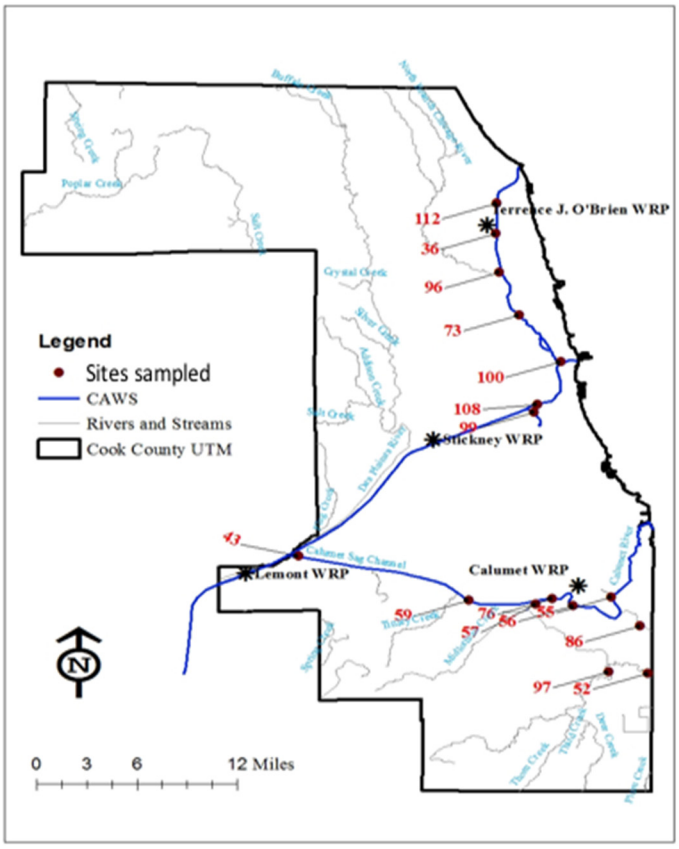
river and South Fork River) of CAWSs during the dry weather events, wet weather events with and without CSOs. Using the results, we emphasize that microbes can be used as biomarkers for anthropogenic impact in these urban waters, with the implication that these microbial biomarkers have a great potential to be used to bio-indicators for water pollution. We also performed detailed correlation analysis between the physicochemical properties of the CAWS and the microbiome to characterize the relationship between water quality indices and the community diversity. We focused on physicochemical properties i.e., pH, dissolved oxygen, nitrate, water temperature, specific conductance, flow, ammonia, total organic carbon, turbidity, chlorophyll, fecal coliforms, and volatile suspended solids. We identified significant correlations (both positive and negative) between these indices and specific microbial taxa with the intent that the findings can be utilized to establish bioindicators for water quality.

1.2 MATERIALS AND METHODS

1.2.1 Assessing Microbial Community Structure in CAWS Samples over Seven Years Using 16S rRNA Amplicon Gene Sequencing

We utilized 16S rRNA gene amplicon sequencing to characterize the microbial communities in CAWS samples during years, 2013–2019. We collected a total of 2,706 samples: 292 effluent samples, 572 sediment samples, 749 water column samples, 92 sewage samples, 472 (bottle, filter, equipment) blanks, and 219 fish samples (Table 1, 2). Sewage and effluent samples were collected from the two WRPs at O'Brien and Calumet (Figure 1, Table1). Disinfection processes were implemented in 2016 at the O'Brien (UV) and Calumet (chlorination/dechlorination) WRPs. In 2016, we also observed the phased implementation of the Tunnel and Reservoir Plan (TARP). Thereafter, we continued sampling treated effluent, water, and sediment samples both upstream and downstream of the two WRP sites (Table 1). Additional information on the number of river water and sediment samples that were processed by sampling site is included in Table 1.

All CAWS locations were sampled monthly for water and sediment by surface grab sampling. Water samples were analyzed for physicochemical parameters including pH, water temperature, alkalinity, total suspended solids, ammonia, nitrate, phosphorus, total metals, dissolved metals, cyanide, phenol, and fecal coliform bacteria, while organic priority pollutants and nonylphenols were sampled semiannually and quarterly, respectively. Both pH and water temperature were measured at each site for all samples. 500 mL of water and 100 g of sediment was collected for microbiome analysis. Sediment samples were stored in polypropylene containers at 4C to be shipped to the lab for DNA extraction and analysis. Two hundred mL of water samples were filtered in duplicate using 0.22 Micron Mixed Cellulose Ester filter, and filters were aseptically transferred to the labeled sterile 50mL tubes and stored at -80C until transferred to the lab on ice for analysis. In the lab all samples were stored at -80C until thawing for processing. Sample collection was conducted by MWRD personnel. Detailed collection and processing protocols for water samples, effluent samples, and sewage samples and processing are described in Appendix D. Fish sample collection and processing are also described in Appendix D.



Site	Address
36	North Shore Channel @ Touhy Ave.
43	Cal-Sag Channel @ Route # 83
52	**Little Calumet River @ Wentworth Ave.**
55	*Calumet River @ 130th St.*
56	*Little Calumet River @ Indiana Ave.*
57	**Little Calumet River @ Ashland Ave.**
59	Cal-Sag Channel @ Cicero Ave.
73	North Branch Chicago River @ Diversey Ave.
76	Little Calumet River @ Halsted St.
86	*Grand Calumet River @ Burnham Ave.*
96	**North Branch Chicago River @ Albany Ave.**
97	**Thorn Creek @ 170th St.**
99	South Fork, South Branch Chicago River @ Archer Ave.
100	Chicago River Main Stem @ Wells St.
108	South Branch Chicago River @ Loomis St.
112	*North Shore Channel @ Dempster Street*

FIGURE 1 CAWS and WRP (*) sampling locations. Tributaries feeding to CAWS are in bold and upstream sites are in italic fonts.

TABLE 1 Details of location for each site with reference to the two WRPs used in the microbiome analyses.

	A. CAWS North		
WRP	O'Brien WRP Disinfected Effluent		UV Disinfected Effluent
112	North Shore Channel (NSC)	Dempster St	~1.5 Miles Upstream from O'Brien WRP
36	North Shore Channel	Touhy Ave.	~0.68 Miles Downstream from O'Brien WRP
73	North Branch Chicago River	Diversey Ave.	~6.5 Miles Downstream from O'Brien WRP
	B. CAWS North Tributary		
96[b]	North Branch Chicago River	Albany Ave.	Tributary River Meets NSC~3.2 Miles from O'Brien WRP
	C. CAWS Main Stem		
100[b]	Chicago River Main Stem	Wells St.	Downtown Chicago River ~11 Miles from O'Brien WRP

TABLE 1 (Cont.)

	D. CAWS South Branch Chicago River		
108	South Branch Chicago River	Loomis St.	~14.5 Miles Downstream from O'Brien WRP
99	SF, South Branch Chicago River	Archer Ave.	South Fork River (~Bubbly Creek receives Racine Avenue Pumping Station Discharge flow)
	E. CAWS Calumet River		
WRP	Calumet WRP Disinfected Effluent		Chlorination/dechlorination Disinfected Effluent
86[b]	Grand Calumet River	Burnham Ave.	Upstream Tributary Meets Little Calumet River ~4.4 Miles from Calumet WRP
56[b]	Little Calumet River	Indiana Ave.	~1 Mile Upstream from Calumet WRP
76	Little Calumet River	Halsted St.	~1.3 Miles Downstream from Calumet WRP
57[b]	Little Calumet River	Ashland Ave.	Tributary Meets Little Calumet River ~1.7 Miles downstream from Calumet WRP
	F. CAWS Cal-Sag Channel		
59	Cal-Sag Channel	Cicero Ave.	~6.4 Miles Downstream from Calumet WRP
43[c]	Cal-Sag Channel	Route #83	~17.2 Miles Downstream from Calumet WRP

[a] Miles for a site along the river which correspond to distance from WRP to the point the tributary joins the CAWS.

[b] Sites on CAWS without influence from O'Brien and Calumet WRPs.

[c] Sites sampled in 2014–2015 to document baseline conditions in the Calumet River System in the two years preceding completion of the Calumet TARP System's Thornton Composite Reservoir. Sites also sampled in 2017–2018 to document conditions in the Calumet River System in the two years following completion of the TCR in the Calumet WRP service area.

TABLE 2 Total number of samples collected per sample type from 2013 to 2019.

Year	Effluent	Sediment	River	Wet/ Dry	Raw Sewage	Controls	Fish	Spiked	Lake Plate	Lake Bypass	Beach	Total
2013	55	78	82	0	0	15	0	0				230
2014	72	84	133	54	9	56	0					408
2015	76	99	109	54	17	103	48	9			7	522
2016	41	104	106	44	19	82	47					443
2017	17	107	107	53	16	83	34		55	2		474
2018	16	100	108	14	16	74	42	0	0	0	0	370
2019	15	0	104	16	15	59	48	0	0	2	0	259
Total	**292**	**572**	**749**	**235**	**92**	**472**	**219**	**9**	**55**	**4**	**7**	**2,706**

1.2.2 Amplicon based Microbial Community Sequencing Analysis

The DNA was extracted from the samples using the protocol described in Appendix A (Protocol#1; (Marotz et al. 2017) and 16S rRNA gene was amplified using the protocol described in Appendix A as Protocol#2 (Minich et al. 2018). Briefly, the V4 region of the 16S rRNA gene (515F-806R) was amplified with region-specific primers that included the Illumina flow cell adapter sequences and a 12-base barcode sequence. Each 25µl PCR reaction contained the following mixture: 12µl of MoBio PCR Water (Certified DNA-Free; MoBio, Carlsbad, USA), 10µl of 5-Prime HotMasterMix (1×), 1µl of forward primer (5µM concentration, 200pM final), 1µl of Golay Barcode Tagged Reverse Primer (5µM concentration, 200pM final), and 1µl of template DNA (Thompson et al. 2017). The conditions for PCR were as follows: 94°C for 3 min to denature the DNA, with 35 cycles at 94°C for 45 s, 50°C for 60 s, and 72°C for 90 s, with a final extension of 10 min at 72°C to ensure complete amplification. Amplicons were quantified using PicoGreen (Invitrogen) assays and a plate reader, followed by clean up using UltraClean® PCR Clean-Up Kit (MoBio, Carlsbad, USA) and then quantification using Qubit readings (Invitrogen, Grand Island, USA. The 16S samples were sequenced on an Illumina MiSeq platform with paired-end sequencing at the Argonne National Laboratory Core Sequencing Facility according to EMP standard protocols (Thompson et al. 2017). The average analyzed amplicon read length was 250bp.

1.2.3 16S rRNA Gene Sequence Analyses

For 16S rRNA gene analysis, the 40 million paired-end reads generated were first joined using join_paired_ends.py script followed by quality-filtering and demultiplexing using split_libraries_fastq.py script in QIIME 1.9.1 (Caporaso et al. 2010). Parameters for quality filtering included 75% consecutive high-quality base calls, a maximum of three low-quality consecutive base calls, zero ambiguous bases, and minimum Phred quality score of 3 as suggested in Bokulich et al., 2013 (Bokulich et al. 2013). Demultiplexed sequences were then selected for ASVs (Absolute Sequence Variants) picking using the DeBlur pipeline (Amir et al. 2017). In the pipeline, *de novo* chimeras were identified and removed, artifacts (i.e., PhiX) were removed, and ASVs in less than 10 samples were removed for further analyses due to low representation. Additionally, we calculated good's coverage index in qiime2 to estimate completeness of sequencing depth. Good's coverage index was between 97 to 100% with average of 99.5% for the data indicating that the number of sequence reads was sufficient to capture most taxa in each sample. 81.2% of the samples ranged between 99 to 100% good's coverage.

1.2.4 Statistical Analyses

Analysis of the resulting biom files was completed in QIIME1.9.1, R3.4.2 (phyloseq and caret packages), and SourceTracker (in QIIME1.9.1) (Caporaso et al. 2010; Knights et al. 2011). As a first step, alpha and beta diversity was estimated between different sample types as well as year-wise. Diversity, defined as the description of "the variety and abundance of species in a defined unit of study," (Magurran, 2004) is a measure often used to describe the complexity of a community. Alpha diversity is defined as species richness (number of taxa) within a single

sample. Diversity indices applied to microbiota data consist of differing weights of two components, richness and evenness (Jost, 2006). Richness is a count of the number of different taxa observed in the community without regard to their frequencies, and evenness refers to the equitability of the taxa frequencies in a community. We here used Shannon and Inverse Simpson indices here. Shannon index equally weights richness and evenness while the Simpson index provides more weight to the evenness. Beta diversity, on the other hand, is defined as diversity in the microbial community between different environmental samples. Beta-diversity was determined using weighted UniFrac distance matrices (Lozupone et al. 2011). UniFrac is a β-diversity measure that uses phylogenetic information to compare environmental samples. UniFrac, coupled with standard multivariate statistical techniques including principal coordinates analysis (PCoA), identifies factors explaining differences among microbial communities. The Weighted UniFrac incorporates the ESV abundances when calculating shared/unshared branch lengths to calculate distance (Lozupone et al. 2011). The statistical significance of the differences in microbial alpha diversity (based on Shannon index) and beta diversity were assessed for significance using paired t-test and permutational multivariate analysis of variance (PERMANOVA), respectively (Anderson Marti J. 2014). Analysis of composition of microbiome (ANCOM) followed by Mann-Whitney U test was used to identify differentially abundant bacterial phyla, genera and ASVs in different sample types across different sampling periods (2013–2019) (p-value cut-off of 0.05 following Benjamini-Hochberg FDR correction) (Mandal et al. 2015). Spearman rank correlation and generalized linear models (GLMs) were used to establish association between the microbiome and other continuous variables in the metadata such as fecal coliform data using microbiomeSeq() (https://github.com/umerijaz/microbiomeSeq) and glm() (https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm) packages in R. All the figures were generated using ggplot2() (https://github.com/tidyverse/ggplot2, lattice() (https://github.com/deepayan/lattice), reshape2() (https://github.com/hadley/reshape), phyloseq() (https://github.com/joey711/phyloseq) and microbiomeSeq() (https://github.com/umerijaz/microbiomeSeq) packages in R scripting language (https://www.r-project.org/). Further, to identify the synergistic impact of physicochemical properties of water on microbial diversity, we used *canonical correspondence analysis* (*CCA*) and *BEST* analyses in microbiomeSeq() package in R.

1.2.5 Dry and Wet Weather Monitoring

Sampling was done during each of the following conditions during the pre- and post-construction monitoring periods:

1. Dry weather (<0.1-inch precipitation). Dry weather was defined by antecedent dry conditions for two days following a 0.25–0.49-inch event, four days following a 0.50–0.99-inch event, and six days following a >1.0-inch event.

2. Wet weather without CSOs (>0.5-inch precipitation). Water sampling occurred within 12 hours of the end of the rain event.

3. Wet weather with CSOs. Water sampling occurred within 12 hours of the end of the rain event.

The impact of rainfall on microbial dynamics of CAWS focused on six sites from Calumet Region, four sites from the O'Brien region, and one site each from the main stem, South Branch Chicago River, and South Fork River. Hence, please note that the microbiome analyses in this report does not cover all the dry, wet weather events with and without CSOs. We selected the microbiome samples from each year that met the categorical requirements for dry weather events, wet weather events without CSOs and wet weather events with CSOs by matching the dates of occurrence of the events. The dry and wet times are from the actual rainfall record maintained by the District. Selection bias was not checked, however samples were collected as part of the monitoring program criteria established for dry/ wet weather events and has been published in several of the District's research report publications—District Report No. 07-79 Fecal Coliform Densities in the Chicago Waterway System During Dry and Wet Weather 2004–2006; District Report Dry and Wet Weather Risk Assessment of Human Health Impacts of Disinfection vs. no Disinfection of the Chicago Area Waterways System (CAWS); District Report No. 11-43 The Chicago Health, Environmental Exposure, and Recreation Study (CHEERS) Final Report; and Gallagher and Wasik (2019).

TABLE 3 List of samples selected for microbiome analyses of dry, wet weather events with and without the CSOs (with dates of sample collection) across the selected sites from the CAWS north region.

CAWS North: Sites 96, 112, 36 and 73			
Year	Dry Weather Events (E1)	Wet Weather Events without CSOs (E2)	Wet Weather Events with CSOs (E3)
2013	4/8/13, 5/13/13, 6/10/13, 7/15/13, 8/12/13, 10/14/13	9/16/13	3/11/13
2014	3/10/14, 8/11/14, 9/15/14, 10/13/14, 11/10/14	0	0
2015	3/9/15, 4/13/15, 7/13/15, 8/10/15, 10/12/15, 11/9/15	0	0
2016	4/11/16, 5/9/16, 6/13/16, 7/11/16, 8/8/16, 9/12/16, 11/14/16	3/14/16, 4/1/16	4/28/16, 5/11/16, 5/12/16, 5/26/16, 6/23/16
2017	3/13/17, 4/10/17, 5/8/17, 6/12/17, 7/17/17, 8/14/17, 9/11/17, 11/13/17	0	6/15/17
2018	3/12/18, 4/9/18, 8/13/18, 9/10/18, 11/13/18	6/11/18, 10/8/18	5/14/18
2019	5/13/19, 6/10/19, 7/15/19, 8/12/19, 9/9/19, 10/14/19, 11/12/19	4/12/19	5/1/19

TABLE 4 List of samples selected for microbiome analyses of dry, wet weather events with and without the CSOs (with dates of sample collection) across the selected sites from the Calumet region, Calumet: Sites 86, 56, 76, 57, 59, 43.

Year	Dry Weather Events (E1)	Wet Weather Events without CSOs (E2)	Wet Weather Events with CSOs (E3)
2013	3/25/13, 5/28/13, 6/24/13, 7/29/13, 8/26/13, 10/28/13, 11/25/13	0	0
2014	3/24/14, 4/28/14, 7/22/14, 7/28/14, 9/29/14, 10/27/14	4/15/14, 8/5/14, 11/24/14	5/21/14, 7/1/14, 8/22/14
2015	3/23/15, 5/21/15, 5/26/15, 6/22/15, 7/27/15, 8/14/15, 8/24/15, 9/28/15, 10/26/15, 11/23/15	4/10/15, 6/11/15, 7/17/15	6/16/15, 7/14/15
2016	3/28/16, 4/25/16, 5/23/16, 6/27/16, 8/22/16, 9/26/16, 11/28/16	7/25/16	0
2017	4/24/17, 7/31/17, 8/28/17, 9/25/17, 11/27/17	5/11/17, 10/11/17, 10/23/17, 10/25/17	3/1/17, 3/31/17
2018	3/26/18, 4/23/18, 5/29/18, 7/30/18, 8/27/18, 9/24/18, 10/22/18	5/15/18, 5/22/18, 11/26/18	0
2019	3/25/19, 4/22/19, 7/29/19, 8/26/19, 11/25/19	5/28/19, 9/23/19	0

TABLE 5 List of samples selected for microbiome analyses of dry, wet weather events with and without the CSOs (with dates of sample collection) across the selected sites from the main stem and south branch sites. 0-No samples collected.

Main Stem and South Branch: Sites 99, 100, 108			
Year	Dry Weather Events (E1)	Wet Weather Events without CSOs (E2)	Wet Weather Events with CSOs (E3)
2013	3/18/13, 4/15/13, 5/20/13, 6/17/13, 8/19/13, 9/23/13, 10/21/13	0	11/18/13
2014	3/17/14, 4/21/14, 5/19/14, 6/16/14, 7/21/14, 8/18/14, 9/22/14, 10/20/14, 11/17/14	0	0
2015	3/16/15, 5/18/15, 7/20/15, 8/17/15, 10/19/15, 11/16/15	0	6/15/15
2016	4/18/16, 5/16/16, 6/20/16, 9/19/16, 11/21/16	0	3/21/16, 3/25/16, 4/28/16, 5/11/16, 5/12/16, 5/26/16, 7/18/16
2017	3/20/17, 5/15/17, 6/19/17, 8/21/17, 9/18/17	4/7/17	10/16/17
2018	3/19/18, 8/20/18, 9/17/18, 10/15/18, 11/19/18	4/16/18	0
2019	3/18/19, 6/17/19, 10/21/19, 11/18/19	8/27/19	4/15/19, 5/1/19, 8/19/19

1.2.6 Assessing Microbial Community Structure and Function Across the CAWS Using Shotgun Metagenomic Sequence Data

While 16S rRNA gene sequencing based data can provide phylogenetic information about the microbial community (i.e., who is there?), shotgun metagenomics aims at surveying

the entire genomes of the microorganisms present in an environment as opposed to single gene, therefore it can capture microbial community at higher resolution. Shotgun metagenomics is also less susceptible to the biases inherent to the target gene amplification and can identify dominant gene pathways (i.e., what is it doing?) that are present in each sample. Shotgun metagenomic data was generated using the DNA extracts following the Illumina TruSeq protocol. The sequencing was performed using a 2 × 100 bp sequencing run on the Illumina HiSeq2500. Paired-end metagenomic reads for 71 samples were quality trimmed using nesoni (http://vicbioinformatics.com/nesoni.shtml) with the following parameters; minimum length = 75, quality cutoff = 30, adapter trimming = yes and ambiguous bases = 0 (https://github.com/Victorian-Bioinformatics-Consortium/nesoni).

To assess the taxonomic diversity, trimmed data were analyzed using MetaPhlAn2 to profile the composition of microbial communities (bacteria, archaea, eukaryotes and viruses) at species level (Overbeek et al. 2014). A database of ~1M unique clade-specific marker genes identified from ~17,000 reference genomes were used in MetaPhlAn2, and BowTie2 was used for reference-based alignment of the reads (Overbeek et al. 2014). Functional profiling was performed using HUMAnN2 which identifies the species profile from shotgun data and aligns reads to their pangenomes, performs translated search on unclassified reads, and quantifies gene families and pathways. HUMAnN2 was used to regroup gene families to MetaCyc reactions (Franzosa et al. 2018). All the figures were generated using ggplot2() (https://github.com/tidyverse/ggplot2), lattice() (https://github.com/deepayan/lattice), reshape2() (https://github.com/hadley/reshape), phyloseq() (https://github.com/joey711/phyloseq) and microbiomeSeq() (https://github.com/umerijaz/microbiomeSeq) packages in R scripting language (https://www.r-project.org/).

1.3 RESULTS AND DISCUSSION

1.3.1 The Unique Microbial Composition of Different Sample Types

Molecular sequencing approaches enable the survey of a large fraction of the microbial community in an ecosystem. The main driver for microbial communities is often the type of environmental medium they inhabit, but examining the within medium variation over time is essential to track dynamics. First, we measure communities using diversity, i.e., the number and proportional abundance of species in a sample, is measured as Alpha Diversity, which can be defined as species richness (number of taxa) within a single sample or the proportional distribution, or evenness, of the sample (Jost, 2006). We employed the Shannon and Inverse Simpson indices, which are positively correlated, but Shannon is weighted towards rare species, and Simpson weighted towards abundant species. Second, we characterize the difference in microbial community structure and composition between samples, known as Beta Diversity. Here we employed Bray Curtis and Weighted UniFrac indices that are weighted towards abundant species, and unweighted UniFrac that is weighted towards rare species.

As expected (Sommers et al. 2019; Yu Wang et al. 2012; Jiang et al. 2006; Feng et al. 2009; Payne et al. 2017), sediment samples, due to inherent complexity of this medium, had the greatest alpha diversity, which was significantly greater than any other medium; effluent samples

had the second greatest alpha diversity (Shannon and Inverse Simpson; $p<0.05$, Figure 2). Alpha diversity was not significantly different between river water, sewage, and beach water samples ($p>0.05$, Figure 2). Fish gut and mucus samples were the least diverse ($p<0.05$, Figure 2). No significant differences in terms of microbial diversity were observed between the effluent samples from O'Brien and Calumet WRPs ($p>0.05$, Figure 2).

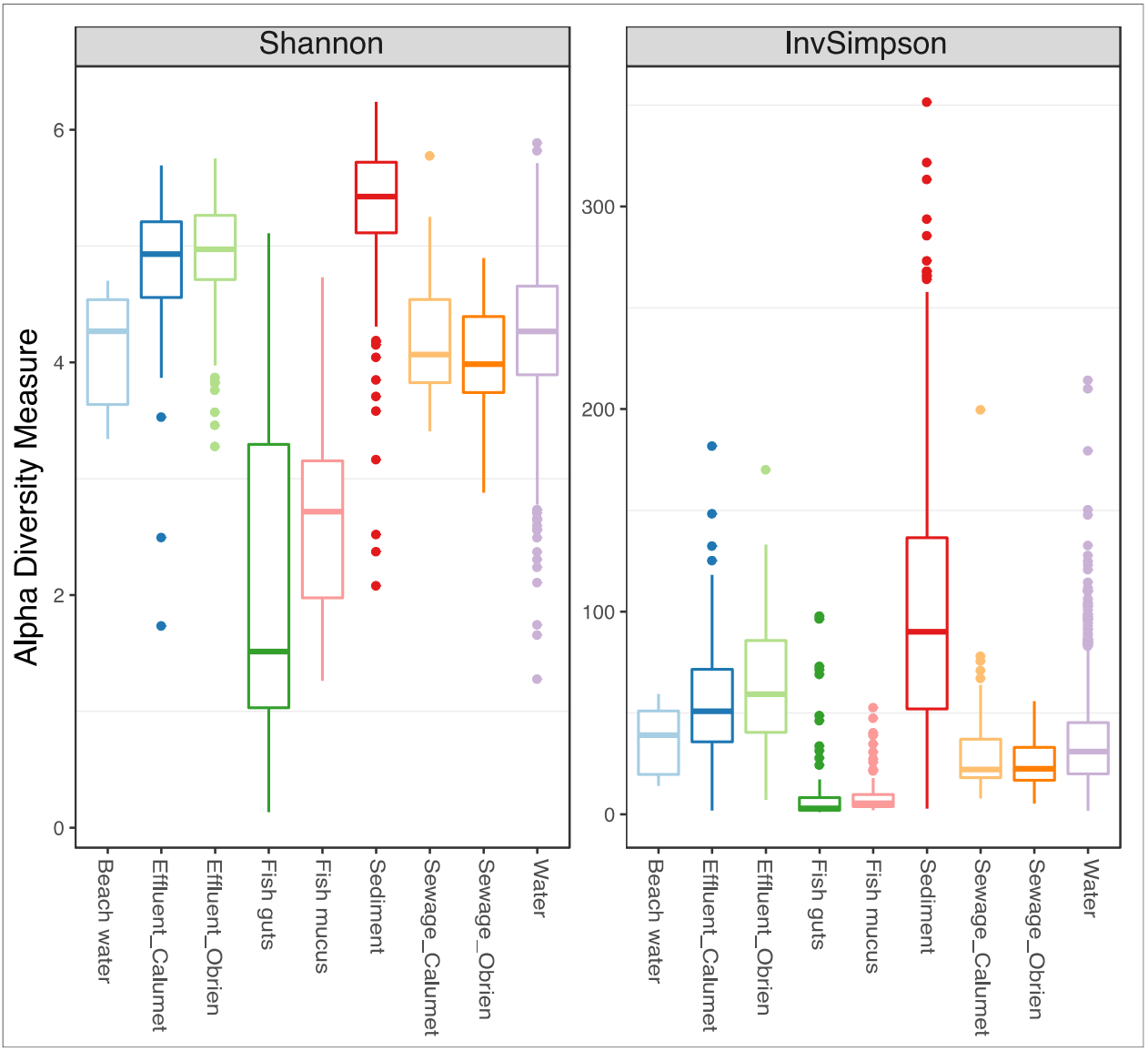


FIGURE 2 The alpha diversity analyses of CAWS samples collected from 2013–2019. The distribution of alpha diversity indices Shannon and Inverse Simpson for each sample type consolidated for all seven sampling years (2013–2019); the sediment samples are the most diverse followed by effluent and water samples, with fish-associated samples being the least diverse. This box and whisker plot demonstrates the quartile range (line) and outliers (dots) for each distribution.

Beta diversity was significantly different between sample types (weighted UniFrac; $p_{PERMANOVA} < 0.05$, Figure 3). CAWS water column samples formed a distinct cluster along with beach water samples that separated from the other sample types ($p_{PERMANOVA} < 0.05$). Sediment and effluent samples ordinated into two separate clusters. While, the alpha diversity (Figure 2), didn't show any significant difference between water, sewage and fish associated samples, beta diversity was significantly different between the three sample types, and they ordinated differently (Figure 3).

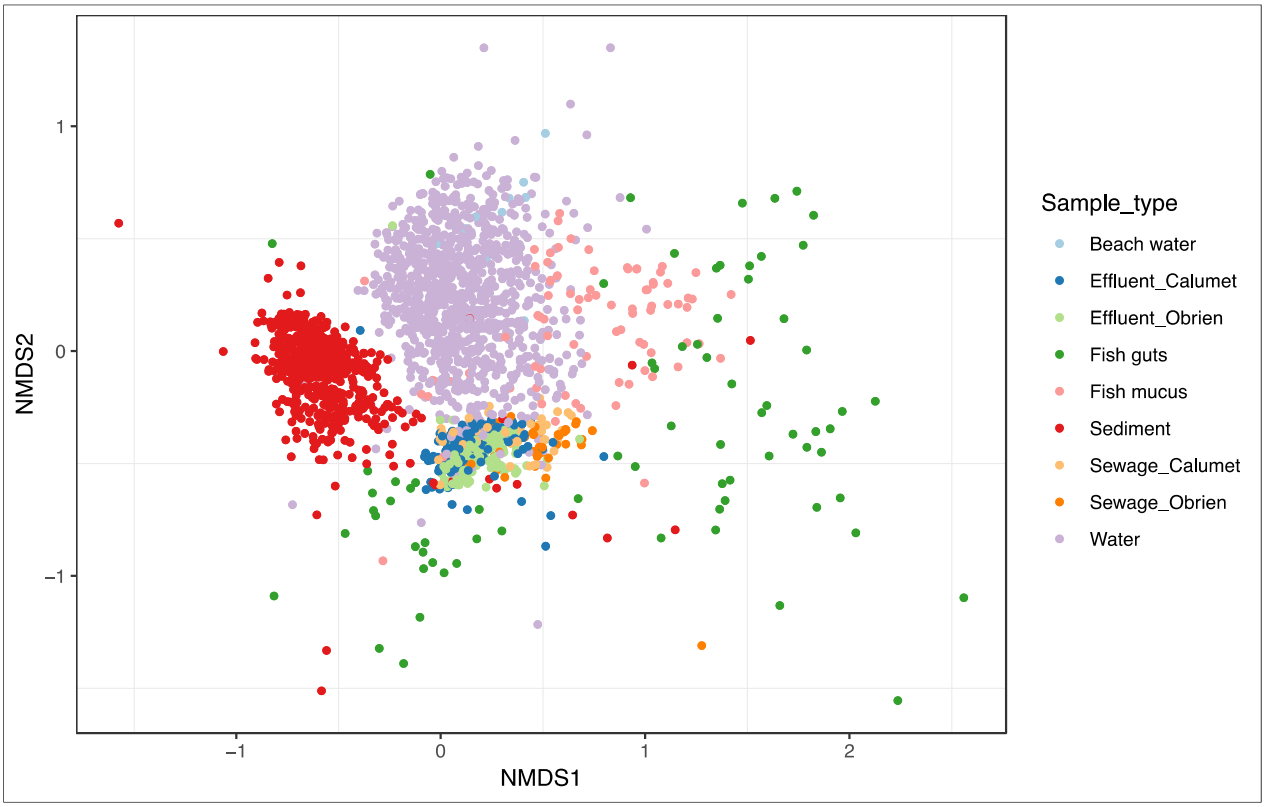


FIGURE 3 Beta diversity analyses of CAWS samples collected from 2013 to 2019. Non-metric multidimensional scaling (NMDS) plot based on the weighted UniFrac distance matrix showing clustering patterns of different sample types, i.e., beach water, effluent, fish guts and mucus, sediment, sewage and river water. The *PERMANOVA* p < 0.05 value suggest significant differences between the sample types. The water, sediment, effluent, and sewage samples form separate distinct clusters with clear and significant segregation (p < 0.05).

We identified 8 microbial genera that were significantly different between the sample types using the Analyses of Composition of Microbiome (ANCOM) algorithm (Figure 4).

Methylotenera was significantly enriched (Mann Whitney U, $p < 0.05$) in the water, sediment and effluent samples, and at greater proportion in the Calumet effluent compared to O'Brien. This genus is an obligatory methylamine-utilizing bacterium which has been reported to be enriched in the effluents from the sewage treatment plants where methanol-enhanced

denitrification is an important step (Mustakhimov et al. 2013). The enrichment of *Methylotenera* at Calumet compared to O'Brien may be due to the differential disinfection methodologies, however, we cannot validate this.

Prosthetobacter was significantly enriched (Mann Whitney U, $p < 0.05$) in the effluent and river water. This genus is often isolated from activated sludge in wastewater treatment plants where it is known to utilize the algal metabolites (J. Lee et al. 2014).

Arcobacter, Trichococcus, Delftia and Dechloromonas, all recognized sewage indicator genera, were enriched in the influent sewage samples at both Calumet and O'Brien and in the effluent; as well as river water to a lesser extent, probably reflecting their enrichment in river samples downstream of effluent sites.

Emticicia was enriched in fish mucus, and is a known symbiont of parasitic ciliates living on fish (Sun et al. 2009).

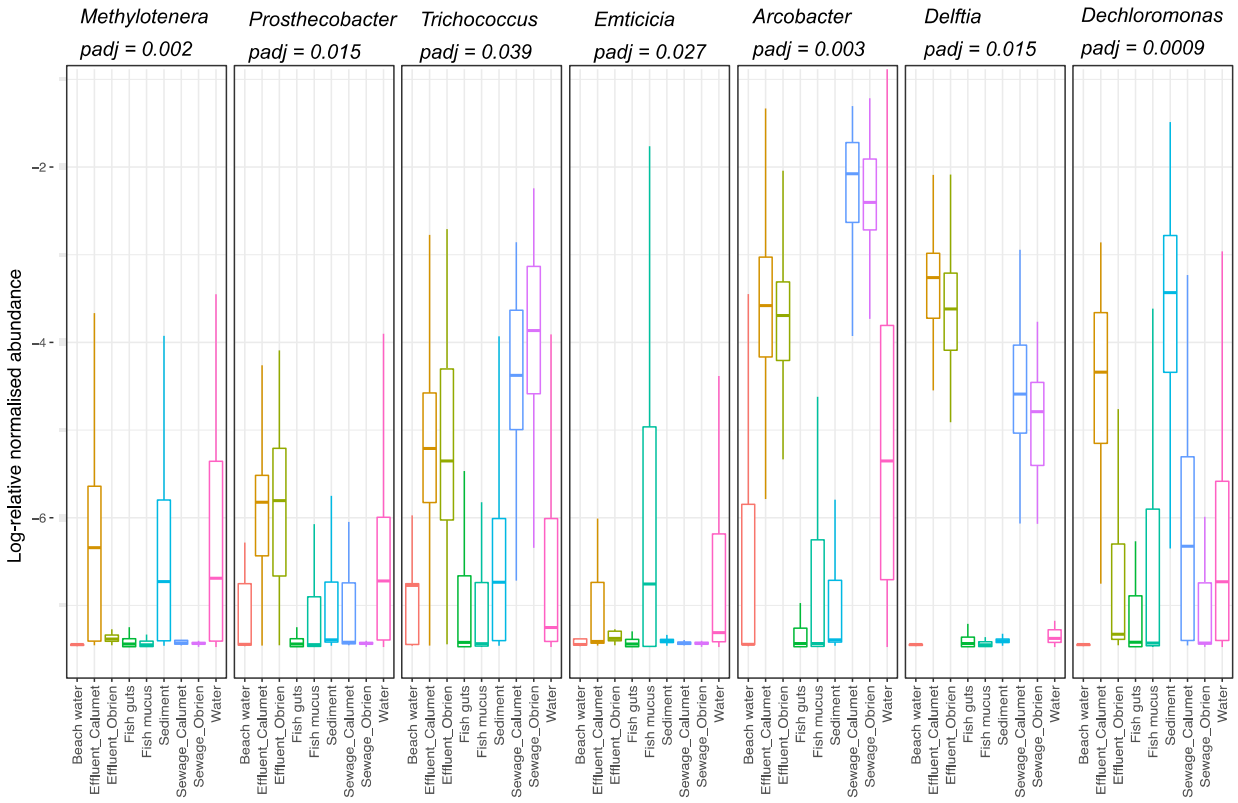


FIGURE 4 Analyses of Composition of Microbiome (ANCOM) between the different sample types. The figure shows a list of seven bacterial genera which significantly (Mann Whitney U, $p < 0.05$) differentiate the sample types with unique enrichment patterns. The statistical significance was tested in ANCOM using Benjamini-Hochberg FDR adjustments to p-values. The adjusted p-values (*padj*) are mentioned for each candidate taxa.

1.3.2 Microbial Composition of Incoming Sewage and Outgoing Effluent across the Seven Years (2013–2019)

Sewage stands for the incoming wastewater that enters the WRPs and effluent stands for the outgoing disinfected discharge flowing out the WRPs. At both the WRPs, the sewage samples were characterized by dominance of phyla Proteobacteria (avg. 70.01%), followed by Firmicutes (avg. 12.31%) and Bacteroidetes (avg. 12.10%) (Figure 5). Phyla Actinobacteria and Fusobacteria were less than 5% abundant across all the years of sampling at both Calumet and O'Brien WRPs (Figure 5). Proteobacteria are a complex and diverse phylum that dominate most natural and man-made environments, and have previously been identified as dominating sewage (Nascimento et al. 2018). At genus level, both Calumet and O'Brien WRPs were dominated by *Acinetobacter* (Cal~18.31%, O'Brien~19.97%) and *Arcobacter* (Cal~ 17.63%, O'Brien~14.79%) across all the years (2013–2019) (Figure 6). These genera are not necessarily fecal associated and are treated as non-fecal component of the sewer systems, however, the organic matter present in sewer pipes provides an ideal energy source for this opportunistic microorganisms (S.L. McLellan et al. 2010).

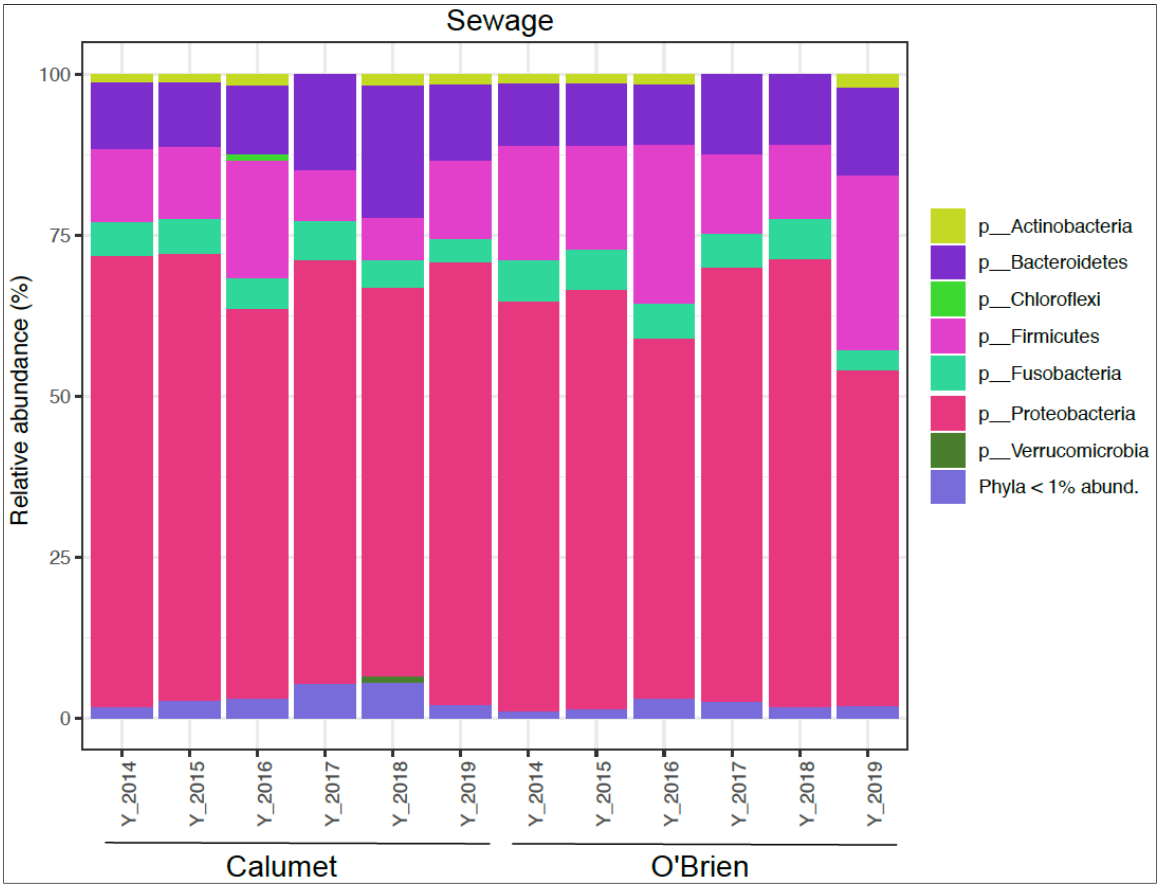


FIGURE 5 Stack plots showing compositional differences between the incoming sewage samples at Calumet and O'Brien WRPs collected from 2014–2019 at phyla level (raw sewage samples were not collected in the year 2013). The most dominant bacterial phyla are shown here with phyla less than 1% of relative abundance collapsed under one group for easier graphical visualization

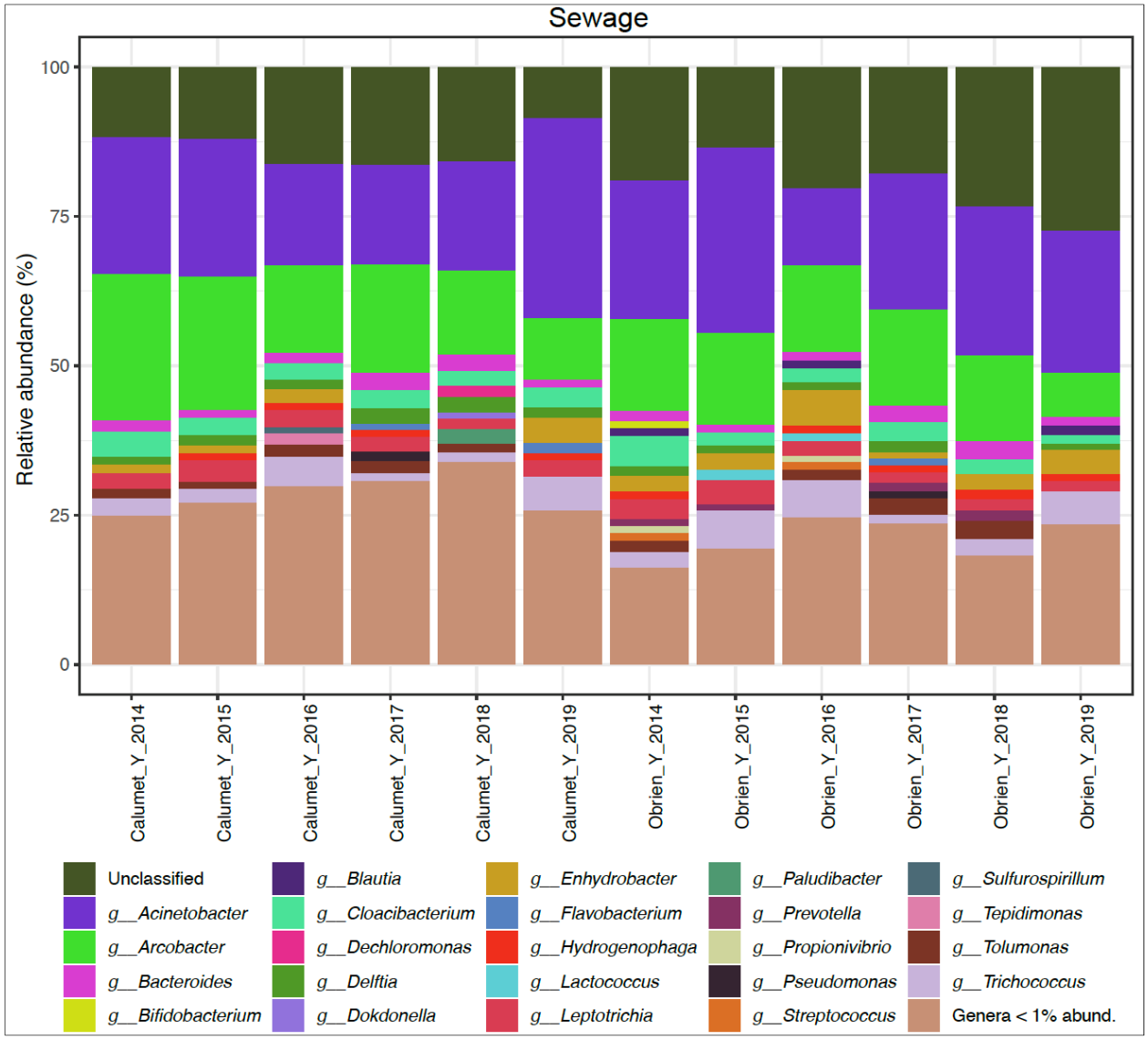


FIGURE 6 Stack plots showing compositional differences between the incoming sewage samples at Calumet and O'Brien WRPs collected from 2014–2019 at genus level (raw sewage samples were not collected in the year 2013). The most dominant bacterial genera are shown here with phyla less than 1% of relative abundance collapsed under one group.

Like the sewage samples, Proteobacteria dominated the effluent samples across all the collection years (2013–2019) (Figure 7). However, unlike the sewage samples where Bacteroidetes and Firmicutes were both almost equally abundant (~12%), Bacteroidetes was the second most dominant (~14%) followed by Firmicutes (~4%), (Figure 5, 7).

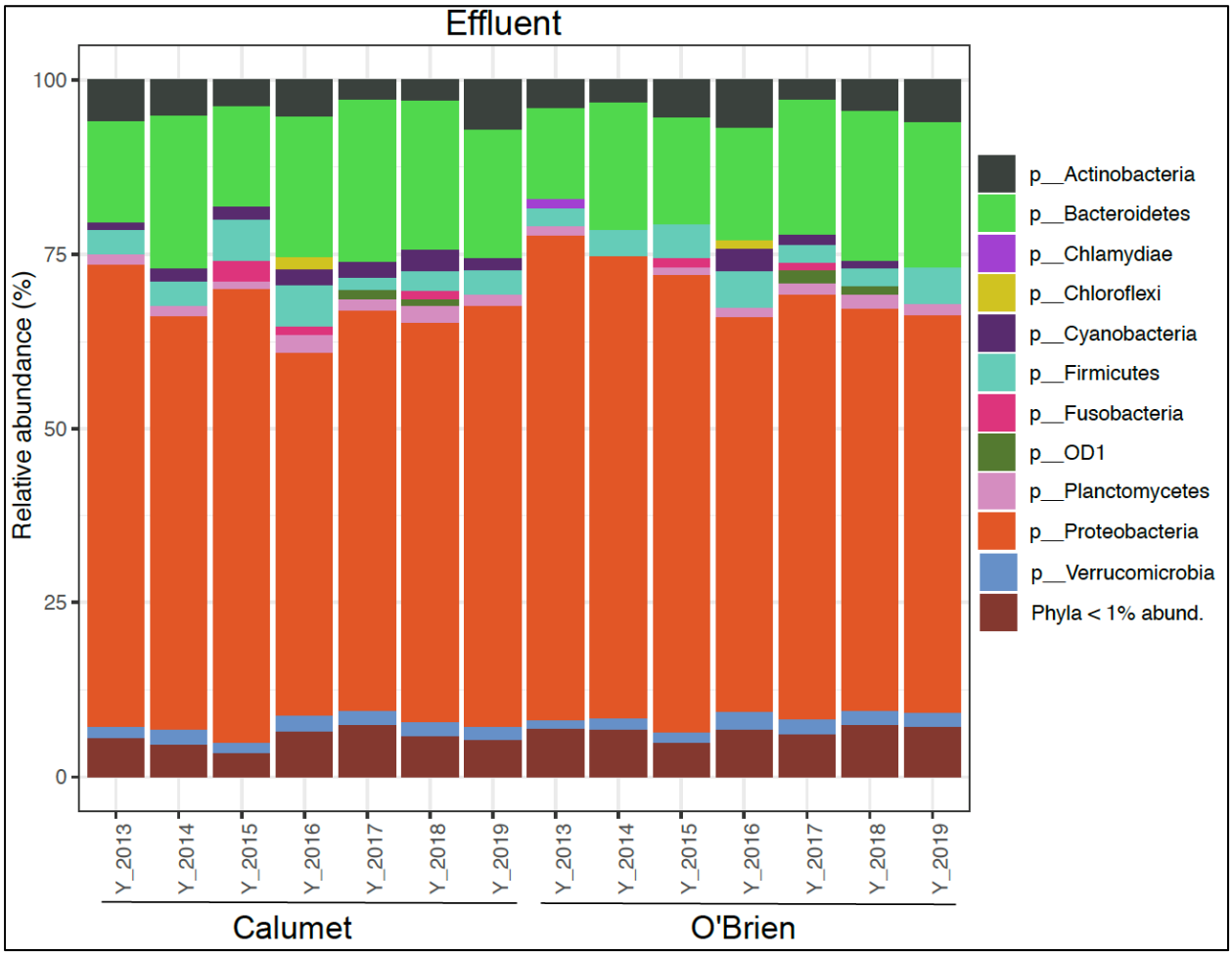


FIGURE 7 Stack plots showing compositional differences between the treated effluent samples at Calumet and O'Brien WRPs collected from 2013–2019 at phyla level. The most dominant bacterial phyla are shown here with phyla less than 1% of relative abundance collapsed under one group.

At the genus level, the effluent (like sewage) was dominated by *Acinetobacter* (Cal~7.52%, O'Brien~6.89) and *Arcobacter* (Cal~5.54%, O'Brien~3.74%), but also included *Bacteroides* (Cal~0.52%, O'Brien~0.51%) (Figure 8); all were significantly reduced (Mann Whitney U, $p < 0.05$) in the post-disinfection years (2016–2019). Both *Acinetobacter* and *Arcobacter* are well established sewage indicators and *Bacteroides* is also known to be associated with human-fecal contamination (Fisher et al. 2014; VandeWalle et al. 2012; Newton et al. 2015). These results thus highlight the efficiency of disinfection technology that was employed at both Calumet and O'Brien WRPs as well as the Calumet TARP system which reduced the number of CSOs significantly especially at the Calumet River system. Interestingly, *Hydrogenophaga* and *Sediminibacterium* were enriched in effluent at O'Brien compared to Calumet, and *Flavobacterium* significantly increased (Mann Whitney U, $p < 0.05$) in the post-disinfection years (2016–2019) in Calumet effluent, but not in O'Brein, which may reflect different disinfection strategies.

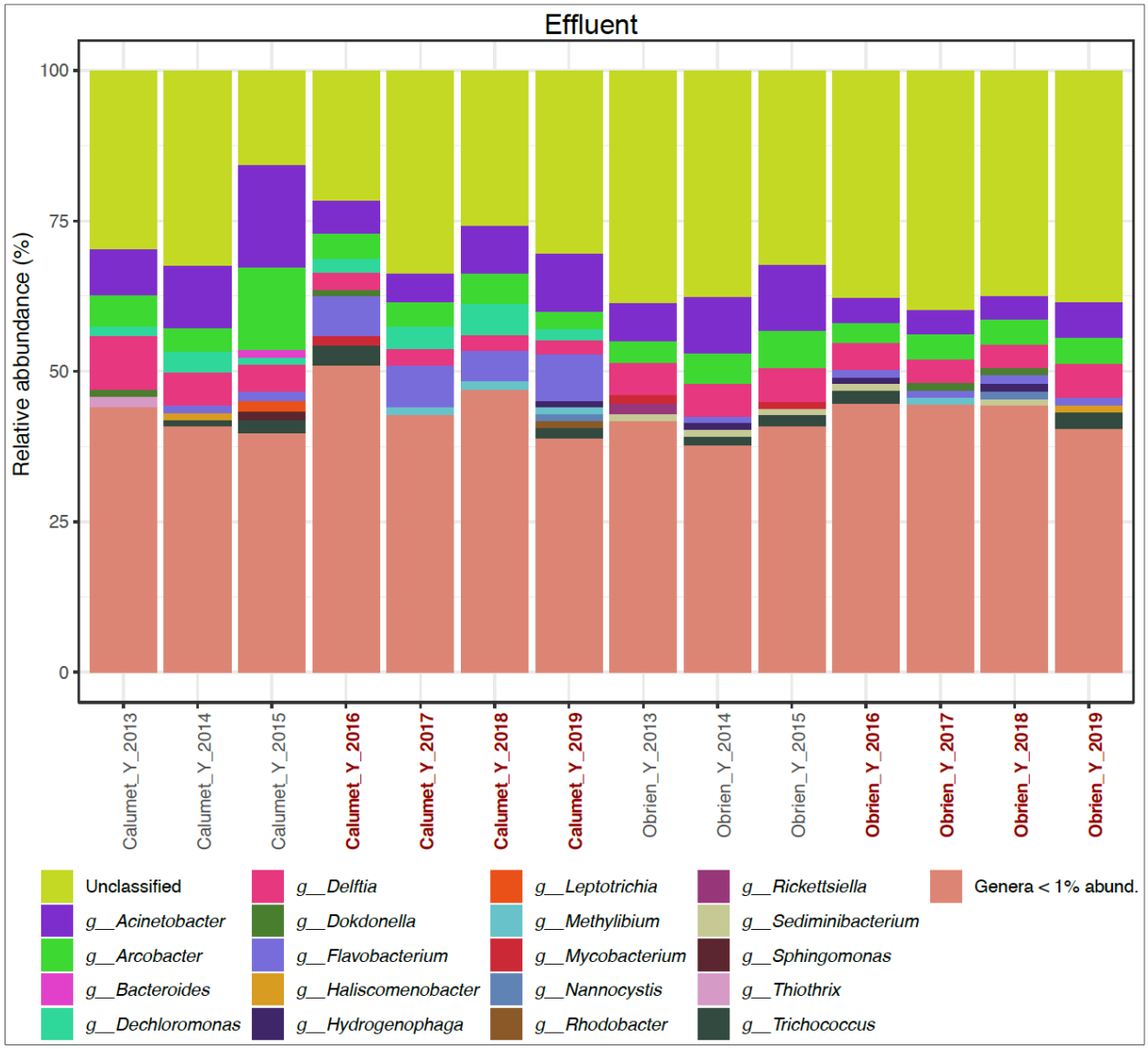


FIGURE 8 Stack plots showing compositional differences between the treated effluent samples at Calumet and O'Brien WRPs collected from 2013–2019 at genus level. The most dominant bacterial phyla are shown here with genera less than 1% of relative abundance collapsed under one group. The post-disinfection years (2016–2019) are highlighted in 'bold red'.

1.3.3 Impact of Disinfection Implemented at Calumet and O'Brien Water Reclamation Plants on Microbial Dynamics of the CAWS

Alpha diversity significantly decreased (t-test, $p<0.05$) in 2016 (the year of disinfection) for the sites downstream of the two WRPs (36 - 0.68 miles downstream of O'Brien; 73 - 6.5 miles downstream of O'Brien; and 76 - 1.3 miles downstream of Calumet; Figure 9). However, alpha diversity also decreased significantly (t-test, $p<0.05$) in two sites in the South Branch Chicago River that were not impacted by O'Brien or Calumet (100 and 108 - >14 miles away from O'Brien; Site 99, which receives CSO during stormflow events from Racine Avenue

Pumping Station, showed no significant decrease in 2016 (Figure 9). For the sites upstream of the two WRPs (i.e., 56 and 112) and tributary sites with no direct influence from the WRPs (i.e., 96, 100, 86), no significant reduction of the microbial diversity was observed in 2016 post-disinfection (Figure 9). Interestingly, sites 57 and 59 are downstream from the Calumet plant but showed no significant decrease in alpha diversity.

Sites 36, 73 and 76 showed a significant increase in Alpha diversity in 2017 and remained stable through 2019. This observation remained consistent for the sediment as well with an exception of sites 100 and 108 which didn't demonstrate any reduction of microbial diversity in 2016 (Figure 10). This increase in 2017 can be attributed to the capability of the microbial community to adapt to the changing environment with continued high growth rates.

Between the sewage and effluent samples collected from Calumet and O'Brien WRPs overall, effluent samples were characterized by higher (t-test, $p<0.05$) microbial diversity (Figure 11). This can be attributed to the fact at the WRPs, the sewage is turned into activated sludge (before the disinfection treatment) which is basically an enrichment culture of microorganisms that can remove the biological oxygen demand and total suspended solids from the WRPs (Shchegolkova et al. 2016). This sludge undergoes processes such as aeration that leads to enrichment of sewage treatment-associated microorganisms, following which, the final disinfection process (UV or chlorination/dichlorination) could alter the microbial diversity of effluent samples (compared to the incoming sewage). Based on Shannon index, Calumet WRP effluent samples demonstrated a significant decrease (t-test, $p<0.05$) in microbial diversity in 2016 as compared to 2015, followed by a significant increase in 2017 (t-test, $p<0.05$), Figure 11). Similarly, the effluent samples from O'Brien also demonstrated a significant reduction ($p < 0.05$, Figure 11) in microbial diversity post-disinfection, followed by a rebound. These results emphasize the efficiency of the disinfection process that was implemented in 2016 at both the WRPs.

The compositional data (combined for post disinfection years i.e., 2016–2019) for water samples collected from immediate downstream sites of both WRPs (Calumet, Site#76; O'Brien, Site#36), showed significant differences compared to the pre-disinfection period (i.e., 2013–2015) with sewage and fecal indicators reducing, highlighting the efficiency of disinfection. At both Calumet and O'Brien WRPs, we observed an increase (Mann Whitney U, $p < 0.05$) in fresh-water indicators such as *Flavobacterium* and a reduction in sewage indicators such as *Arcobacter* and *Acinetobacter* (Figure 12).

At Calumet we also observed an increase (Mann Whitney U, $p < 0.05$) in ASVs belonging to Comamonadaceae and Pelagibacteraceae. Family Comamonadaceae is another river water indicator known to be found ubiquitously in river water habitats (Moon et al. 2018). The family Comamonadaceae is one of the dominant bacterial groups in river water environments. The family Pelagibacteraceae has also been reported to be abundant in river water systems with many studies focused on its strong seasonal variability in the river water systems (Heinrich, Eiler, and Bertilsson 2013). This can be attributed to both disinfection technology that was implemented as well as the TARP TCR in the Calumet WRP service area.

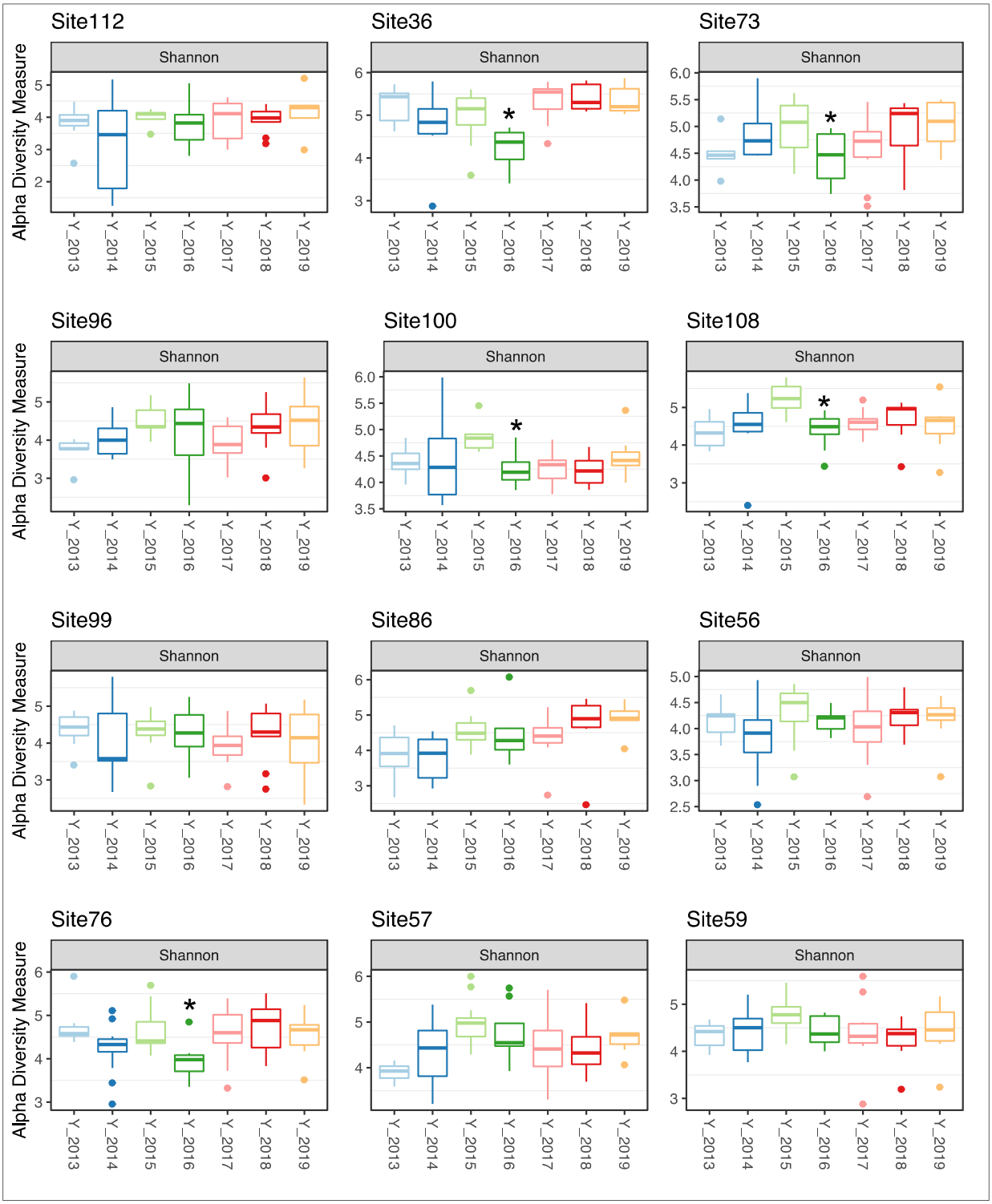


FIGURE 9 Alpha diversity analyses for water samples collected at different sites over a period of seven years (2013–2019). The distribution of Shannon diversity index is shown for the river water sites. The boxplots are grouped by sampling year. The sites with asterisks (*) in 2016 represent the comparisons between 2015 and 2016 that show statistical differences (t-test, $p<0.05$). Between the years post-disinfection, we didn't observe any significant differences between 2017, 2018 and 2019.

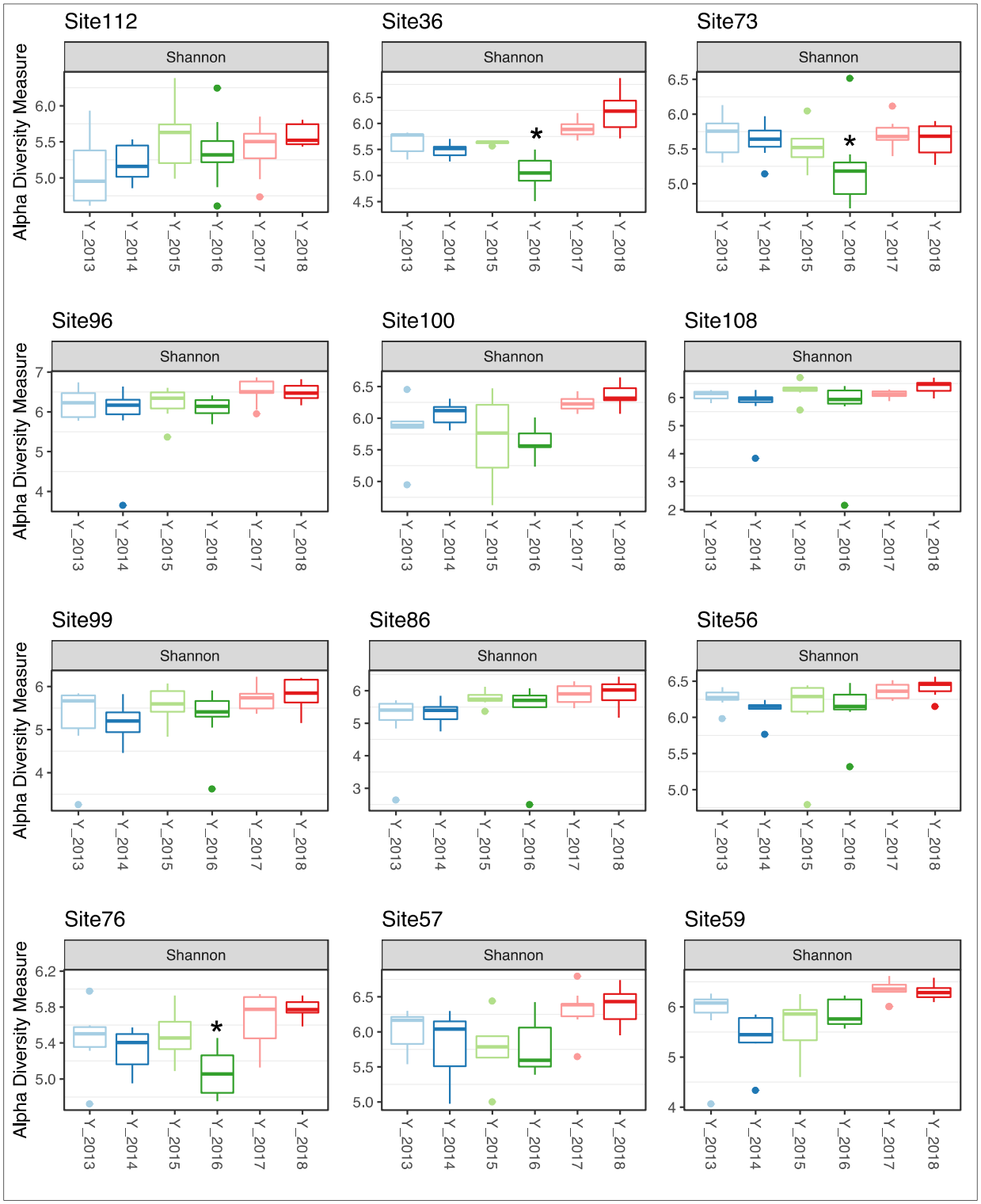


FIGURE 10 Alpha diversity analyses for sediment samples collected at different sites over a period of six years (2013–2018). Please note that sediment samples were not collected for the year 2019. The distribution of Shannon diversity index is shown for the sediment sites. The boxplots are grouped by sampling year. Between the years post-disinfection, we didn't observe any significant differences between 2017, and 2018.

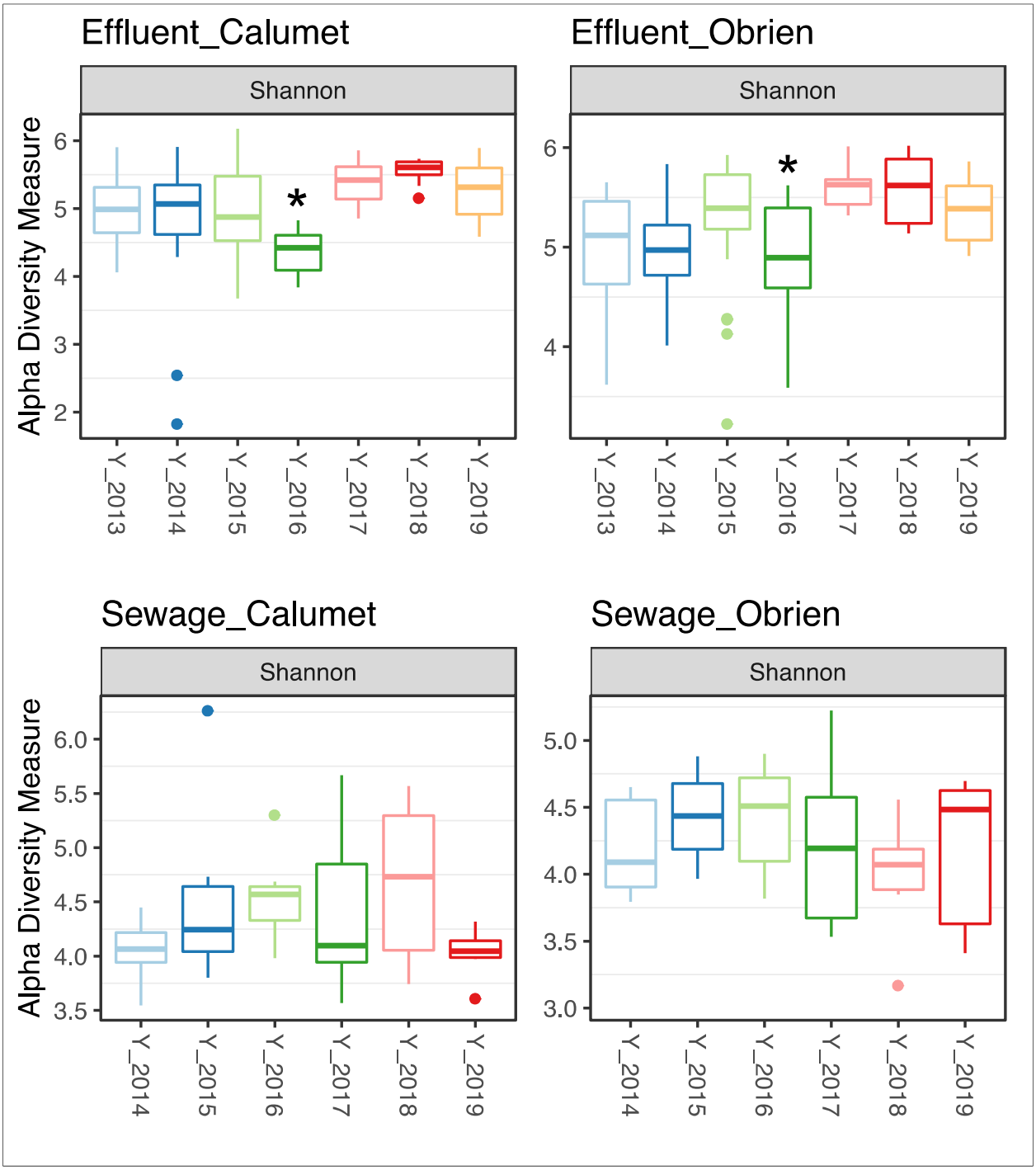


FIGURE 11 Alpha diversity analyses for sewage and effluent samples at Calumet and O'Brien WRPs. The distribution of Shannon diversity index is in the form of boxplots. The boxplots are grouped by sampling year.

At O'Brien we observed an increase in *Leadbetterella* post-disinfection which is another river water-associated bacteria that has been reported from Lake Michigan (Chopyk et al. 2018; P. O. Lee et al. 2015) (Figure 12). Therefore, we identified additional river water indicators which increased significantly post-disinfection (Mann Whitney U, $p < 0.05$) in the downstream waterways of WRPs with a further reduction in fecal indicators.

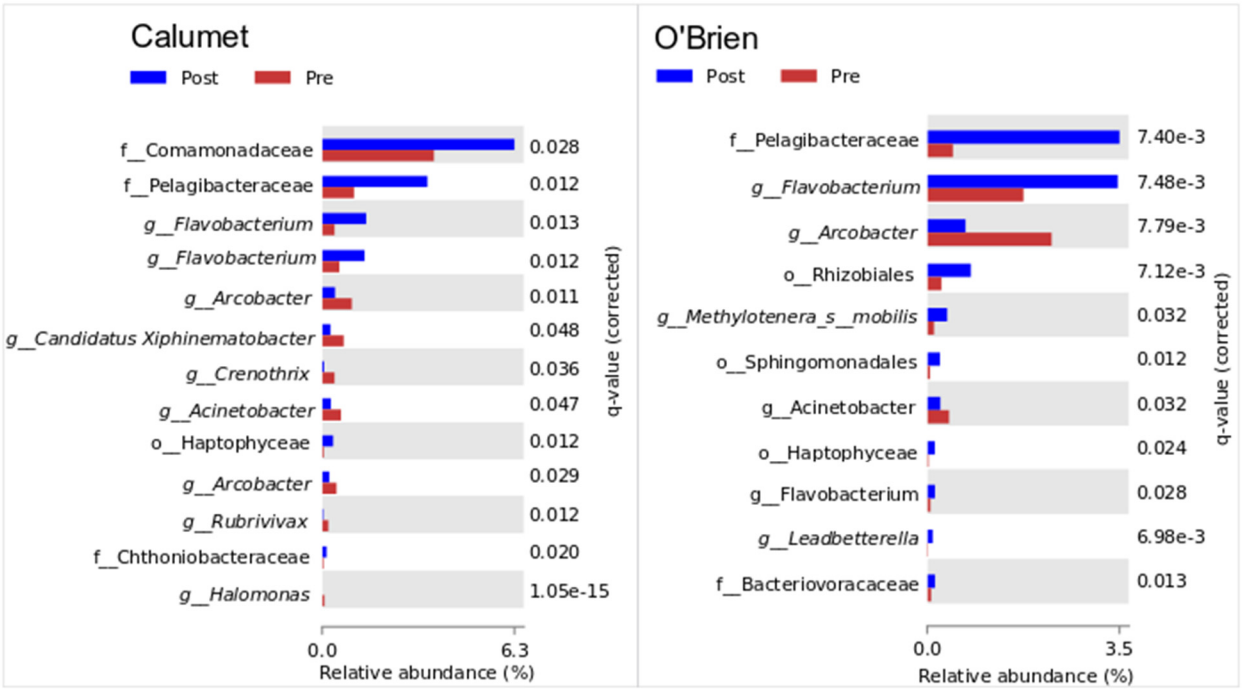


FIGURE 12 The sewage and wastewater microbes decrease significantly post-disinfection compared to pre-disinfection at the downstream sites for the O'Brien (Site#36) and Calumet (Site#76) WRPs. The river water microbes increase significantly post-disinfection. A few of the indicators for sewage like *Acinetobacter,* and human fecal material like *Arcobacter*, further reduce significantly. This figure shows list of statistically differential bacterial genera with Benjamini-Hochberg FDR corrected p-values (< 0.05) labelled for each taxon. Note that Figure 12 represents ESVs (sub-species level assignments) but many of these can only be annotated (due to the short length of the 16S rRNA fragment) to a genus or family, hence different ESVs can have the same annotation.

At sites immediate downstream of the plants, i.e., Site#36 (North Shore Channel) at O'Brien and Site#76 (Little Calumet River) at Calumet, we observed that river water indicators (e.g., Commamonadaceae, Pelagibacteraceae, *Flavobacterium*, and *Polynucleobacter)* continue to increase proportionally during the post-disinfection phase and sewage/fecal indicators (*Arcobacter, Acinetobacter, Cloacibacterium, Bacteroides, Trichococcus* etc.) continue to decrease (2016–2019; Figure 13). In addition to the fresh-water indicators, we also observed an increase (Mann Whitney U, $p < 0.05$) in genera such as *Dechloromonas* and *Hydrogenophaga* at Site# between 2015 and 2016 (Figure 13). *Hydrogenophaga* has been identified as a heterotrophic bacterium which can utilize carbon under aerobic conditions and under anaerobic

conditions through a denitrification metabolism and has been previously isolated from the activated sludge at wastewater treatment plants (Hoshino et al. 2005). The genus *Dechloromonas* is capable of reducing perchlorate and chlorate, which is associated with nitrate reductase. Moreover, *Dechloromonas* is frequently reported as a phosphate accumulating organism in enhanced biological phosphorus removal reactors (Y. Liu, Zhang, and Fang 2005). *Dechloromonas* may be associated with the chlorination/dechlorination sterilization approach employed at Calumet. Between 2017 and 2018 the river water indicators continued to increase in proportion, and while there were some fluctuations between 2018 and 2019, the trend continued (Figure 13). While chlorine and UV disinfection appeared to have similar impacts on microbial communities, there were in fact nuanced fluctuations in local river water associated bacteria, as a result of biogeographic differences between the North and Calumet regions. That both disinfection procedures impacted the proportion of fecal-coliform associated taxa should be an indication of their equal efficacy. More specific variation in effluent microbial signatures could not be detected adequately due to the lack of statistical power for these samples.

The MWRD fecal coliform (FC) plate count results validated the observation of a reduction in sewage and fecal indicator organisms. Applicable primary contact recreation designated CAWS waters are subject to general use water quality standards (Illinois Administrative Code Title 35, section 302.209 for fecal coliforms) however the recreational season is considered March through November according to Effluent Disinfection standards outlines in Illinois Administrative Code Title 35, Section 304.224) which specify that 1) the geometric mean of five samples should not exceed 200 CFU/100 ml within 30 days and 2) no more than 10% of samples should exceed 400 CFU/100 ml during any 30-day period. These standards do not apply to incidental contact or non-contact recreational water. Because samples were collected monthly for assessment purposes, overall annual geomean of 200 CFU/ 100 mL and the limit of 400 CFU/100 mL was used to evaluate the compliance.

The geometric means of the fecal coliform concentrations were significantly lower in the post-disinfection/TARP implementation period (2016–2019) compared to the pre-disinfection/TARP implementation period (2013 to 2015) in effluent water collected at the O'Brien WRP ($p<0.0001$) and Calumet WRP ($p<0.0001$) (Tables 6 and Table 7). Similarly, the number of FC exceedances of 400 CFU/100 mL occurred in almost all samples collected between 2013 and 2015, but exceedances were only found in a small fraction of samples collected in 2016–2019 (Table 6).

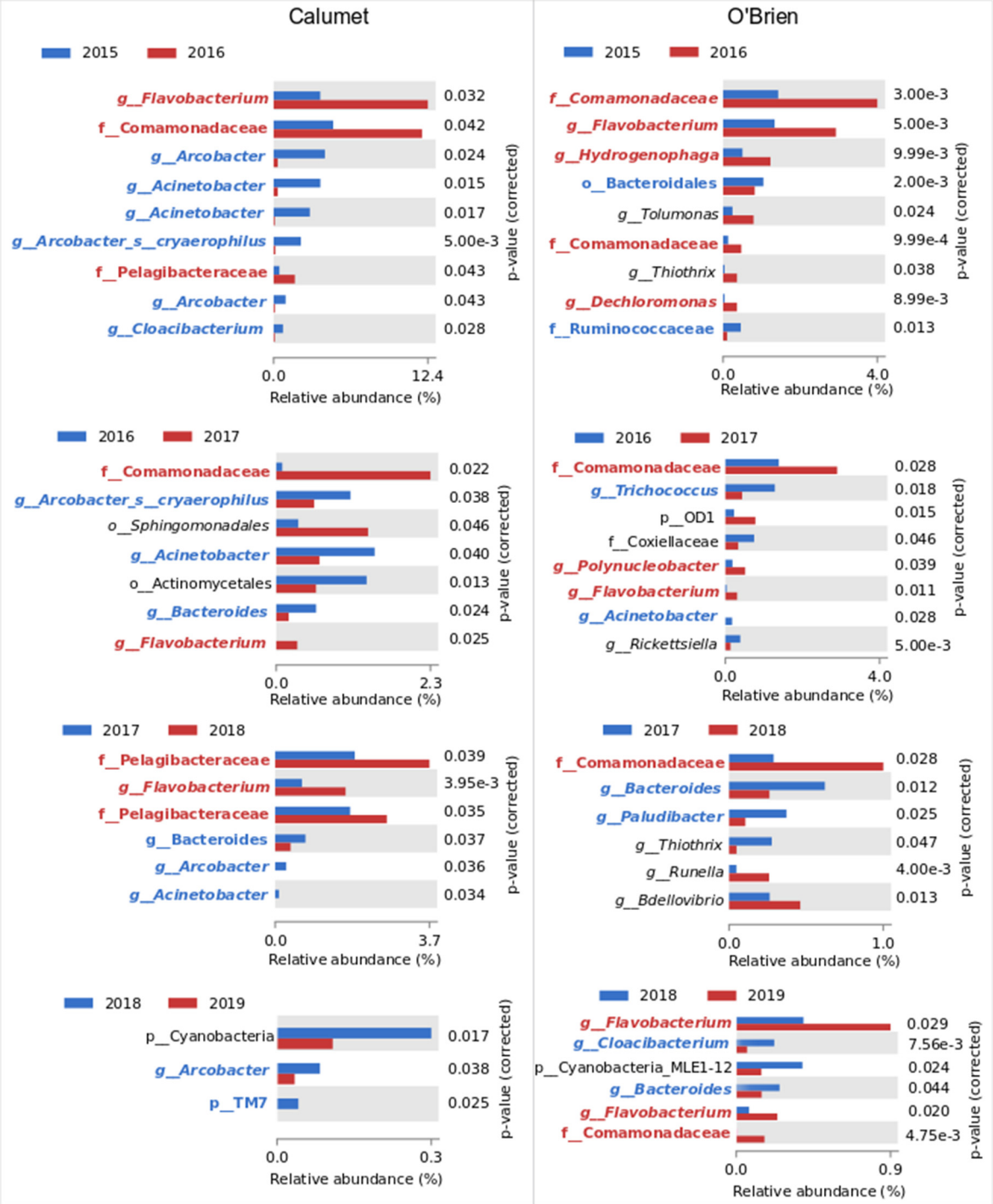


FIGURE 13 Ongoing decrease in sewage indicators and increase in fresh-water indicators downstream of the Calumet and O'Brien WRPs between the four years of disinfection (2016, 2017, 2018, 2019). This figure shows list of statistically differential bacterial Absolute Sequence Variants (ASVs) between the three years of the post-disinfection period with Benjamini-Hochberg FDR corrected p-values (< 0.05) labelled for each taxon. Taxa that are well-known sewage/fecal indicators and fresh-water indicators as per literature are highlighted in 'bold blue' and 'bold red', respectively.

TABLE 6 Summary of fecal coliform[1] concentrations in final disinfected effluents at the O'Brien and Calumet WRPs during pre (2013–2015) and post (2016–2019) disinfection/TARP implementation period.

WRP	2013	2014	2015	2016	2017	2018	2019
O'Brien							
GM[a]	13,520	14,494	9,110	50	70	71	64
#>400[b]	39/39	39/39	38/38	3/168	2/196	4/195	1/190
Calumet							
GM	8,749	7,520	2,672	16	19	15	22
#>400	39/39	39/39	42/64	0/168	3/197	3/197	0/196

[a] Fecal Coliform CFU/100 mL Geometric Mean

[b] Number of samples above 400 FC CFU/100 mL/per total number of samples collected.

TABLE 7 Statistical comparison of fecal coliform (LOG(FC)) concentrations during pre (2013–2015) and post (2016–2019) disinfection/TARP implementation period.

WRP		Period	n[a]	GM[b]	STD[c]	*p*[d]
O'Brien	Pre	2013–2015	116	12,161	0.86	0.000
	Post	2016–2019	749	64	0.93	
Calumet	Pre	2013–2015	142	4,917	2.28	0.000
	Post	2016–2019	759	18	1.02	

[a] n = Number of samples under all weather conditions;

[b] GM = Geometric Mean of the Sample data were calculated from UMV (Uniformly Minimum Variance Unbiased Estimator);

[c] STD = Standard Deviation of the Sample data were calculated from UMV;

[d] *p* = Significance Probability of natural logarithm transformed means at being equal (significant if $p \leq 0.05$).

To examine spatial and temporal changes in fecal coliform concentrations within the CAWS we compared fecal coliform concentrations in CAWS surface water upstream and downstream of the WRPs before and after disinfection/TARP implementation. We examined fecal coliform concentrations under three conditions:

1. E1 – sample collected under that did not meet categorical requirements, including dry weather (<0.1-inch precipitation) as defined by antecedent dry conditions for two days following a 0.25–0.49-inch event, four days following a 0.50–0.99-inch event, and six days following a >1.0- inch event.

2. E2 – wet weather without Combine Sewer Overflows (> 0.5-inch precipitation), and

3. E3 – wet weather with Combine Sewer Overflows, to the extent such events occur, includes the Pumping Station, if discharging.

In the North Shore Channel and North Branch of the Chicago River sites downstream of the O'Brien WRP (sites 36 and 73), the geometric mean fecal coliform concentrations were significantly lower in the 2016–2018 period (Table 8). When examined by the three weather conditions, comparisons of the 2013–2015 and 2016–2019 periods could only be made under E1 conditions because only one sample was taken during E2 and E3 conditions from 2013 to 2016. For E1 conditions, the GM FC concentration and the number of exceedances of 400 CFU were lower at these sites in 2016–2019 compared to 2013–2015 (Table 9).

TABLE 8 Statistical comparison of fecal coliform (LOG(FC) concentrations during the pre (2013–2015) and post (2016–2019) disinfection and tarp implementation monitoring at the CAWS north locations under all weather conditions.

Sample Location	Site		Period	n[a]	GM[b]	STD[c]	*p*[d]
North Shore Channel	Dempster St.	112	2013–2015 2016–2019	25 42	274 160	2.37 2.77	0.423
North Shore Channel	Touhy Ave.	36	2013–2015 2016–2019	27 43	11,694 269	0.79 1.76	0.000
North Branch Chicago River	Albany Ave.	96	2013–2015 2016–2019	26 43	779 827	1.33 2.06	0.894
North Branch Chicago River	Diversey Ave.	73	2013–2015 2016–2019	27 43	5,588 806	1.28 2.00	0.000

[a] n = Number of samples under all weather conditions;

[b] GM = Geometric Mean of the Sample data were calculated from UMV (Uniformly Minimum Variance Unbiased Estimator);

[c] STD = Standard Deviation of the Sample data were calculated from UMV;

[d] *p* = Significance Probability of natural logarithm transformed means at being equal (significant if $p \leq 0.05$).

TABLE 9 Summary of North Shore Channel and North Branch of the Chicago River system fecal coliform concentration during pre- and post-disinfection and tarp implementation.

		Pre-Disinfection/TARP 2013–2015			Post-Disinfection/TARP Phases 2016–2019		
Site		E1[a]	E2[b]	E3[c]	E1	E2	E3
WW 112	#>400[d]	6/23	1/1	1/1	1/29	1/5	7/8
1.3 miles Upstream	GM/C*	188	3,300	130,000	57	45	15,019
WW 36	#>400	25/25	1/1	1/1	3/30	0/5	5/8
0.7 miles Downstream	GM/C*	11,099	21,000	24,000	155	181	2,721
WW 73	#>400	25/25	1/1	1/1	15/30	3/5	8/8
6.7 miles Downstream	GM/C*	4,307	81,000	260,000	338	515	27,837
WW 96	#>400	15/24	1/1	1/1	12/30	5/5	8/8
Tributary River	GM/C*	609	10,000	22,000	311	1,466	22,644

[a] E1 – Monthly sampling event that did not meet categorical requirements, includes also the Dry weather (<0.1-inch precipitation). Dry weather was defined by antecedent dry conditions for two days following a 0.25–0.49-inch event;

[b] E2 – Wet weather without Combine Sewer Overflows (> 0.5-inch precipitation);

[c] E3 – Wet weather with Combine Sewer Overflows, to the extent such events occur, includes the Pumping Station, if discharging;

[d] Number of samples above 400 FC CFU/100 mL per event(s);

* Geometric Mean, For one event sample, actual count reported.

Geometric mean fecal coliform concentrations in the mainstem Chicago River and South Branch of the Chicago River (sites 100 and 108) were not significantly different between the pre and post disinfection/TARP implementation periods (Table 10). This finding may be due to the distance of these sites from the WRPs, strongly suggesting fecal coliform levels at these sites are influenced by sources other than WRP discharge. When examined by weather conditions, geometric mean fecal coliform concentrations and 400 CFU exceedances were similar in the pre and post-disinfection period for E1 conditions. A comparison could not be made for E2 and E3 conditions because too few samples were collected between 2013 and 2015 (Table 11).

TABLE 10 Statistical comparison of fecal coliform (LOG(FC) concentrations during the pre (2013–2015) and post (2016–2019) disinfection and tarp implementation monitoring at the mainstem and South Branch Chicago River under all weather conditions.

Sample Location		Site	Period	n[a]	GM[b]	STD[c]	*p*[d]
South Branch Chicago River	Archer Ave.	99	2013–2015	27	769	3.78	0.903
			2016–2019	41	866	3.98	
Chicago River Main Stem	Wells St.	100	2013–2015	27	933	1.95	0.251
			2016–2019	42	1624	1.94	
South Branch Chicago River	Loomis St.	108	2013–2015	27	434	1.43	0.330
			2016–2019	39	744	2.59	

[a] n = Number of samples under all weather conditions;

[b] GM = Geometric Mean of the Sample data were calculated from UMV (Uniformly Minimum Variance Unbiased Estimator);

[c] STD = Standard Deviation of the Sample data were calculated from UMV;

[d] *p* = Significance Probability of natural logarithm transformed means at being equal (significant if $p \leq 0.05$).

TABLE 11 Summary of Mainstem and South Branch of the Chicago River fecal coliform concentrations in samples collected during each of the categorical weather events for the pre- and post-disinfection monitoring period.

		Pre-Disinfection/TARP 2013–2015			Post-Disinfection/TARP Phases 2016–2019		
Site		E1[a]	E2[b]	E3[c]	E1	E2	E3
WW 100 Mainstem	#>400[d]	16/25	–[**]	2/2	17/28	3/3	10/11
	GM[e]	683	–	45,826	842	2,038	8,135
WW 108 SBCR	#>400	14/25	–	2/2	10/25	1/3	9/11
	GM	333	–	11,912	309	1,696	4,384
WW 99 SBCR	#>400	9/25	–	2/2	8/27	1/3	8/11
	GM	447	–	670,820	280	670	14,835

[a] E1 – Monthly sampling event that did not meet categorical requirements, includes also the Dry weather (<0.1-inch precipitation). Dry weather was defined by antecedent dry conditions for two days following a 0.25–0.49-inch event;

Footnotes continued on next page.

TABLE 11 (Cont.)

[b] E2 – Wet weather without Combine Sewer Overflows (> 0.5-inch precipitation);

[c] E3 – Wet weather with Combine Sewer Overflows, to the extent such events occur, includes the Pumping Station, if discharging;

[d] Number of samples above 400 FC CFU/100 mL per event(s);

[e] Geometric Mean;

[*] Note no wet without CSO and two wet weather events (with CSO) was sampled during 2013–2015 (pre-disinfection/TARP) compared to 3 wet with no CSO and 11 wet with CSO condition in post disinfection period (2016–2019);

[**] No sample collected.

The GM FC level (including samples collected during all weather conditions) was significantly lower in the 2016–2019 period at sites 76 (Little Calumet River) and 59 (Cal-Sag Channel), which are mainstem sites below the Calumet WRP (Table 12). The number of FC exceedances was also much lower in 2016–2019 compared to 2013–2015. The GM FC concentration at Site 43, located, 17 miles from the Calumet WRP, was not significantly different between the two periods. GM FC concentrations in tributary sample sites downstream of the Calumet WRP were not significantly different between in 2013–2015 compared to 2016–2019 (Table 13).

TABLE 12 Statistical comparison of fecal coliform (LOG(FC)) concentration during pre (2013–2015) and post (2016–2019) disinfection and tarp implementation at the Calumet sites under all weather conditions.

Sample Location		Site	Period	n[a]	GM[b]	STD[c]	*p*[d]
Grand Calumet River	Burnham Ave.	86	2013–2015 2016–2019	40 42	1,961 433	2.93 2.09	0.008
Calumet River	130th St.	55	2013–2015 2016–2019	31 23	30 15	1.46 0.62	0.042
Little Calumet River	Indiana Ave.	56	2013–2015 2016–2019	38 42	65 23	2.08 1.13	0.006
Little Calumet River	Halsted St.	76	2013–2015 2016–2019	40 41	3,330 61	1.92 0.94	0.000
Little Calumet River	Ashland Ave.	57	2013–2015 2016–2019	40 42	1,126 673	1.60 2.02	0.206
Little Calumet River	Wentworth Ave.	52	2013–2015 2016–2019	31 25	3,129 1,804	1.74 1.52	0.218
Thorn Creek	170th St.	97	2013–2015 2016–2019	31 25	1,785 1,050	1.80 1.80	0.278
Cal-Sag Channel	Cicero Ave.	59	2013–2015 2016–2019	40 42	2,027 152	2.12 2.09	0.000
Cal-Sag Channel	Routh #83	43	2013–2015 2016–2019	40 42	248 105	2.25 2.43	0.101

[a] n = Number of samples under all weather conditions;

[b] GM = Geometric Mean of the Sample data were calculated from UMV (Uniformly Minimum Variance Unbiased Estimator);

[c] STD = Standard Deviation of the Sample data were calculated from UMV;

[d] *p* = Significance Probability of natural logarithm transformed means at being equal (significant if $p \leq 0.05$).

TABLE 13 Summary of mainstem and south branch of the Calumet River system fecal coliform concentrations in samples collected during each of the categorical weather events for the pre- and post-disinfection monitoring period.

		Pre-Disinfection/TARP 2013–2015			Post-Disinfection/TARP Phases 2016–2019		
Site		E1[a]	E2[b]	E3[c]	E1	E2	E3
WW 86	#>400[d]	17/29	5/6	4/5	9/30	7/10	2/2
Upstream Tributary	GM[e]	893	6,368	45,702	205	1,562	50,160
WW 55	#>400	1/20	1/6	1/5	0/14	0/7	0/2
Upstream Tributary	GM	17	45	157	11	26	14
WW 56	#>400	3/27	0/6	3/5	0/30	0/10	0/2
1 mile Upstream	GM	48	34	719	16	72	22
WW 76	#>400	25/29	6/6	5/5	0/29	0/10	0/2
1.3 miles Downstream	GM	1,863	7,378	37,200	46	139	65
WW 57	#>400	18/29	6/6	5/5	12/30	10/10	2/2
Downstream Tributary	GM	626	2,246	14,859	266	7,053	6,142
WW 52	#>400	17/20	6/6	5/5	13/16	6/7	2/2
Downstream Tributary	GM	1,806	5,424	14,591	842	6,449	9,249
WW 97	#>400	16/20	5/6	5/5	6/16	7/7	2/2
Downstream Tributary	GM	816	8,452	6,308	350	7,489	7,099
WW 59	#>400	22/29	5/6	5/5	3/30	9/10	2/2
6.4 miles Downstream	GM	1,206	1,868	45,437	59	1,426	3,098
WW 43	#>400	7/29	4/6	4/5	1/30	9/10	2/2
17.2 miles Downstream	GM	116	518	8,162	31	2,295	1,497

[a] E1 – Monthly sampling event that did not meet categorical requirements, includes also the Dry weather (<0.1-inch precipitation). Dry weather was defined by antecedent dry conditions for two days following a 0.25–0.49-inch event;

[b] E2 – Wet weather without Combine Sewer Overflows (> 0.5-inch precipitation);

[c] E3 – Wet weather with Combine Sewer Overflows, to the extent such events occur, includes the Pumping Station, if discharging;

[d] Number of samples above 400 FC CFU/100 mL per event(s);

[e] Geometric Mean;

[*] Note 6 wet weather events (with no CSO) and 5 with CSO was sampled during 2013–2015 (pre-disinfection/TARP) compared to 10 wet with no CSO and 2 wet with CSO condition in post disinfection period (2016–2019).

1.3.4 General Assessment of Dry and Wet Weather Events with Stormwater Flow Conditions and Impact of Calumet TARP Construction on Microbial Diversity of CAWS

Chicago and 51 suburbs in the District's service area have combined sewer systems which discharged approximately 100 times a year to area waterways prior to the implementation of TARP. For an average rain event, the pollution load from these CSOs was equivalent to the organic waste loading from a population of four million people (Polls et al., 1998). During the most severe events, it was necessary to reverse the flow in the Chicago Area Waterway System (CAWS) by opening the sluice gates at the controlling works and discharging stormwater and raw sewage to Lake Michigan. Hence, in response to the problem of flooding and maintaining water quality, the TARP initiative was implemented with huge underground tunnels excavated beneath the city to intercept CSO discharges in excess of interceptor sewer capacity, and convey them to three large open surface reservoirs for storage. Following a storm, the captured CSOs were then pumped out of the tunnels and reservoirs to WRPs for treatment, and later discharged as treated effluent to area waterways. The TCR in the Calumet WRP service area became operational on November 26, 2015.

The microbiome analysis focused on 13 sites from the Calumet WRP region, O'Brien WRP region, main stem, South Branch Chicago River, and South Fork River system. Please note that the microbiome analyses in this report doesn't cover all the wet weather events with and without CSOs (see the Tables 3,4, and 5 for details of each event). We selected the microbiome samples from each year that met the categorical requirements for dry weather events, wet weather events without CSOs and wet weather events with CSOs by matching the dates of occurrence of the events as provided in the MWRD's Post TARP Construction Monitoring Report (Gallagher and Wasik, 2019). Briefly, in the Calumet region, no wet events (with and without CSOs) samples were collected in 2013, no wet weather with CSOs were recorded in 2016, 2018 and 2019. Only two wet weather with CSOs were recorded in 2017 (Table 4). This reduction in number of CSOs can be significantly attributed to the construction of the TCR in the Calumet WRP service area. In the O'Brien region, samples were collected during dry and wet weather events without CSOs in 2013; 2014 and 2015 recorded only dry weather events; 2016 recorded 5 CSO events; 2017, 2018 and 2019 recorded only one CSO event (Table 3). The main stem and south branch region recorded only one CSO event each in the years 2013, 2015, 2017, 2018, three in 2019 and seven CSO events in 2016 (Table 5). Tables 3, 4 and 5 describe the number of dry (E1), wet weather events without CSOs (E2) and wet weather events with CSOs (E3) and dates of occurrences. The impact of dry and wet events (with and without CSOs) on the CAWS microbial dynamics was analyzed at both the phylum and genera level during the pre- and post-TARP implementation phases.

In the Calumet region (covering sites 86, 56, 76, 57, 59 and 43), the overall trend at both the upstream and downstream sites (of the WRP) was an increasing proportion (Mann Whitney U, $p < 0.05$) of Bacteroidetes and Firmicutes during the wet events (both with and without CSOs). However, during wet events Bacteroidetes and Firmicutes were more abundant (Mann Whitney U, $p < 0.05$) with a CSO event compared to without (Figure 14). In the post-disinfection/TARP phase, Bacteroidetes and Firmicutes were both significantly proportionally reduced (Mann Whitney U, $p < 0.05$) in comparison to the pre-disinfection/TARP phase

(Figure 14). An increase in the proportion of Bacteroidetes and Firmicutes was associated with rainfall events in all CAWS sites investigated. Both phyla are found in the microbial community associated with stormwater and sewage water, and were both significantly reduced (Mann Whitney U, $p < 0.05$) throughout the CAWS post-treatment. As the post-TARP phase overlaps with the post-disinfection phase this reduction can be attributed to the implementation of disinfection practices at the WRPs. These results are also in good agreement with reduction of CSO events to zero in the years 2016, 2018, and 2019 (only 2 CSO events were recorded in 2017 (Gallagher and Wasik, 2019)).

Sewage indicators such as *Arcobacter* and *Acinetobacter* increased (Mann Whitney U, $p < 0.05$) in proportion during the wet weather events (with and without CSOs) (Figure 15). The fecal indicator bacteria *Bacteroides* increased in proportion during the wet weather events with CSOs compared top without at sites 86 (upstream), 76, 59 (downstream), and 57 (a downstream tributary); but it was a comparatively rare taxon (average of 1.51% of total). *Flavobacterium* increased significantly (Mann Whitney U, $p < 0.05$) in the post-disinfection and post-TARP phase at downstream sites (2016–2019), compared to the pre-disinfection/TARP phase which could be attributed to both disinfection as well as TARP implementation. These results in the Calumet region thus highlight the efficiency of Calumet TARP reservoir implementation in regulating the sewer microbial signature across this region during the heavy rainfall events, precisely characterized with CSOs.

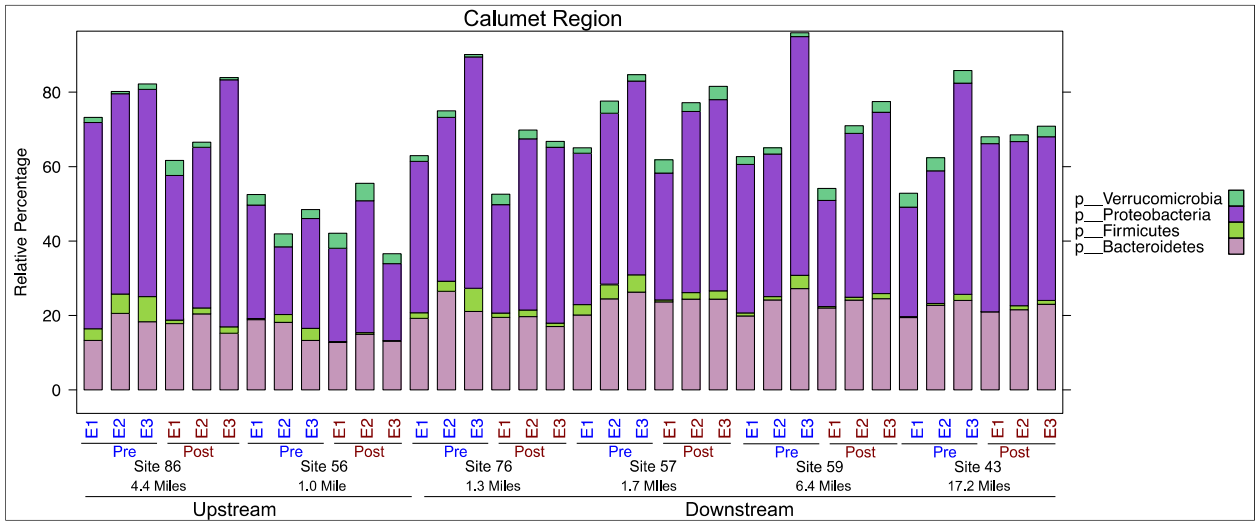


FIGURE 14 Stack plots showing the abundance patterns of the dominant phyla (>1% relative abundance) at sites 86, 56, 76, 57, 59 and 43 of the Calumet region with samples collected from Pre (2013–2015) and Post- Disinfection/TARP (2016–2019) phases during dry weather events, wet weather events without (E2) CSOs and with CSOs (E3). Site 57 is a tributary and not directly downstream of the Calumet WRP.

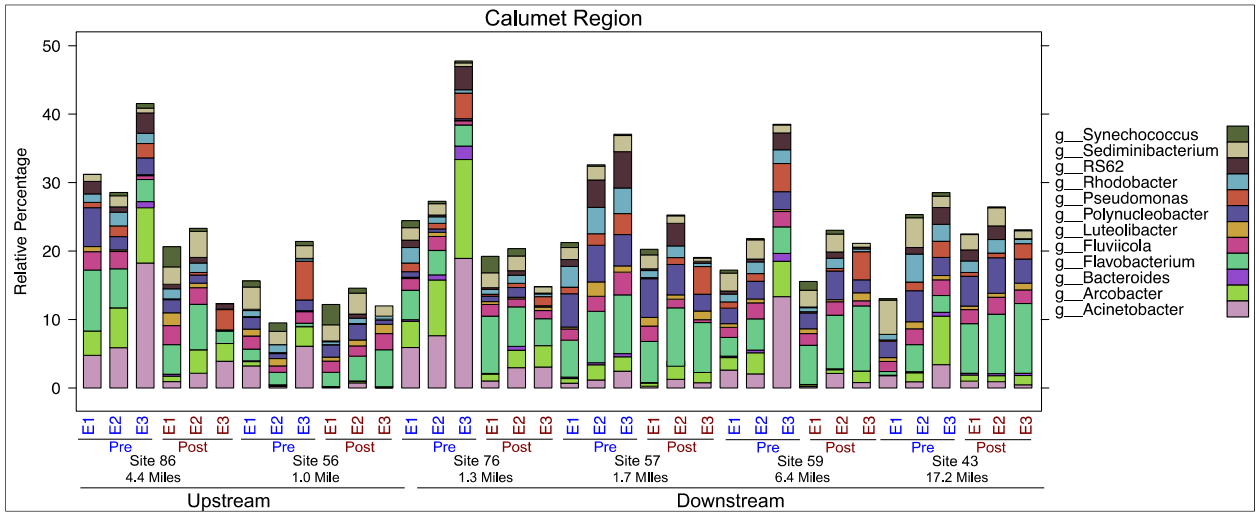


FIGURE 15 Stack plots showing the abundance patterns of the dominant genera (>1% relative abundance) at sites 86, 56, 76, 57, 59 and 43 of the Calumet region with samples collected from Pre (2013–2015) and Post- Disinfection/TARP (2016–2019) phases during dry weather events, wet weather events without (E2) and with CSOs (E3). Site 57 is a tributary and not directly downstream of the Calumet WRP.

In the O'Brien region (sites 96, 112, 36, 73), the phyla Firmicutes increase (Mann Whitney U, $p < 0.05$) in proportion in the wet events (Figure 16), and with CSO events during wet events. Post-disinfection, Firmicutes decreased (Mann Whitney U, $p < 0.05$) in proportion at the downstream sites (36 and 73), but not the upstream site 112. This further suggests that this reduction at the downstream sites can thus be attributed to the disinfection process (2016–2019). At site # 96 which is a tributary feeding to the North Shore Channel downstream of the O'Brien WRP (~3.2 miles), Firmicutes remained at similar levels throughout the study, irrespective of wet and dry events, likely because this site has no direct contribution from the O'Brien WRP. *Acinetobacter* and *Arcobacter* were significantly proportionally enriched (Mann Whitney U, $p < 0.05$) during wet weather events with CSOs compared without CSOs (Figure 17), but only at the downstream sites. The upstream site # 112, showed an increase (Mann Whitney U, $p < 0.05$) in only *Arcobacter.* Site#96 (tributary river) had a unique microbial signature with no significant enrichment of sewer or fecal associated bacteria.

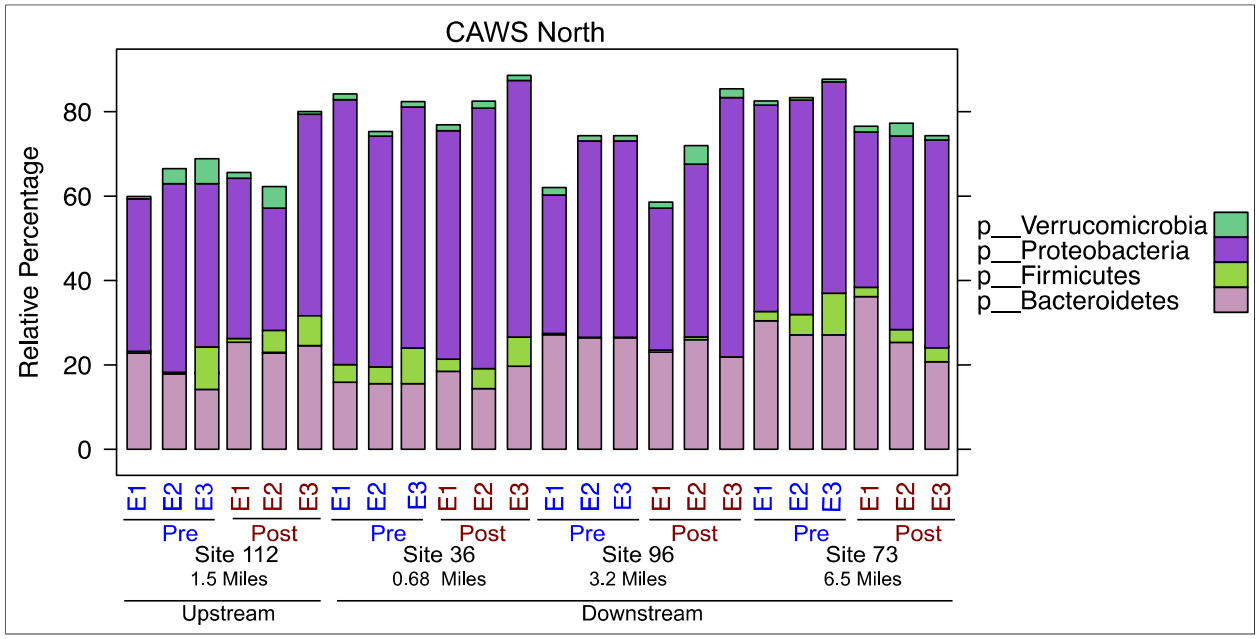


FIGURE 16 Stack plots showing the abundance patterns of the dominant phyla (>1% relative abundance) at sites 96, 112, 36, and 73 of the O'Brien North region with samples collected from Pre (2013–2015) and Post-TARP (2016–2019) phases during dry weather events (E1), wet weather events without (E2) and with CSOs (E3). Site 96 is a tributary and not directly downstream of the O'Brien WRP.

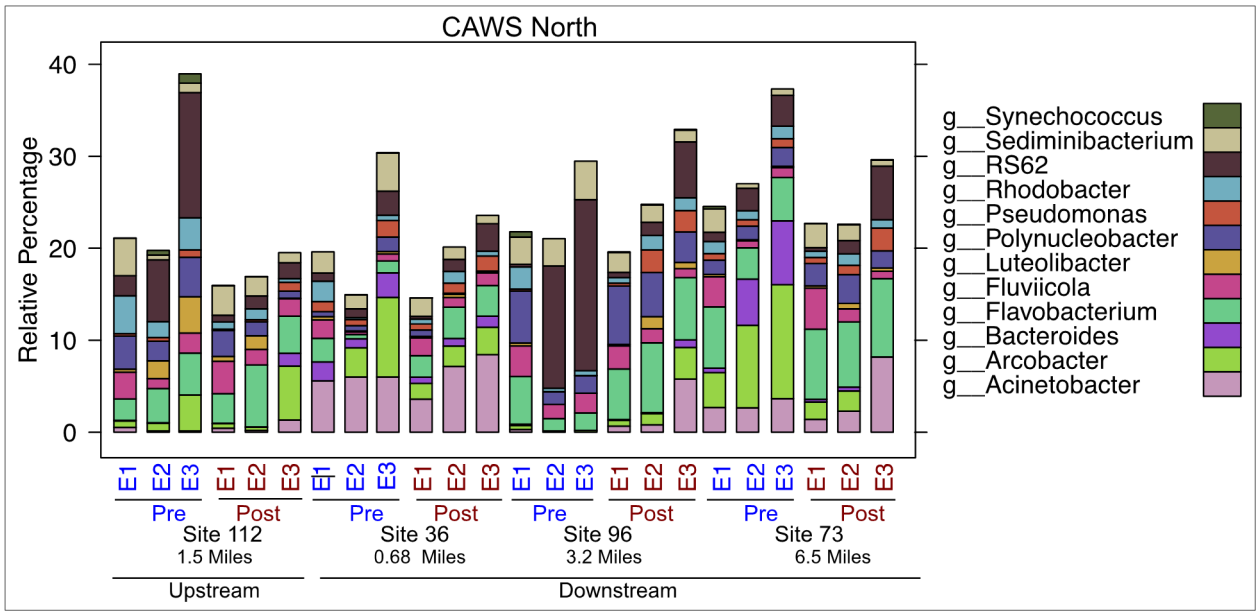


FIGURE 17 Stack plots showing the abundance patterns of the dominant genera (>1% relative abundance) at sites 96, 112, 36, and 73 of the O'Brien North region with samples collected from Pre (2013–2015) and Post-Disinfection (2016–2019) phases during dry weather events (E1), wet weather events without (E2) and with CSOs (E3). Site 96 is a tributary and not directly downstream of the O'Brien WRP.

In the main stem and south branch region (sites 100, 108, and 99), Firmicutes significantly proportionally increased (Mann Whitney U, $p < 0.05$) during the wet events (Figure 18). However, at site # 100, which is located downtown and doesn't have any direct inflow from the O'Brien WRP, Firmicutes had very low abundance, even compared to the other south branch sites. Firmicutes proportionally decreased in the post-disinfection phase. CSO events were characterized by increased (Mann Whitney U, $p < 0.05$) sewer associated taxa such as *Acinetobacter* and *Arcobacter* (Figure 19). But no specific trend (i.e., further reduction of these microbes) was observed between the pre- and post- TARP/disinfection years at the main stem and south branch sites.

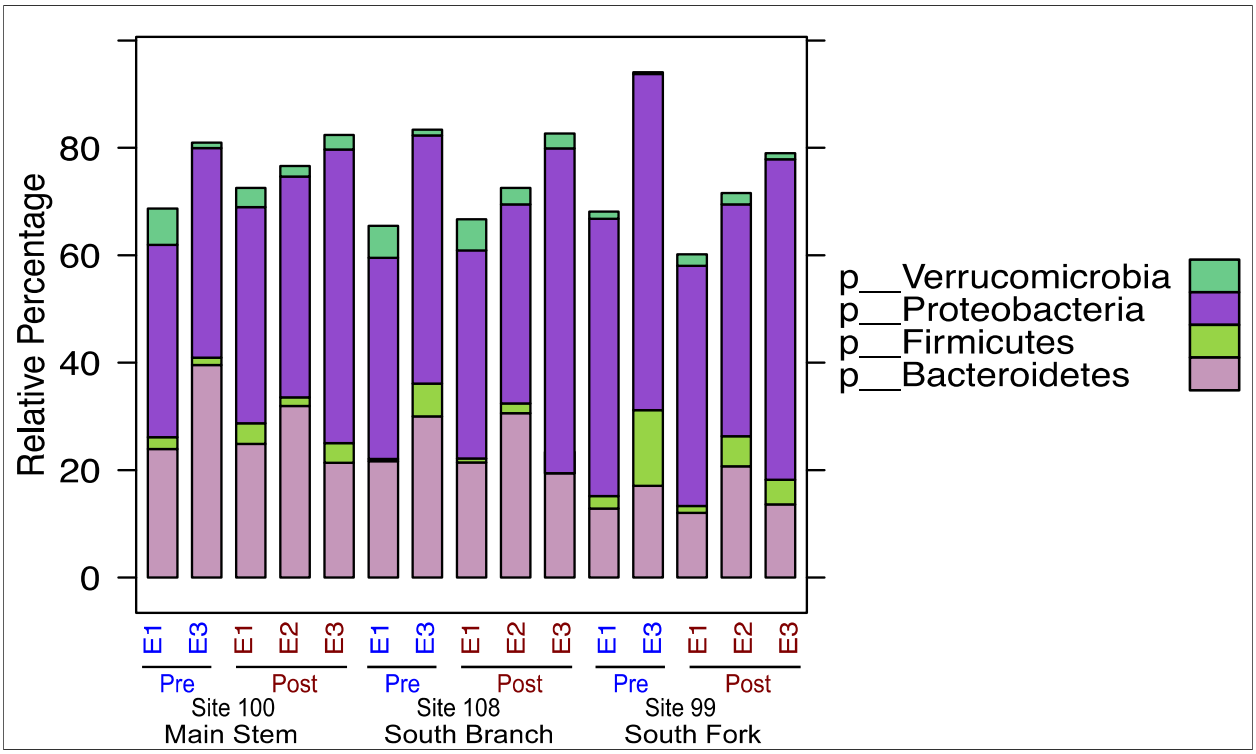


FIGURE 18 Stack plots showing the abundance patterns of the dominant phyla (>1% relative abundance) at sites 99 & 108 from the south branch and site 100 form the main stem of CAWS with samples collected from Pre (2013–2015) and Post-TARP (2016–2019) phases during dry weather events (E1), wet weather events without (E2) and with CSOs (E3). Please note that no wet weather events without CSOs were recorded in the pre-TARP years (2013–2015).

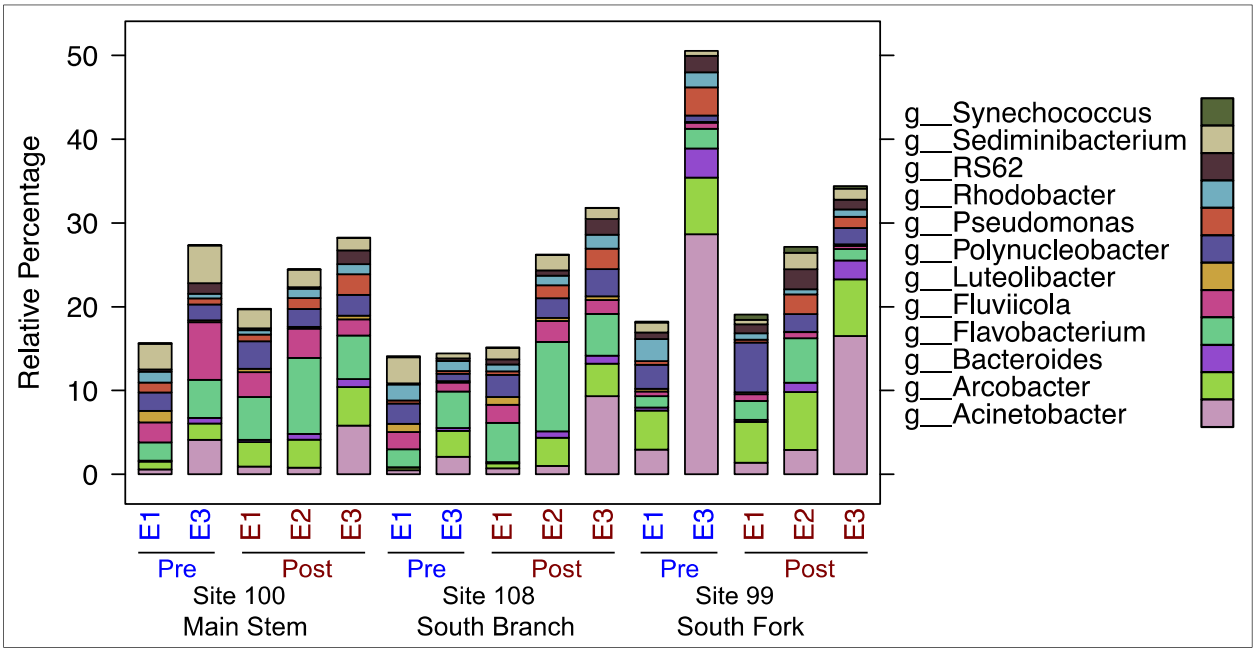


FIGURE 19 Stack plots showing the abundance patterns of the dominant genera (>1% relative abundance) at sites 99 & 108 from the south branch and site 100 form the main stem of CAWS with samples collected from Pre (2013–2015) and Post-Disinfection/TARP (2016–2019) phases during dry weather events (E1), wet weather events without (E2) and with CSOs (E3). Please note that no wet weather events without CSOs were recorded in the pre-TARP years (2013–2015).

1.3.5 Specific Microbial Community Strongly Correlates with Fecal Coliform Data

Quantitative fecal coliform data (Table 14, 15, 16) were correlated against the different taxonomic groups in the molecular microbiome data. Overall, linear regression fitting of microbiome data against fecal coliform data for all the sites, demonstrated that sewage associated phyla (Firmicutes, Bacteroidetes and Proteobacteria) were positively correlated (Spearman correlation, $p<0.05$) with increasing fecal coliforms. Verrucomicrobia was negatively correlated (Spearman correlation, $p<0.05$) with increasing coliforms. Fecal coliforms counts were also positively correlated with Firmicutes and Bacteroidetes proportions during the wet events compared to dry events (Figures 14, 16 and 18). *Acinetobacter, Arcobacter, Blautia, Bacteroides,* and *Prevotella*, etc. were positively correlated with fecal coliform counts. Whereas *Synechococcus, Sediminibacterium, Fluviicola* (river water indicators) were negatively correlated with coliforms (Figure 20).

Candidate taxa with correlation values (R2) greater than 0.6 for investigating their dynamics across different sites of CAWS during dry, wet and CSO events. *Acinetobacter, Arcobacter, Bacteroides, Cloacibacterium,* and *Trichoccus* were enriched (ANCOM, Mann Whitney U, $p < 0.05$) in wet events both with and without CSOs in the Calumet region (Figure 21), and these genera were reduced in proportion under the post-TARP phase when compared to pre-TARP phase. This trend was consistent between both the upstream and downstream sites highlighting the efficacy of both Calumet TARP reservoir (explained by

reduction of sewer microbiome at the upstream sites) as well as disinfection (explained by reduction at the downstream sites). *Flavobacterium, Sediminbacterium, Fluviicola* were significantly increased (ANCOM, Mann Whitney U, $p < 0.05$) in proportion during the wet weather events (rainfall) compared to the dry weather events both upstream and downstream sites with a significant reduction during the post-disinfection/TARP phase (Figure 21). Between the wet weather events with and without CSOs, we observed a significant increase (ANCOM, Mann Whitney U, $p < 0.05$) of the sewer microbes and a significant decrease (ANCOM, Mann Whitney U, $p < 0.05$) of river water indicators during the CSO events in contrast to the wet weather conditions without CSOs.

In the CAWS north region, we identified an increase in bacteria that correlated positively with fecal coliform data during the wet weather events (with and without CSOs) compared to the dry weather events and a decrease in the river water indicators (correlating negatively with the fecal coliform data) during the wet weather events (Figure 22). Across the downstream sites, we observed a decrease in the positively correlated bacteria during the post-disinfection phase but not across upstream sites. These results highlight the impact of disinfection process. In the main stem and south branch region a significant increase (ANCOM, Mann Whitney U, $p < 0.05$) in fecal coliform associated bacteria was observed during the wet weather events (compared to the dry events) although no specific trend was observed between the pre- and post-TARP/ disinfection phase (Figure 22). This can be explained by the fact that these sites are also distant from the O'Brien WRP and sites 99 and 100 have no direct O'Brien WRP influence.

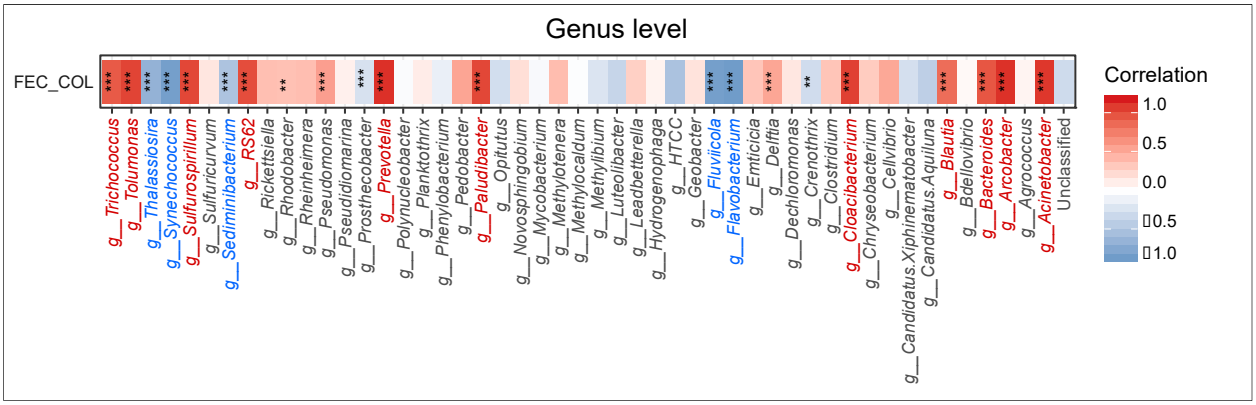


FIGURE 20 Correlation analyses based on Spearman Rank coefficient between microbial phyla, genera and fecal coliform data. * represents all the statistically significant correlations. The red color and gradient represent all positive correlations and the blue color along with its gradient represents negative correlations. The correlation scale varies between -1 to +1 representing R2 value of -100% to +100%. The * represents Benjamini-Hochberg FDR adjusted p-values ≤ 0.05. The ** represents Benjamini-Hochberg FDR adjusted p-values ≤ 0.01. The * represents Benjamini-Hochberg FDR adjusted p-values ≤ 0.001. The candidates listed in blue color (negative) and red color (positive) are the ones with correlation value greater than 0.6 and were hence selected for further community structure analyses during wet, dry and CSO events across multiple sites of CAWS.**

TABLE 14 Fecal coliform (CFU/100ml) per site before (Pre-) and after Disinfection/ TARP monitoring phase (Post-) at the Calumet region.

Site	Mean	Median	St.Dev.
Site_86_Post_E1	702.7	105.0	1608.4
Site_86_Post_E2	6038.8	3850.0	6718.2
Site_86_Post_E3	54000.0	54000.0	28284.3
Site_86_Pre_E1	18891.8	470.0	57648.4
Site_86_Pre_E2	2003.9	405.0	4454.0
Site_86_Pre_E3	28333.6	42000.0	24542.0
Site_56_Post_E1	10.7	5.0	12.0
Site_56_Post_E2	116.9	105.0	105.7
Site_56_Post_E3	27.5	27.5	31.8
Site_56_Pre_E1	302.0	25.0	774.0
Site_56_Pre_E2	1768.6	65.0	3535.0
Site_56_Pre_E3	3000.2	1100.0	4278.8
Site_76_Post_E1	46.4	40.0	38.1
Site_76_Post_E2	197.5	200.0	129.1
Site_76_Post_E3	65.0	65.0	7.1
Site_76_Pre_E1	10704.8	2200.0	21935.6
Site_76_Pre_E2	6117.5	2200.0	8310.4
Site_76_Pre_E3	33373.6	120.0	57700.2
Site_57_Post_E1	235.6	125.0	319.0
Site_57_Post_E2	7700.0	6900.0	2772.3
Site_57_Post_E3	6300.0	6300.0	1979.9
Site_57_Pre_E1	5704.5	390.0	16266.1
Site_57_Pre_E2	2781.4	2800.0	2440.6
Site_57_Pre_E3	3667.3	4000.0	3510.9
Site_59_Post_E1	53.7	30.0	69.2
Site_59_Post_E2	1890.0	1800.0	1343.8
Site_59_Post_E3	3400.0	3400.0	1979.9
Site_59_Pre_E1	1471.2	510.0	1985.2
Site_59_Pre_E2	5652.5	1050.0	12360.1
Site_59_Pre_E3	7733.6	3200.0	10742.8
Site_43_Post_E2	4000.0	3900.0	1367.5
Site_43_Post_E3	1500.0	1500.0	141.4
Site_43_Pre_E1	310.1	40.0	481.0
Site_43_Pre_E2	740.1	380.0	800.7
Site_43_Pre_E3	16533.7	1600.0	27262.3

TABLE 15 Fecal coliform (CFU/100ml) per site before and after Disinfection monitoring phase at CAWS North region

Site	Mean	Median	St.Dev
Site_112_Post_E1	85.1	60.0	103.9
Site_112_Post_E2	170.0	50.0	261.5
Site_112_Post_E3	202405.7	14000.0	450566.9
Site_112_Pre_E1	243.7	110.0	389.0
Site_112_Pre_E2	3300.0	3300.0	0.0
Site_36_Post_E1	191.9	150.0	136.8
Site_36_Post_E2	190.0	200.0	38.3
Site_36_Post_E3	55004.3	1300.0	134600.7
Site_36_Pre_E1	10859.1	10500.0	9184.3
Site_36_Pre_E2	21000.0	21000.0	0.0
Site_96_Post_E1	619.5	170.0	1266.5
Site_96_Post_E2	2225.0	2350.0	1239.3
Site_96_Post_E3	83871.4	50000.0	136805.3
Site_96_Pre_E1	380.3	290.0	520.4
Site_96_Pre_E2	10000.0	10000.0	0.0
Site_73_Post_E1	451.9	380.0	572.4
Site_73_Post_E2	605.0	655.0	222.5
Site_73_Post_E3	160942.9	20000.0	361861.1
Site_73_Pre_E1	3288.0	2400.0	3649.5
Site_73_Pre_E2	81000.0	81000.0	81000.0

TABLE 16 Fecal coliform (CFU/100ml) per site before and after Disinfection monitoring phase at Main Stem and South Branch Chicago River System.

Region	Site	Mean	Median	Stdev
Main stem	Site_100_Post_E1	4443.6	1040.0	6941.1
	Site_100_Post_E2	9065.0	7250.0	8737.9
	Site_100_Post_E3	11614.3	11000.0	8800.1
	Site_100_Pre_E1	7611.8	505.0	23087.1
	Site_100_Pre_E3	4200.0	4200.0	0.0
South Branch River System	Site_108_Post_E1	755.5	70.0	1396.5
	Site_108_Post_E2	10502.5	455.0	20332.5
	Site_108_Post_E3	59257.1	3600.0	90123.8
	Site_108_Pre_E1	396.1	300.0	435.8
	Site_108_Pre_E3	23150.0	23150.0	28072.1
South Fork River System	Site_99_Post_E1	3191.7	25.0	9467.6
	Site_99_Post_E2	26758.8	15015.0	36354.2
	Site_99_Post_E3	1831510.0	160000.0	4226156.6
	Site_99_Pre_E1	42094.7	25.0	134078.3
	Site_99_Pre_E3	675000.0	675000.0	106066.0

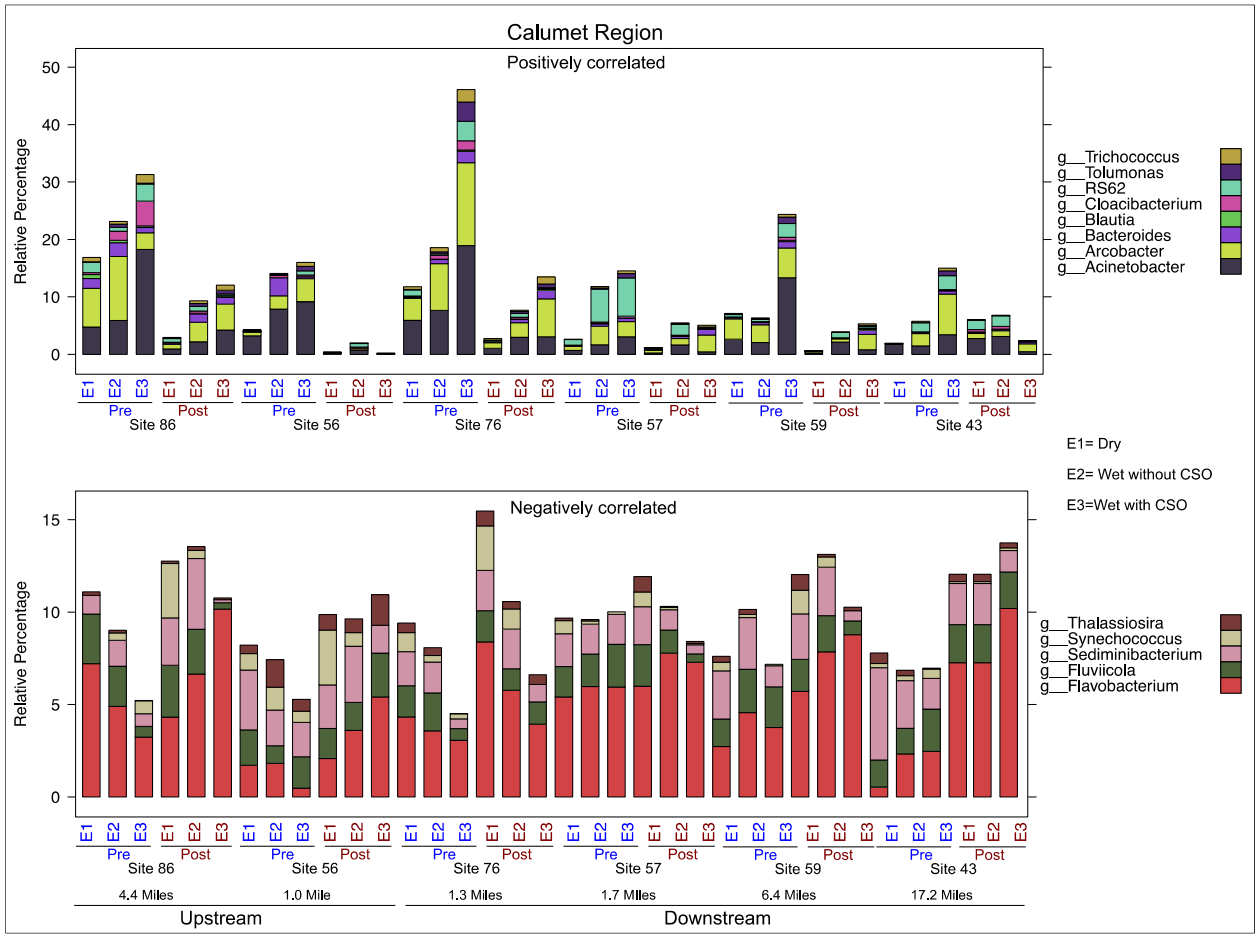


FIGURE 21 Stack plots showing distribution pattern of phyla that demonstrate significant correlation (positive and negative) with fecal coliform data, across the 6 sites of Calumet WRP, 4 sites of O'Brien WRP and 3 sites of the main stem of CAWS both pre- and post- Disinfection/ TARP. The distance in miles (from the WRPs) is mentioned along with each site. Site 57 is a tributary and not directly downstream of the Calumet WRP.

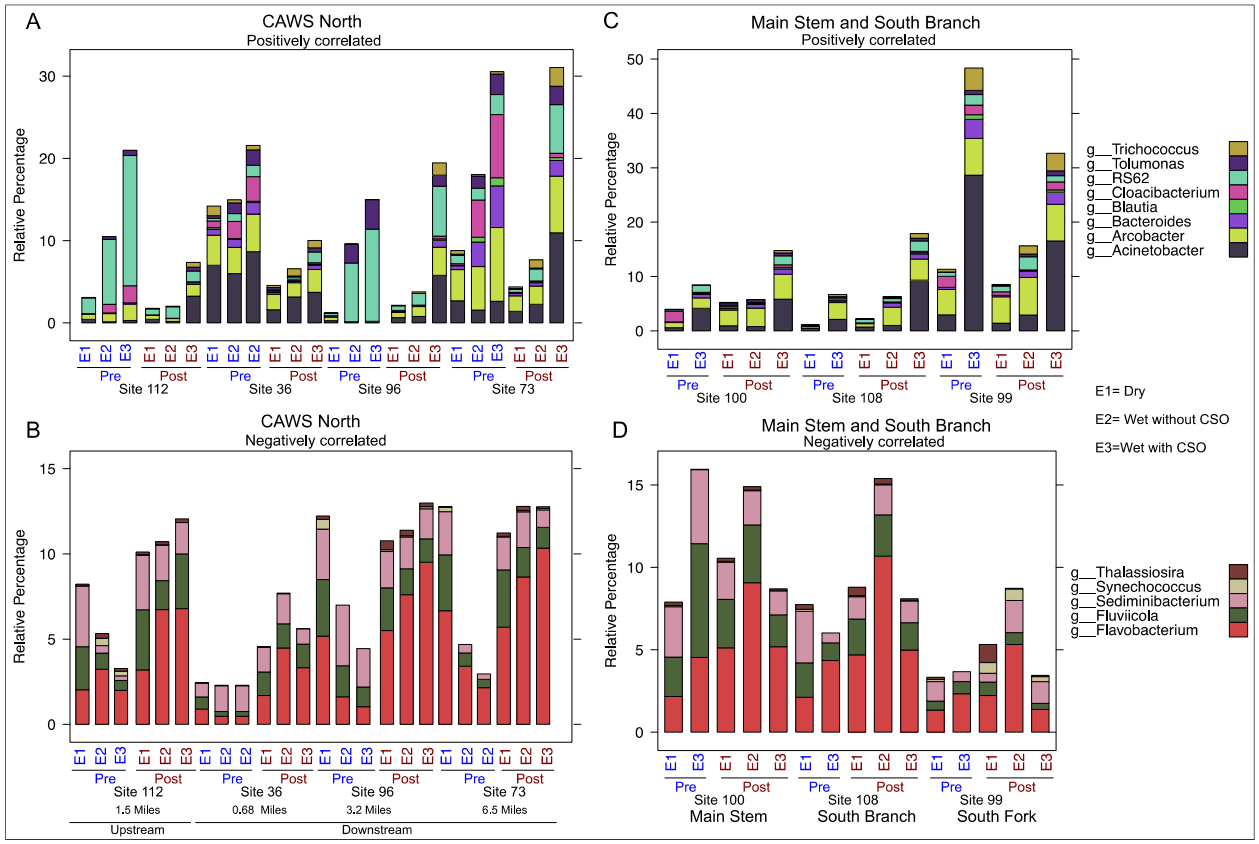


FIGURE 22 Stack plots showing distribution pattern of genera that demonstrate significant correlation (positive and negative) with fecal coliform data, across the 6 sites of Calumet WRP, 4 sites of O'Brien WRP and 3 sites of the main stem of CAWS both pre- and post-Disinfection/ TARP. The distance in miles (from the WRPs) is mentioned along with each site. Site 96 is a tributary and not directly downstream of the O'Brien WRP.

1.3.6 Sources of Microbial Organisms across the CAWS Sites

Fecal contamination of recreational waters is an increasing global health concern and therefore tracing the source of the contamination is an important step towards maintaining water quality. In this 7-year long microbiome study of Chicago Area Waterways System (CAWS) using 16S rRNA gene amplicon data, we employed the Markov-Chain Monte Carlo based analytical tool, SourceTracker, for microbial source tracking (MST). While conventional MST methods such as qPCR track specific microorganisms, this molecular sequencing approach enables broad spectrum identification of the probability that specific strain-level sequences originated from one of several different sources comprising 16S rRNA data. This provides a less accurate, but more comprehensive assessment of the impact of different sources on the composition of bacteria in the CAWS. We developed a manually curated database of different sources using CAWS samples (effluent, incoming sewage, sediment, fish gut and mucus) and an additional ~200,000 samples from the Earth Microbiome Project (EMP) latest release version (qiita.microbio.me/emp). The sources from the EMP database included- animal feces, fresh water, soil, human feces, mucus, and stream sediment. We tracked these sources for data from all

sites across the 7 years of observation from the Calumet region, O'Brien region, main stem, north and south branch Chicago river, and south fork river system.

SourceTracker analysis indicates that the sources of microbial diversity across all the river water samples can be largely attributed to effluent, sewage, CAWS sediment, river water, and fish associated samples (Figure 23). The three CAWS regions i.e., North (O'Brien), Main, and Calumet, have a unique compilation of potential sources that best explain the microbial signatures in those regions. For example, river water samples collected from the Calumet region show approximately equal contributions from fish mucus, effluent, and sewage samples; while river water samples collected from the North region have a dominant effluent signature. One caveat for this analysis is that we had very few fish associated samples (n=219) and effluent samples (n=92), which can bias quantitative source estimation. The EMP database is the largest collection of animal feces till date including sources from bats, primates, anteaters, sloths, wolf, buffalo, birds, insects, bear, bison, cow, kangaroo, emu, monkeys, dogs, rabbits, deer, fox, goose, geese, gulls etc. A separate database also exists for the aquatic biome covering different fish in EMP. The human fecal sources cover both urban as well as village settlements. In validation of the previous analysis suggesting low number of potential feces and sewage associated bacterial ASVs, the contribution made by human fecal matter and animal feces across all water column samples was extremely low, i.e., an average of 0.03% and 0.07% of all taxonomic units in each sample, respectively. However, ~40% of bacterial taxonomic diversity in the CAWS that cannot be reliably attributed to a 'source' (Figure 23). This is possibly because of an absence of urban river microbiomes in our 'source' databases. Additionally, it is likely that these are endemic but extremely rare taxa that are only found in the Chicago River system.

SourceTracker results from samples immediately downstream of the WRP outfalls, e.g., site#36, suggest that the majority of the ASVs can be mapped to the treated effluent (Figure 23); however both sites at O'Brien and Calumet had low concentrations of human or animal feces-associated taxa. For instance, well established FIBs such as *Bacteroides* (Cal~0.52%, O'Brien~0.50%), *Bifidobacterium* (Cal~0.11%, O'Brien~0.20%, *Blautia* (Cal~0.12%, O'Brien~0.15%), *Cloacibacterium* (Cal~0.49%, O'Brien~0.56%), *Prevotella* (Cal~0.15%, O'Brien~0.15%), were at <1% proportion across all seven years. *Acinetobacter* (Cal~7.52%, O'Brien~6.89) and *Arcobacter* (Cal~5.54%, O'Brien~3.74%) were the most dominant microbial signature in the effluent samples which are considered as 'sewer and treatment-plant associated' microbes, not necessarily fecal-associated (Fisher et al. 2014). Our results are validated by previous work suggesting that that human microbiome contributes only ~10% of the sewage and treated effluent community, while the sewer pipe and treatment associated bacteria comprise ~80% (VandeWalle et al. 2012).

The CAWS water column microbiome was primarily sourced from its sediment. Sediment samples were enriched for *Dechloromonas* (Calumet region: 5.42%, CAWS north: 7.42%, Main stem and south branch: 5.40%), *Thiobacillus* (2.73%, 2.06%, 0.74%), *Acinetobacter* (0.53%, 0.87%, 2.41%), *Geobacter* (0.78%, 1.43%, 0.63%), *Syntrophobacter* (1.08%, 0.86%, 0.79%), *Desulfococcus* (0.90%, 0.54%, 0.43%). Interestingly, the second largest contribution was from fish-associated mucus (avg. 16%), which comprised mostly *Acinetobacter* (avg. 5.1%), *Clostridium* (4.0%), *Polynucleobacter* (3.2%), *Pseudomonas* (2.8%), *Emticicia* (2.2%) and *Flavobacterium* (1.8%). Fish guts on the other hand contributed an average of 2%

towards the water column community, with *Mycoplasma* comprising 13% of the fish gut community.

While this technique is promising, the lack reference databases, geographic variability, the cost and time to build high-throughput sequencing libraries are major limitations to widespread application. Hence, 16S rRNA surveys need to be supplemented with shotgun metagenomics and qPCR that provide greater taxonomic and functional resolution of the microbial community, enabling more accurate analysis of fecal indicator bacteria. While we were unable to apply qPCR, we have explored the use of shotgun metagenomics below.

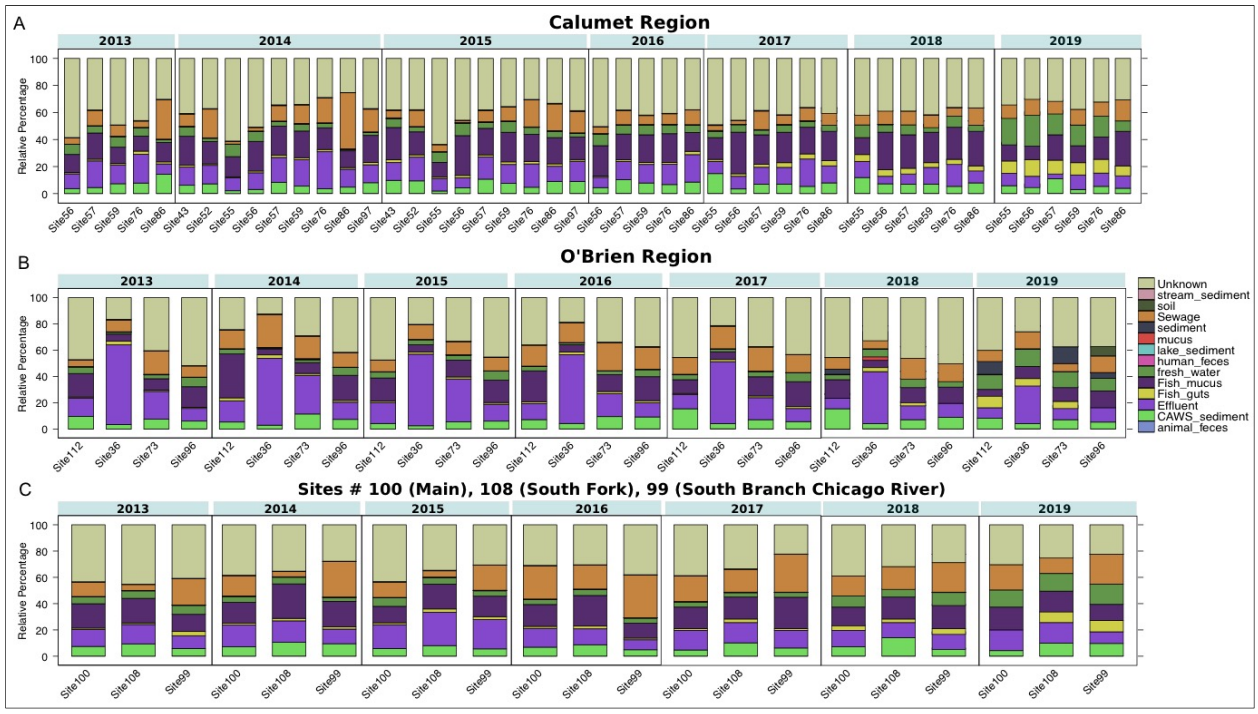


FIGURE 23 CAWS Microbial Community Sources Using Earth Microbiome Project Database. SourceTracker 2.0 analysis of water column samples by sampling site for years 2013–2019 using a curated database for (A) Calumet region, (B) O'Brien region, and (C) Sites # 100 from Main, 108 from the South Branch Chicago River, 99 from South Fork River System. A curated database was built using CAWS samples (i.e., effluent, sewage, sediment, fish gut and mucus) and additional ~200,000 samples from the Earth Microbiome Project (EMP) latest release version (qiita.microbio.me/emp). The sources from the EMP database included- animal feces, human feces, fresh water, soil, and stream sediment.

1.3.7 Employing Shotgun Metagenomics to Enhance Taxonomic Resolution and Examine Virulence and Resistance Factors across the CAWS during Dry and Wet Weather Events

Metagenomic sequencing enables the analysis of all genomic DNA in a sample, which provides enhanced taxonomic resolution when compared to the amplicon sequencing analysis of a single gene such as 16S rRNA; but metagenomics also provides access to the analysis of all the functional genes which code for proteins and enzymes in the bacterial genomes, enabling us to ask novel questions that 16S rRNA analysis can only infer answers to. However, due to cost-limitations, we only selected a subset of samples (n=71) for deep shotgun metagenomic sequencing and characterization. The sample set included upstream (sites 56, 112) and downstream (76, 36) sites at Calumet and O'Brien WRPs and sites from main stem (100), south branch (sites 99, 108) sampled from 2013–2019 including pre- (2013–2015) and post-disinfection (2016–2019) phases and dry and wet weather events. Please note some of the samples failed sequencing due to low starting biomass (i.e., low concentration of DNA for metagenomic library preparation and sequencing). These samples are listed in the index of each figure. Overall, the shotgun data for all the sites supported the 16S rRNA analysis data, suggesting a low proportion of human fecal indicators such as *Bacteroides* (~0.08%), *Prevotella* (0.05%), *Bifidobacterium* (0.01%), *Cloacibacterium* (0.06%) etc. in the CAWS water (Figures 24, 25, 26).

In the Calumet region, *Limnohabitans* spp. and *Acinteobacter johnsonii* were at significantly greater (ANCOM, Mann Whitney U, $p < 0.05$) proportion in downstream site#76 (Figure 24). *Limnohabitans* (Betaproteobacteria, Comamonadaceae) is considered a key bacterioplankton freshwater genus known to use dissolved organic carbon (DOC) from algae, and provide substrates for photolytic chemical transformation for other bacterial species (Kasalický et al. 2013) (Jezberová et al. 2017). *Acinetobacter johnsonii* is a well-known activated sludge-associated bacterium involved in phosphate removal (Kim, Hao, and Wang 1997). At both the upstream and downstream sites, *Thiomonas* was significantly enriched (ANCOM, Mann Whitney U, $p < 0.05$) during wet weather events (Figure 24); *Thiomonas* is a sulfur oxidizing bacteria often associated with sludge and sewage water (Huber et al. 2016). In downstream site#76, the proportion of *Arcobacter* increased (ANCOM, Mann Whitney U, $p < 0.05$) during the wet weather, which validates the 16S rRNA amplicon data.

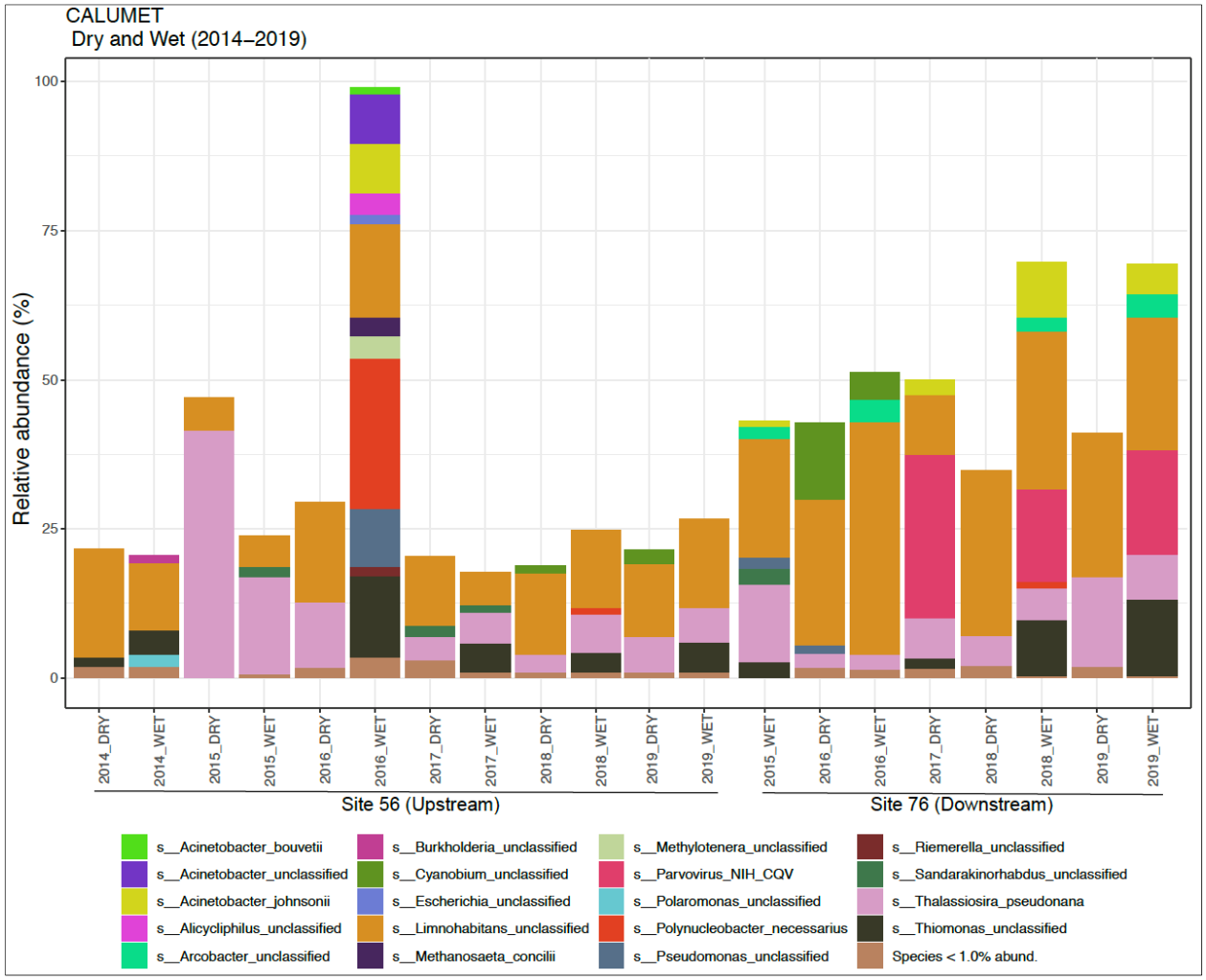


FIGURE 24 Stack plots showing the abundance patterns of the 20 most abundant species including species with <1% proportion at the upstream (site#56) and downstream (site#76) sites of the Calumet WRP with samples collected from 2014–2019 during the dry and wet weather events. Samples collected during a wet weather event (in 2017) and dry weather event (in 2015) from site 76 failed sequencing. The year 2013 in the Calumet region had no recorded wet events, hence the samples were not used in shotgun sequencing. Species whose proportional abundance is less than 1% are merged as one stack.

At O'Brien, the proportion of *Pseudomonas* significantly decreased, (ANCOM, Mann Whitney U, $p < 0.05$) while *Acinetobacter*, *Arcobacter* and *Thiomonas* significantly increased (ANCOM, Mann Whitney U, $p < 0.05$) in the downstream site#36, compared to upstream site #112 (Figure 25). *Acinetobacter*, *Arcobacter* and *Thiomonas* are all well-established sewage and sludge associated bacteria. *Acinetobacter* and *Arcobacter* species significantly increased in proportion during the wet weather events at the downstream site#36.

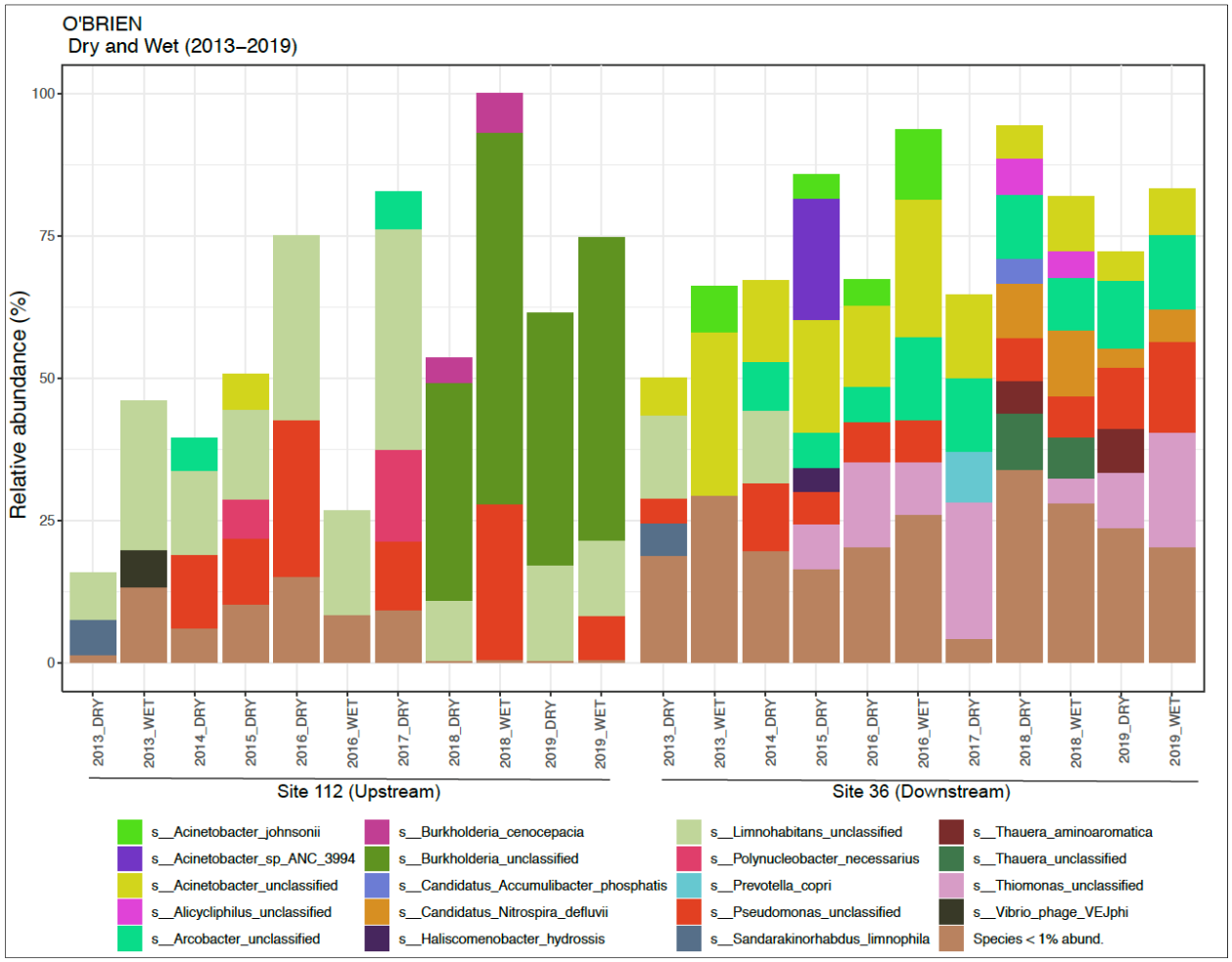


FIGURE 25 Stack plots showing the proportional trends of the 20 most abundant species including species with <1% proportion at the upstream (site#112) and downstream (site#36) sites of the O'Brien WRP with samples collected from 2013–2019 during the dry and wet weather events. No wet weather collections occurred during 2014 and 2015. Species whose proportional abundance is less than 1% are merged as one stack.

Microbial dynamics between the dry and wet weather events at site#100 from the main stem region, and sites#99 and #108 from the south branch river system (Figure 26) was also evaluated. The proportion of *Pseudomonas* increased (ANCOM, Mann Whitney U, $p < 0.05$) during the wet weather events at all the three sites, but only in the years 2016 and 2017. At site #100 specifically, *Burkholderia cenocepacia* and *Arcobacter* spp. significantly increased (ANCOM, Mann Whitney U, $p < 0.05$) in proportion during wet weather events. At site#108, like site #100, *Burkholderia* species significantly increased (ANCOM, Mann Whitney U, $p < 0.05$) during the wet weather events compared to the dry events. *Burkholderia* species occur in particle-rich waters, which suggest it may have come from eroded soils washed into the river during heavy rainfall (Zimmermann et al. 2018). At site #99 (a tributary of the south branch), there was an increase in *A. johnsonii* and *Arcobacter* species during the wet weather events.

Site#99 was enriched compared to the other sites in *Polynucleobacter necessaries,* a free-living river water bacterium (Hahn et al. 2016; Sangwan et al. 2016).

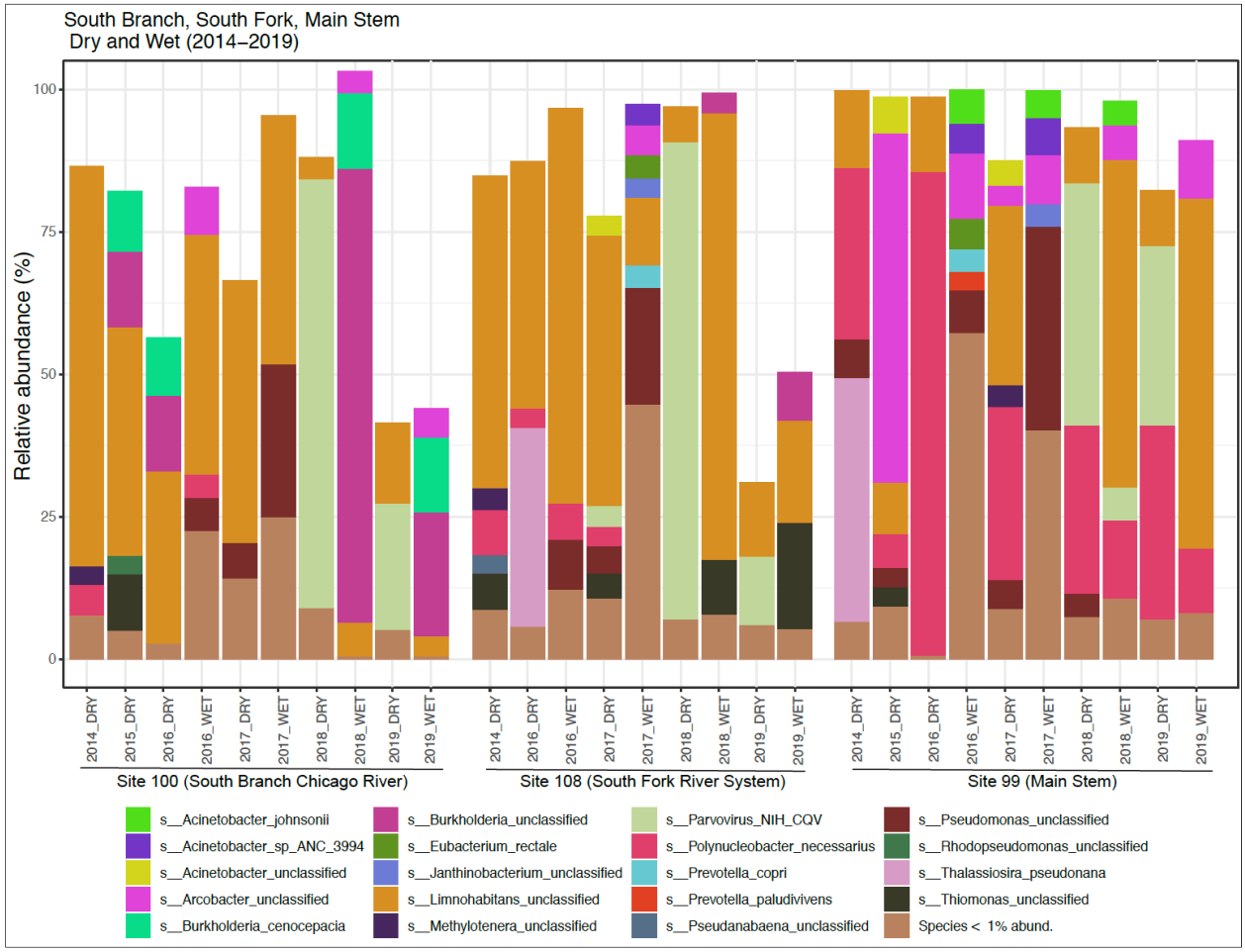


FIGURE 26 Stack plots showing the abundance patterns of the 20 most abundant species including species with <1% proportion at the sites 99 & 108 from the south branch river system and site#100 from the main stem of CAWS with samples collected from 2014–2019 during the dry and wet weather events. The 2015 dry weather event sample failed sequencing, and 2014 and 2015 had no sampled wet weather events. Species whose proportional abundance is less than 1% are merged as one stack.

While 16S rRNA amplicon sequencing data can be used to predict the proportion of genes that encode proteins and enzymes in a microbiome, shotgun metagenomics can be used to observe the actual metabolic potential. As observed in other environments (Fierer et al. 2012, Huttenhower et al. 2012, Tully et al. 2018), the majority of functional metabolic traits in the genome of the microbiome were conserved across sites. The top 20 most prevalent metabolic pathways encoded by the CAWS metagenome included amino acid biosynthesis (e.g., valine, isoleucine, arginine), aerobic respiration (cytochrome c based), pyruvate fermentation, nucleoside biosynthesis (guanosine, adenosine, pyrimidine), acetyl CoA biosynthesis, and

photosynthetic light reactions (Figures 27, 28 and 29); most of which are essential core metabolic processes found in the majority of bacteria. River waters are important transport pathways for solutes and matter fluxes of relevance to nitrogen, carbon and oxygen and thus primary productivity and functioning of the river ecosystems depend largely on the aerobic respiration which is driven by availability of carbon and oxygen (Vieweg et al. 2016). However, the wastewater treatment systems use anaerobic chambers for bacterial digestion of sludge/sewage which leads to enrichment of anaerobic pathways such as pyruvate fermentation. The presence of carbon and other hydrocarbons in the water also leads to enrichment of beta-oxidation process such as acetyl CoA biosynthesis. The metagenomic survey presented here validates a previous metatranscriptomic analysis of the Detroit river (Falk et al. 2019).

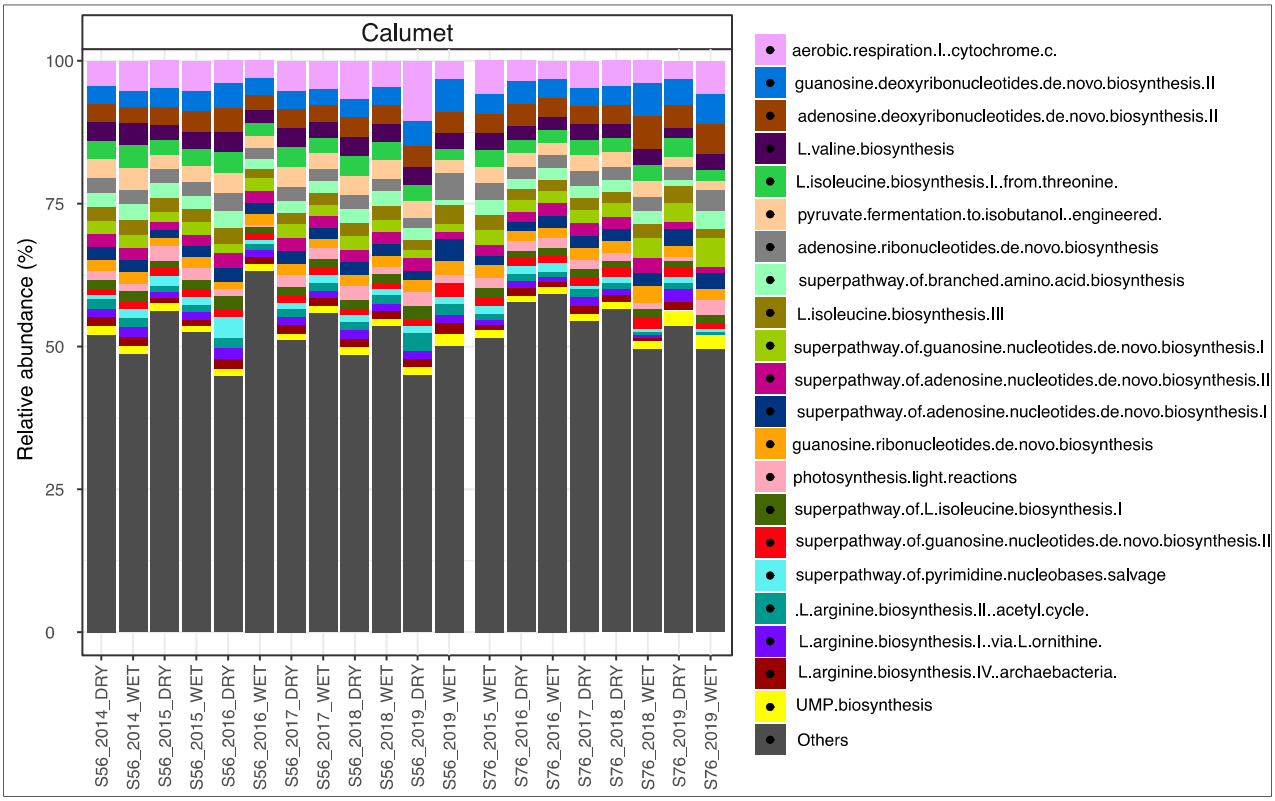


FIGURE 27 Stack plots showing the abundance patterns of the 20 most prevalent metabolic pathways from the upstream (site#56) and downstream (site#76) sites of the Calumet WRP with samples collected from 2014–2019 during the dry and wet weather events. Sample collected during a 2017 wet weather event and a 2015 dry weather event from site 76 failed sequencing. In 2013 the Calumet region had no recorded wet events.

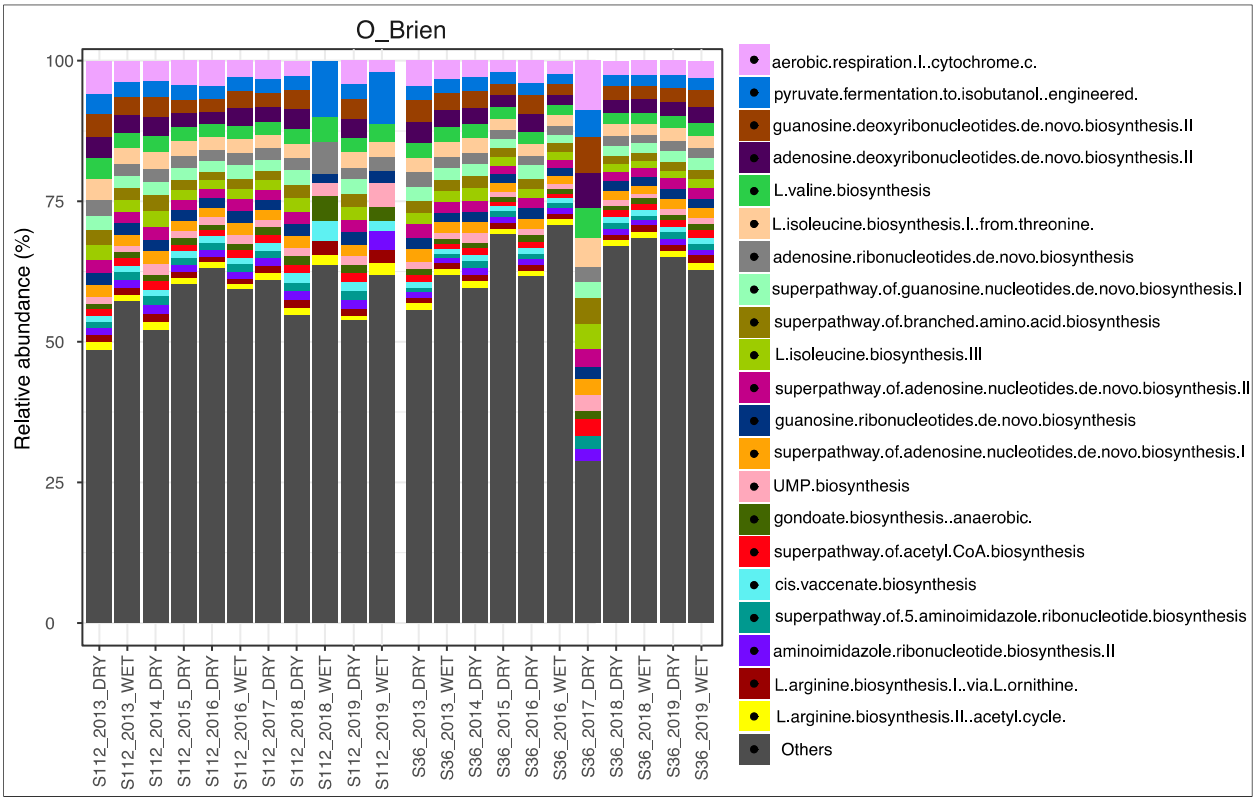


FIGURE 28 Stack plots showing the abundance patterns of the 20 most prevalent metabolic pathways from the upstream (site#112) and downstream (site#36) sites of the O'Brien WRP with samples collected from 2013–2019 during the dry and wet weather events. 2014 and 2015 had no recorded wet weather events.

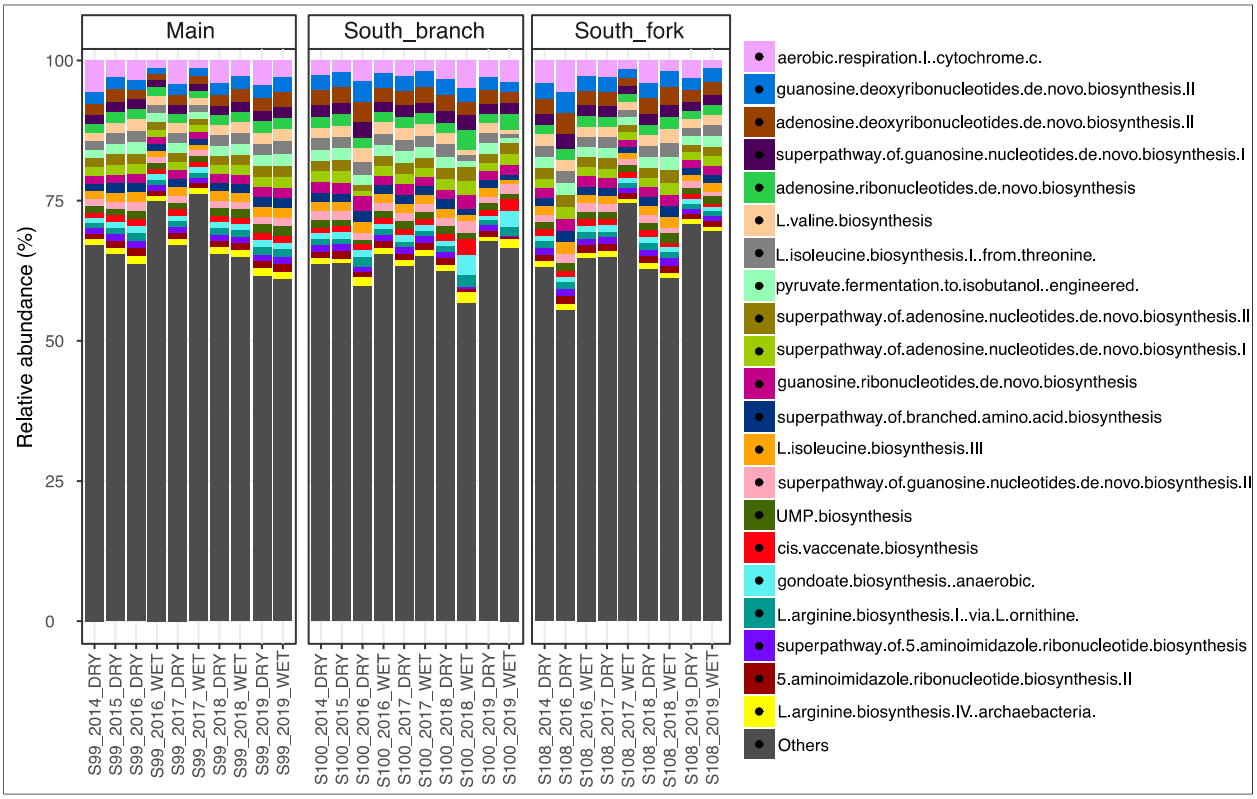


FIGURE 29 Stack plots showing the abundance patterns of the 20 most dominant pathways annotated at the sites 99 & 108 from the south branch river system and site#100 from the main stem of CAWS with samples collected from 2014–2019 during the dry and wet weather events. A sample from the 2015 dry weather event failed sequencing. 2014 and 2015 had no wet weather events recorded.

Outside of central metabolism pathways, the proportional variance of 30 virulence associated genes were also characterized, which overall were quite rare (0-0.25%; Figures 30, 31, 32). These genes demonstrated no proportional trends between the dry and wet weather events. While the low proportion of virulence genes may suggest low proportions of potential pathogens, it is also critical to note that other potential pathogens such as viruses and protozoa may not have been adequately characterized in this analysis. While, the shotgun data based functional annotations includes bacteria, viruses, protists and fungi, the overall representation of genomes of non-bacterial taxa still remains limited in the databases in contrast to the bacterial clades. Additionally, pathogens can be infectious even in low cellular or partial abundance, and our analysis does not quantify these abundances not provide information on infectivity rate.

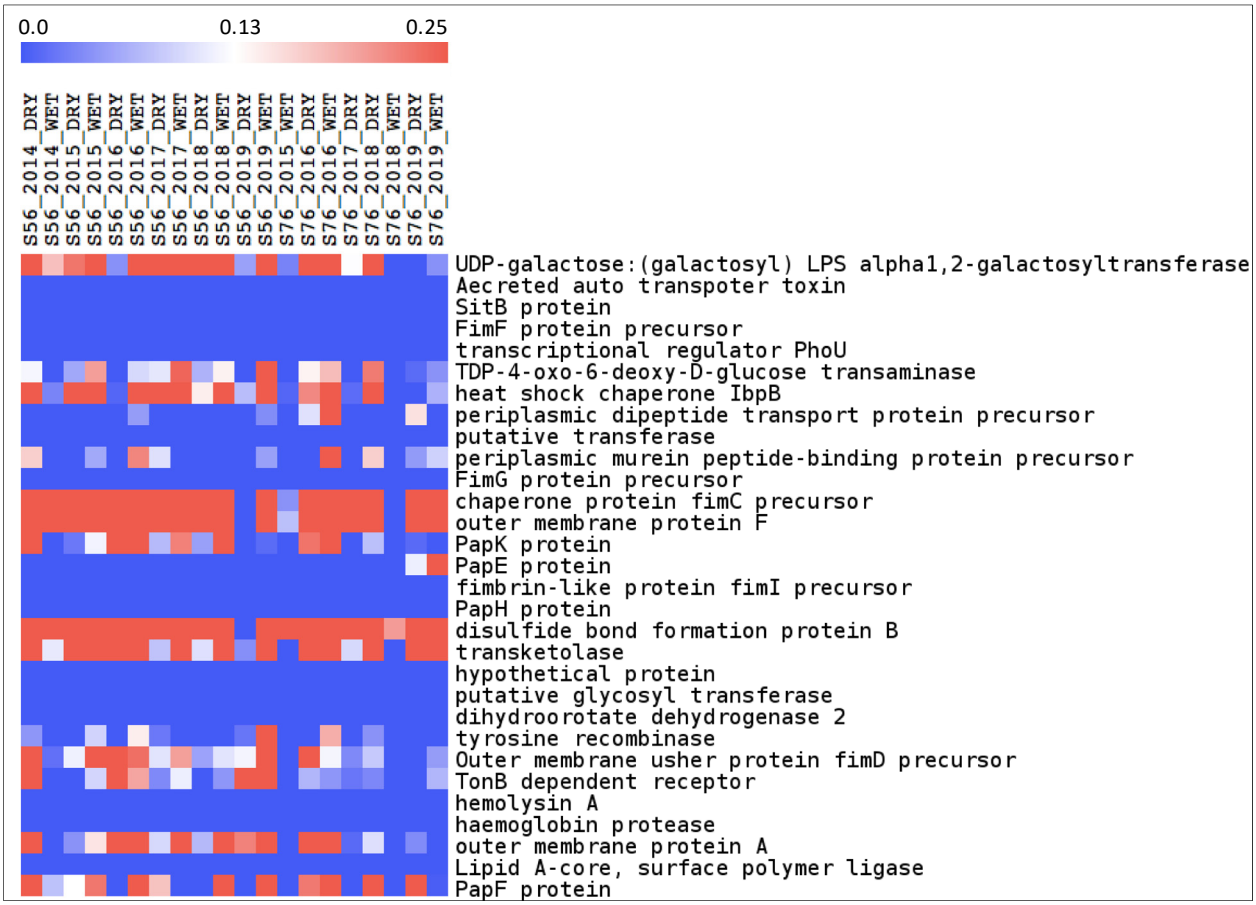


FIGURE 30 Heatmap showing the distribution of 30 virulence genes from the upstream (site#56) and downstream (site#76) sites of the Calumet WRP with samples collected from 2014–2019 during the dry and wet weather events. Samples from a 2017 wet weather event and 2015 dry weather event failed sequencing. 2013 in the Calumet region had no recorded wet events. The color scale stands for the relative abundance range of blue (low, 0.0%) to red (high, 0.25%).

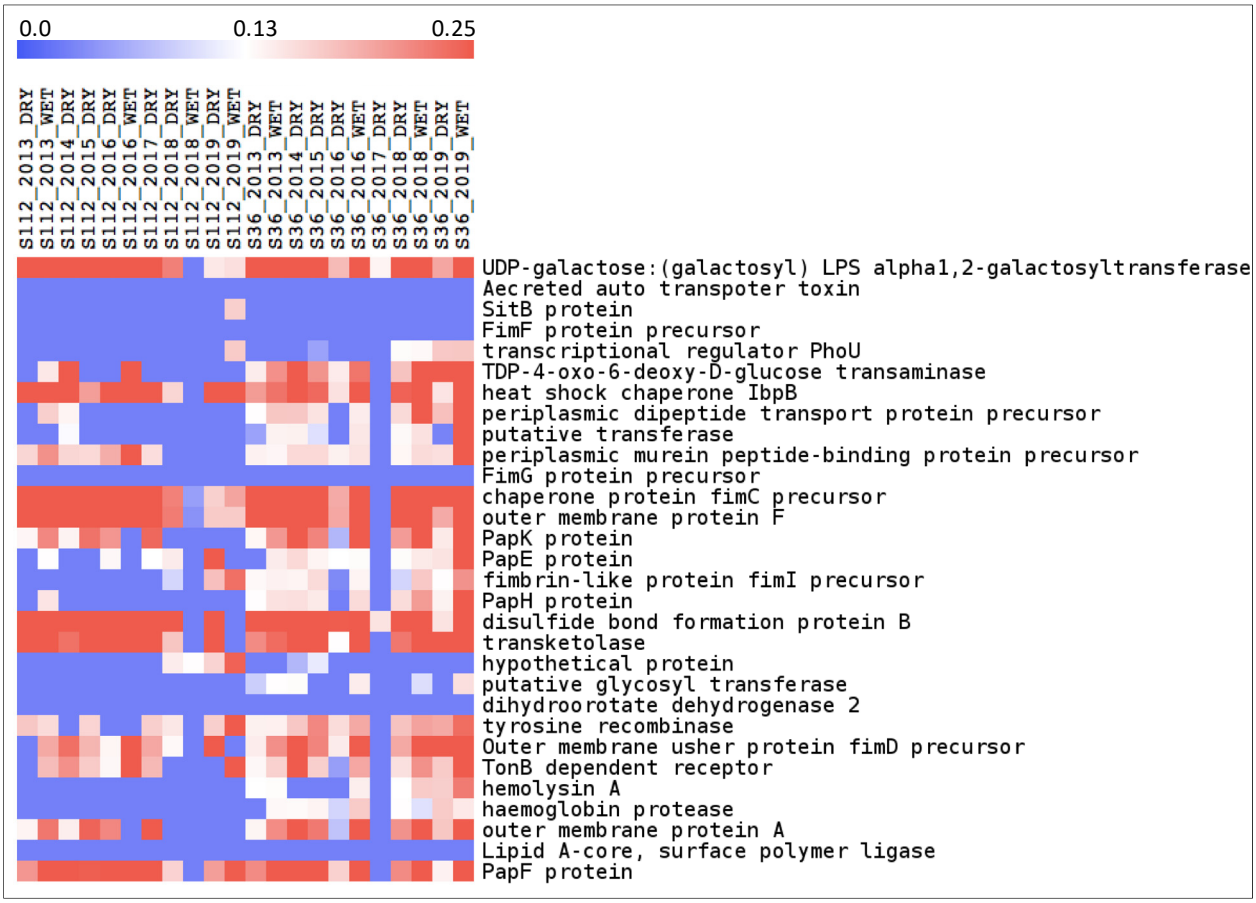


FIGURE 31 Heatmap showing the distribution of 30 virulence genes from the upstream (site#112) and downstream (site#36) sites of the O'Brien WRP with samples collected from 2014–2019 during the dry and wet weather events. 2014 and 2015 had no recorded wet weather events. The color scale stands for the relative abundance range of blue (low, 0.0%) to red (high, 0.25%).

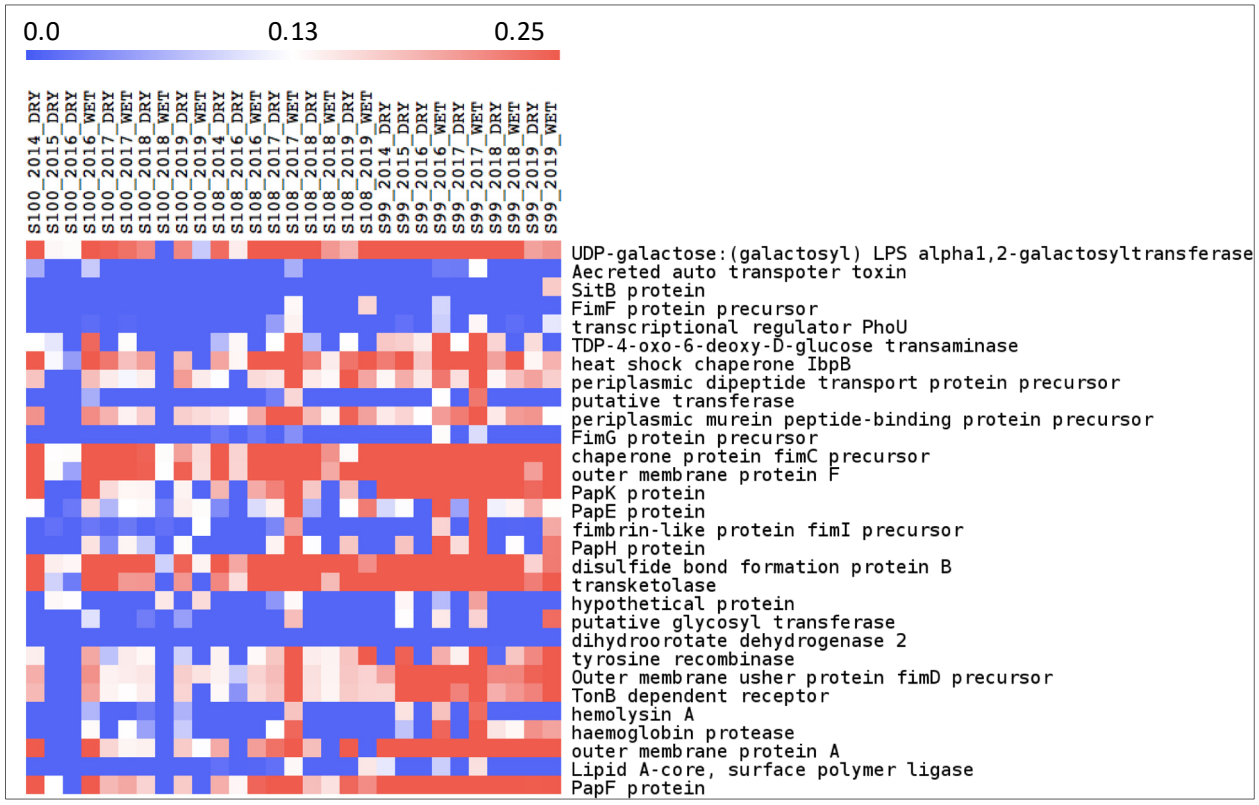


FIGURE 32 Heatmap showing the distribution of 30 virulence genes from sites 99 & 108 (South branch river system) and site#100 (main stem) of the CAWS with samples collected from 2014–2019 during the dry and wet weather events. The sample from the 2015 dry weather event failed sequencing. 2014 and 2015 had no recorded wet weather events. The color scale stands for the relative abundance range of blue (low, 0.0%) to red (high, 0.25%).

1.3.8 Strong Correlation between the Microbiome and Riverine Physicochemical Properties

The bacterial community interacts with the ambient environment and can be used as a bioindicator to reflect anthropogenic activities in aquatic ecosystems (Gibbons et al 2014). By correlating the physicochemical water quality indices with microbial community diversity and composition in the CAWS river ecosystem areas we aimed to determine if anthropogenic disturbance could be characterized. Correlation analysis was performed between microbial metrics and physicochemical properties of the water including pH, dissolved oxygen, nitrate, water temperature, specific conductance, flow, ammonia, total organic carbon, turbidity, chlorophyll, and volatile suspended solids.

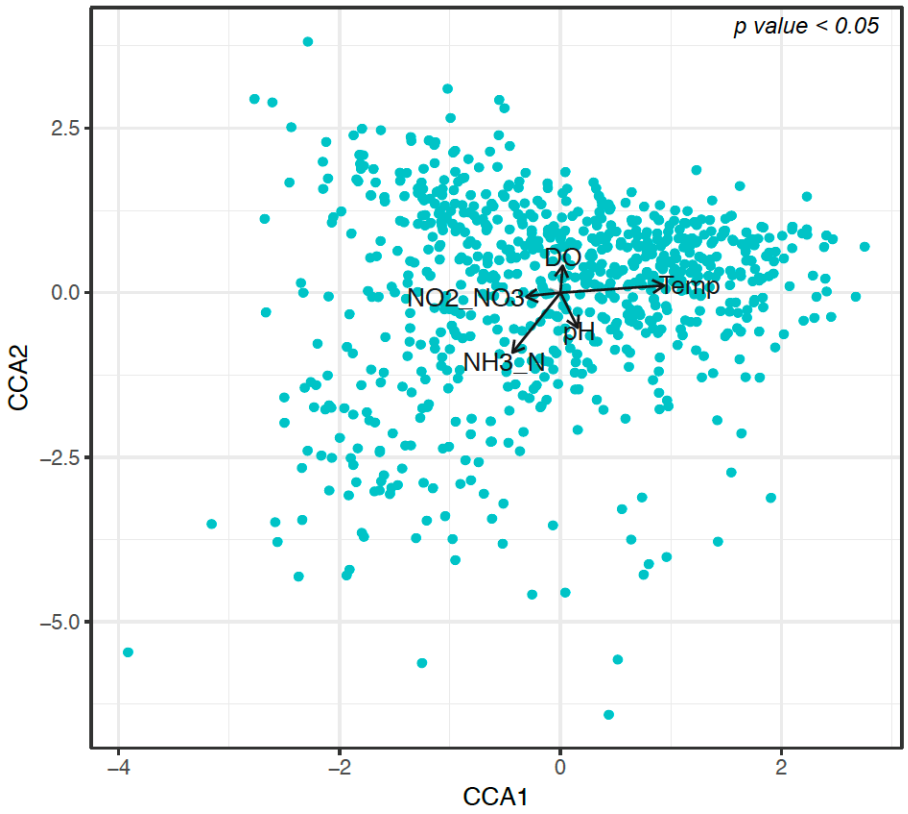


FIGURE 33 Canonical Correspondence Analysis (CCA) between microbiome and the physicochemical properties. It shows five most influential factors driving the microbiome changes.

Canonical Correspondence Analysis (CCA) revealed that temperature, NH3_N, DO, pH and NO2_NO3 are the most influential properties driving microbial changes (Figure 33). Using BEST analyses in microbiomeseq () package, we identified combination of variables (i.e., fecal coliform, VSS, Chlorophyll A, NH3_N, and DO) to best describe the variance in dominant phyla (i.e., Proteobacteria, Fusobacteria, Firmicutes, Bacteroidetes, and Cyanobacteria) using combined data from 2013–2019. However, when analyzed separately the pre-disinfection (2013–2015) and post-disinfection datasets (2016–2019) revealed slight variations in the best suite of physicochemical properties describing microbiome variance. While, during pre-disinfection phase, fecal coliforms, VSS, temperature, DO, chlorophyll A, NO2_NO3 remained the most influential set of variables, during the post-disinfection VSS, TOC, NH3_N, and chlorophyll A were the most important in describing variation in microbial diversity. Fecal coliforms data was identified as one of the most influential variables in driving microbial changes overall, however, in comparison to pre-disinfection phase, fecal coliforms data was not among the best set of variables in highlighting the change in total microbial diversity during post-disinfection phase.

Further using Spearman correlations, at the level of bacterial phylum (Figure 33), as expected Cyanobacteria (which contain chlorophyll A for photosynthesis) significantly positively correlated with chlorophyll A concentration**.** Both Cyanobacteria and Fusobacteria were significantly positively correlated with dissolved oxygen concentration. Fusobacteria and Firmicutes were positively correlated with ammonia concentration, while Chlamydiae and TM6 positively correlated with nitrate and nitrite concentration (Figure 34). At the genus level (Figure 34), chlorophyll A concentration was significantly positively correlated with the cyanobacterium *Synechococcus*; interestingly, chlorophyll A concentration was significantly negatively correlated with genera associated with fecal and sewage biomarkers (*Acinetobacter, Arcobacter, Bacteroides, Chryseobacterium*, and *Sulfurospirillum*). Conductivity positively correlated with *Thalassiosira*, *Rhodobacter*, and *Leadbetterella*. Dissolved oxygen concentration positively correlated with *Thalassiosira* and negatively correlated with *Methylocaldum, Dechloromonas, Cloacibacterium*. Ammonia concentration was positively correlated with *Trichococcus, Pseudomonas, Clostridium* and *Acinetobacter*, suggesting greater concentration associated with sewage outfall. Nitrate and nitrite concentrations were positively correlated with *Mycobacterium* and *Methylibium*, the latter being potentially associated with environmental pollution. pH only had weak correlations, likely due to the fact that it barely fluctuated. Water temperature was significantly positively correlated with *Synechococcus* and *Candidatus* Aquiluna, likely due to cyanobacterial blooms in the summer, and negatively correlated with *Trichococcus, Rhodobacter, Pseudomonas, Paludibacter, Flavobacterium, Clostridium*, and *Arcobacter*, which were generally more abundant during wet-weather events, which may associated with reduced temperatures due to stormwater input. Phosphate concentration positively correlated with *Mycobacterium*, maybe due to association with effluent outfall. Similarly, turbidity and VSS, which increases during wet weather events and at the treated effluent outfall, was positively correlated with *Trichococcus, Pedobacter, Clostridium, Cloacibacterium***.**

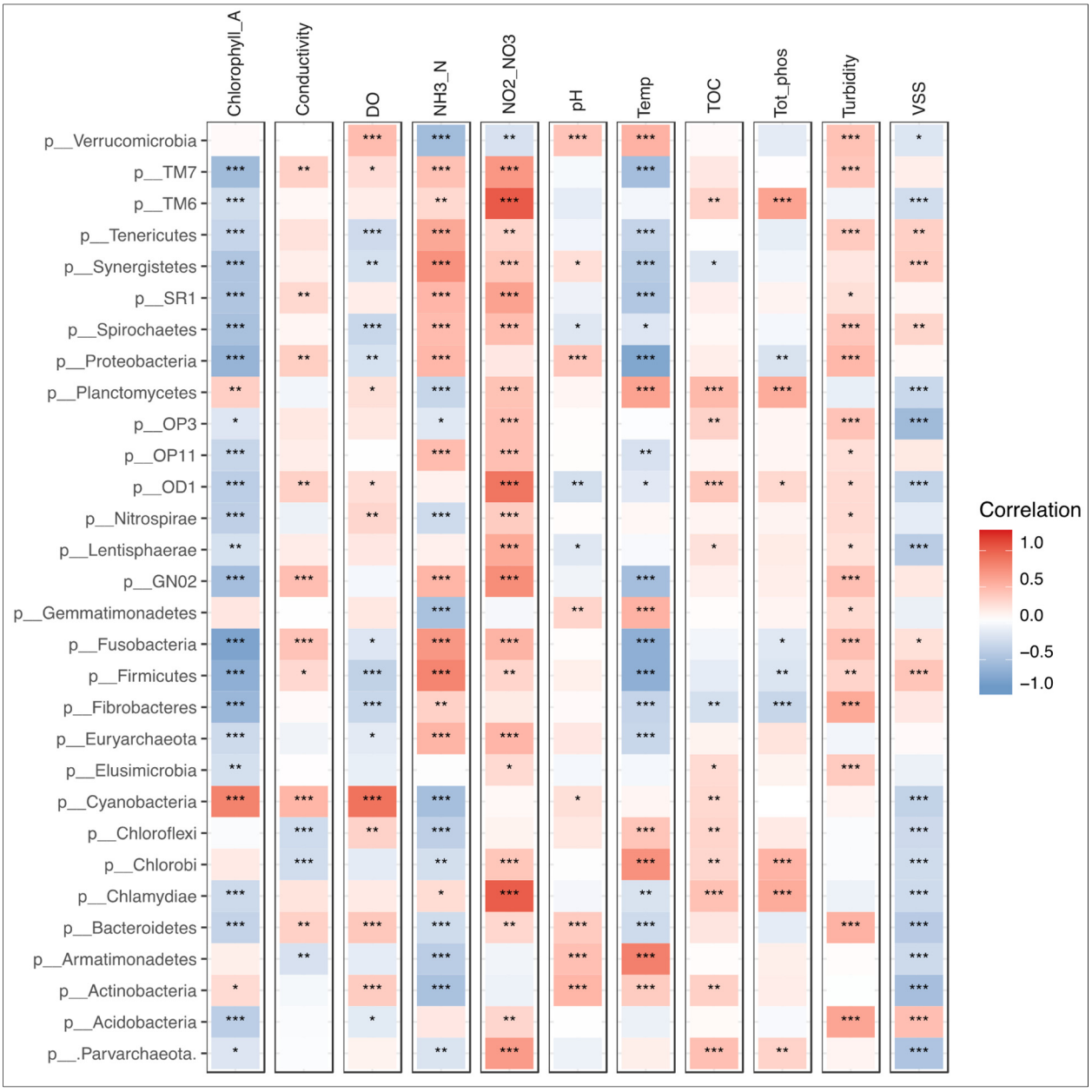


FIGURE 34 Correlation analyses based on Spearman Rank coefficient between microbial phyla and physicochemical properties of water samples. * represents all the statistically significant correlations. Red represents positive correlations, and blue represents negative correlations. The correlation scale varies between -1 to +1 representing R2 value of -100% to +100%. The * represents Benjamini-Hochberg FDR adjusted p-values ≤ 0.05. The ** represents Benjamini-Hochberg FDR adjusted p-values ≤ 0.01. The * represents Benjamini-Hochberg FDR adjusted p-values ≤ 0.001.**

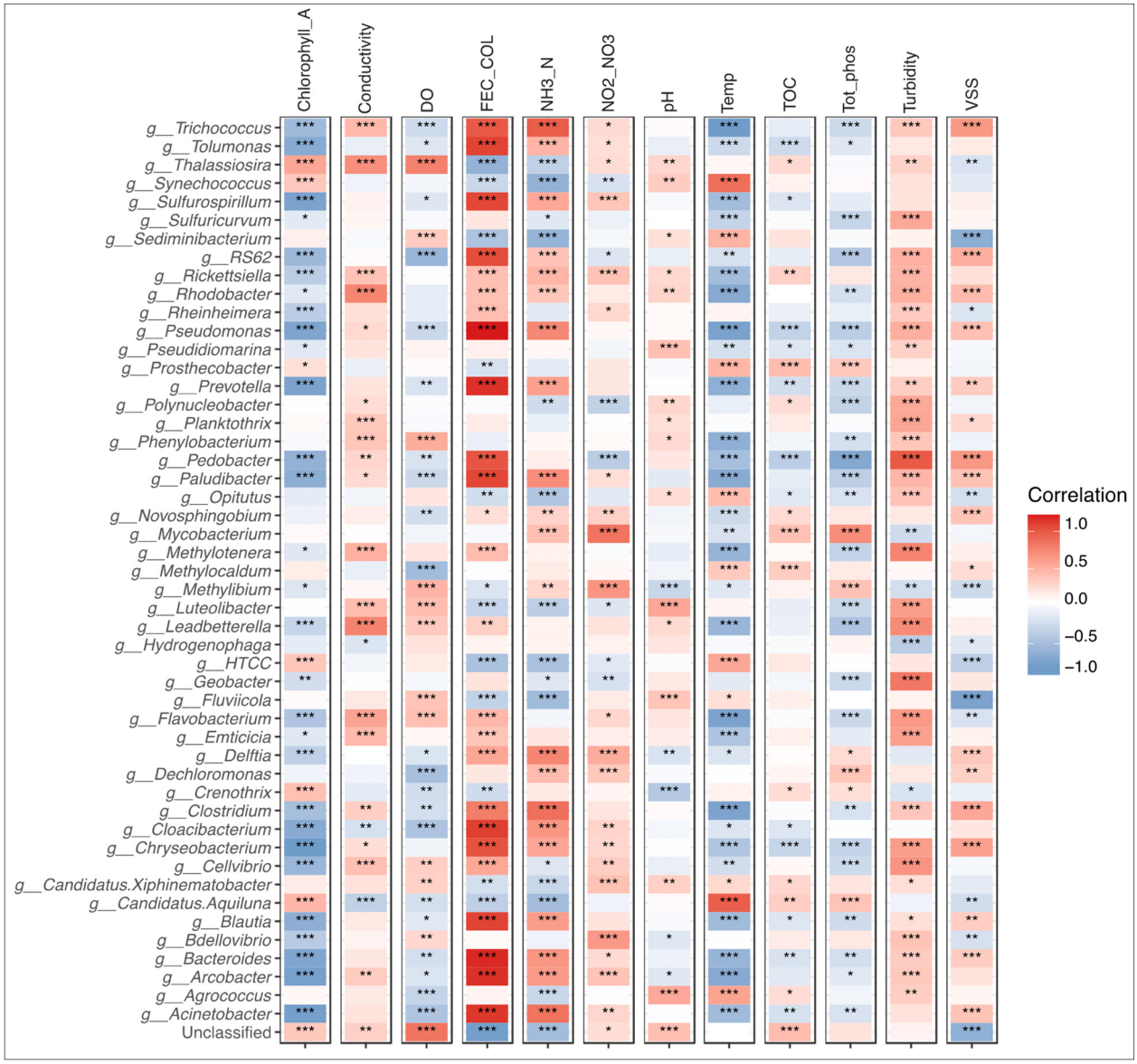


FIGURE 35 Correlation analyses based on Spearman Rank coefficient between microbial genera and physicochemical properties of water samples. *represents all the statistically significant correlations. Red represents positive correlations, and blue represents negative correlations. The correlation scale varies between -1 to +1 representing R2 value of -100% to +100%. The * represents Benjamini-Hochberg FDR adjusted p-values ≤ 0.05. The ** represents Benjamini-Hochberg FDR adjusted p-values ≤ 0.01. The * represents Benjamini-Hochberg FDR adjusted p-values ≤ 0.001.**

We further identified correlations between ASVs and the above-mentioned properties (except pH which had no significant correlations):

- **Temperature:** ASVs from *Methylotenera mobilis* were negatively correlated, while family Sphingobacteriaceae were positively correlated with increasing

temperature *Methylotenera mobilis* is involved in denitrification, but is grows best at room temperature (Mustakhimov et al. 2013). Sphingobacteriaceae has previously been shown to increase in abundance during summer months in river water (Bowers et al. 2012).

- **Dissolved Oxygen:** ASVs belonging to *Gemmatimonas* and *Fluviicola* were positively correlated with dissolved oxygen. *Gemmatimonas* has no known association, but *Fluviicola* has previously been associated with elevated dissolved oxygen concentrations (J. Liu et al. 2015). Reduced dissolved oxygen tends to happen when nutrient concentrations promote heterotrophic bacterial blooms that elevate biological oxygen demand and create anaerobic conditions.

- **Ammonia:** ASVs from genus *Arcobacter* and *Acinetobacter* were positively correlated with ammonia, whereas, ASVs belonging to *Fluviicola* and Oxalobacteraceae are negatively associated with ammonia. Activated sludge processing employs a mixed microbial consortium under aerobic conditions to remove carbon and nitrify ammonia, and is enriched in *Arcobacter* and *Acinetobacter* (Saunders et al. 2016). Fluviicola is a known N2 fixer and assimilates ammonium through ammonification and denitrification (Bentzon-Tilia et al. 2015).

- **Total organic phosphorous:** ASVs belonging to Sphingobacteriaceae, Actinomycetales, *Paludibacter* were positively correlated. *Paludibacter,* is a common activated sludge bacterium involved in phosphate metabolism (Zhou et al. 2016). Similarly, taxa associated with Sphingobacteriaceae and Actinomycetales are associated with biological phosphorous removal (Kamika, Azizi, and Tekere 2018). Hence, these bacteria are likely enriched at effluent outfall sites.

- **Volatile Suspended Solids:** ASVs belonging to Rhodobacteraceae, Cloacibacterium, and *Desulfobulbus* are positively correlated, while ASVs belonging to *Sediminibacterium* are negatively correlated with VSS. The biomass solids in a biological wastewater reactor comprise total suspended solids (TSS) and volatile suspended solids (VSS). Rhodobacteraceae are known to be dominant across activate biomass and sediments in wastewater treatment plants (Pohlner et al. 2019). *Cloacibacterium* is involved in activated sludge and biomass in wastewater treatment plants, producing the exopolysacchrides that comprise 95% of the sludge (Klai et al. 2015). *Desulfobulbus and Sediminibacterium* are also abundant in activate sludge. *Desulfobulbus* performed S reduction in WRPs (Nascimento et al. 2018).

- **Total Organic Carbon:** ASVs belonging to Oxalobacteraceae and Actinomycetales are negatively correlated with total organic carbon (TOC), while ASVs from *Prevotella copri* and Comamonadaceae are positively correlated with TOC. TOC is often used as a non-specific indicator of water

quality, with lower concentrations being better. *Prevotella* is usually associated with the human and animal gut, and so elevated levels of TOC and *Prevotella* might represent fecal contamination (Franke and Deppenmeier 2018). Oxalobacteraceae and Actinomycetales can both metabolize complex carbon compounds, but their negative correlation with TOC may indicate negative association with TOC pollution.

- **Chlorophyll:** ASVs from Actinomycetales, Mycobacterium are positively correlated with chlorophyll. Mycobacterial species are known to grow on aquatic plants and therefore may associate with algal blooms in the summer (Mougin, Tian, and Drancourt 2015).

- **Conductivity:** ASVs from Gemmatimonas, Rhodobacter, Methylotenera are positively correlated with conductivity. Rhodobacter species are non-sulfur bacteria which are known to convert dinitrogen into ammonia with production of H2, which in turn increases the water conductivity (Tao et al. 2012). *Methylotenera* is a methane utilizing bacteria, which indicate at high methane concentrations which in turn means increased conductivity of water (Wright et al. 2017). Gemmatimonas is known to be associated with conductivity although in case of soil/sand and not in water (Shi et al. 2017).

- **Fecal coliforms:** ASVs from Enterobacteriaceae, *Cloacibacterium, Parabacteroides, Acinetobacter, Pseudomonas, Bacteroides, Agrobacterium, Dialister*, and *Hydrogenophaga* are positively correlated with fecal coliforms. Enterococci, *Bacteroides, Acinetobacter, Cloacibacterium* and *Parabacteroides* are already well known fecal indicators and have been isolated from human stool (Leight, Crump, and Hood 2018; Fisher et al. 2015; García-Bayona and Comstock 2019; Zhang, He, and Yan 2015). These results supplement the previous reports provide a definite assessment of recreational water quality and human health risk. *Agrobacterium*, although not a fecal indicator bacteria, is known to exist in wastewater treatment plants where it is involved in denitrification under aerobic conditions (Ma et al. 2016). These results also highlight the potential of using microbial taxa as bioindicators for fecal nitrogen as water quality indicators.

- **Turbidity:** ASVs belonging to *Pseudomonas, Geobacter, Agrobacterium, Cellvibrio* are positively correlated with turbidity. *Pseudomonas* is normally associated with turbid events after rainfall and stormwater influx (Flores Ribeiro et al. 2014). Similarly, *Geobacter* is associated with wastewater treatment plants (Tejedor-Sanz et al. 2018), and so this correlation could be associated with effluent from the plant.

1.4 CAWS MICROBIOME RECOMMENDATIONS FOR WATER QUALITY MANAGEMENT FOR RIVER HEALTH AND RECREATION USE

Decreasing water quality, for example through riverine eutrophication due to agricultural run-off, is a constant threat to the global human population. In the United States, southern Californian droughts are a major threat, promoting water reuse through wastewater treatment plants. Reuse and treatment of municipal wastewater instead of discharge to surface waters augments the water supply of communities reliably, safely, and economically. However, recycling wastewater promotes public concern over contamination and pollution, and therefore requires system engineering and monitoring to provide chemical and pathogen control. Diagnostic monitoring for bacterial and eukaryotic pathogens within wastewater effluent still uses traditional methods such as coliform plate counting (Fu et al., 2010); while still valuable, these approaches could be augmented and contextualized using DNA sequencing, which can broaden detection to include microorganisms and viruses that are not detected by traditional cultivation approaches (Staley and Konopka, 1985; Santos et al., 2009; Leddy et al., 2017). The use of high-throughput DNA sequencing for microbial water quality analysis was recently summarized with respect to potable water reuse (Leddy et al., 2018). Essentially, while knowing the changing abundance of specific pollution markers (e.g., sewage and fecal-associated bacteria) is extremely powerful, being able to characterize the broader microbial and viral community, including the presence of functional metabolic pathways, including resistance and virulence markers, can significantly improve existing efforts by further stratifying water quality estimates and improving the assessment of engineering and treatment innovation performance. This study presents a systematic effort to determine the potential of 16S rRNA amplicon and shotgun metagenomic sequencing to detect the broad-spectrum influence of water treatment innovations on the microbiome of a complex urban waterway, as well as to provide direct comparison against traditional coliform plate count data. This study represents the most comprehensive and longest characterization of the microbiome of an urban waterway yet attempted.

This study concludes that the implementation of state-of-the-art disinfection, and establishment of the TCR in the Calumet WRP service area, have fundamentally altered the microbial composition and dynamics of the CAWS. This change primarily consists of a significant reduction in the proportion of sewage and fecal-associated bacteria in the river system around WRP effluent outfalls. Traditional techniques validated these results; so, what did the molecular approaches add to the coliform counts analysis? For the most part, the use of 16S rRNA amplicon sequencing and shotgun metagenomics has provided more detail. While the traditional analysis evidenced the removal of potential pathogens, the molecular approach has provided categorical identification of those potential pathogens and identified their virulence and resistance pathways. These extra data provide more sophisticated diagnostic monitoring, especially as multi-drug resistant infectious disease transmission has become a serious public health concern. Identifying which bacterial species are being transmitted through the sewage system, and which antibiotics they are resistant to, can provide an early warning system for the spread of such infections in the human population. While the analysis of the water system is likely too far downstream, this proof of principle study supports the application of these approaches further upstream in the wastewater collection system. However, the characterization of exactly which potential pathogens are present in a water body, and how their proportion is linked to physicochemical characteristics and season, could potentially be informative in

management the recreational use of that system, provide a more detailed water quality assessment. To validate this potential, future studies should attempt to combine these sophisticated analyses with epidemiological studies, to deliver accurate risk of infection assessments.

While improving the accuracy and specificity of potential pathogen detection is valuable, the ability to survey a broader range of microbial elements using molecular approaches, provides a compelling advantage over traditional plate counts. This study demonstrated the microorganisms and their functions that associate with reduced sewage pollution, both across the CAWS prior to implementation of disinfection and the TCR, and the specific enrichment of these markers following these implementations. Diagnosis of urban river 'health' on the basis of the abundance of potential pathogens alone misses out on the potential to characterize ways to optimize metrics that define 'health'. For example, just because potential pathogen load has decreased does not mean that a river system is functional under optimal ecological conditions. Many other anthropogenic stressors can impact riverine ecology, and the microbiome acts as a highly responsive sentinel (a 'canary in a coal mine') to these effectors. As such, monitoring the broader microbial community and its functional metabolic potential provides a more detailed assessment of river ecosystem quality, which should be optimized to ensure maximum benefit to urban society. While this study provides a solid baseline against which to identify compelling biomarkers of riverine health, further studies will need to be done to determine how fine-scale manipulation of physicochemical parameters impacts these health metrics. Also, the emergent properties of improved riverine health, such as nutrient recycling, fish stocks, control of insect vectors of disease, will need to be quantified in association with such interventions. However, the potential for using molecular characterization of the microbiome as a whole to create such metrics of health is great and could dramatically transform the management of urban waterways from reactive to proactive management.

The other major innovation is the ability of these molecular approaches to create a more detailed exploration of the origin of different microbial taxa in the CAWS water column. Source tracking of potential pollutants and microbial contaminants can be done accurately and quantitatively using traditional qPCR methods, but this requires prior knowledge of the target organism and the molecular markers that can differentiate it from other, potentially very similar species. Using 16S rRNA amplicon sequencing, with sequences identified to the level of exact-sequence variant (ASVs; 100% nucleotide identify fragments) it was possible to widely characterize the probable source of different organisms in the CAWS data. Using a Bayesian algorithm called SourceTracker we can identify the source of between 60-80% of the thousands of ASVs in a sample, creating a detailed snapshot of the sources which contribute to the diverse microbial community in that sample. Traditional approaches cannot come close to this level of resolution without huge financial investment. While not as accurate as traditional approach, and certainly not quantitative, it does augment such data, contextualizing the potential sources of key organisms, and providing a broader understanding of the origin of specific microbial pollution. Additionally, it may help elucidate situations whereby qPCR approaches cannot differentiate between sources of fecal pollution, for example, mammalian wild animals and human feces. This is helped by access to vast 'source' databases, such as those held by the Earth Microbiome Project. These can help to identify nuanced pollution, such as that associated with migrating birds, or seasonally differentiated runoff waste from urban grasslands and agricultural fields.

In summary, 16S rRNA amplicon and metagenomic sequencing of the microbial communities in the CAWS has provided a detailed characterization that substantially improves our understanding of the complexity of the Chicago river ecosystem, and how it responds to implementation of disinfection and the TCR by the District. By providing better characterization of pathogens and their virulence/resistance profile, expanding the options of stratification of water quality to include 'healthy' microbial biomarkers for optimization, and the increased breadth of molecular source tracking, these approaches augment existing water quality assessment tools. However, more research is needed if we are to be able to implement these tools effectively and reliably to produce actionable evidence to the District.

1.5 SUMMARY AND DISCUSSION

Microbial communities are key players in maintaining the health of the CAWS. Traditional laboratory-culture methods such as fecal bacteria count and PCR-based methods have been extensively used to characterize the CAWS microbial quality for regulatory purposes; however, these methods are limited in their ability to resolve the source of fecal and/or sewage contamination. Additionally, these methods do not completely describe the diversity of microbial communities present in the CAWS. This study, which started in 2013, aimed to better understand the composition, biogeography and sources of the microbial community associated with the CAWS using state-of-the-art 16S rRNA gene amplicon- and metagenome-based sequencing. The study also aims at determining if environmental physicochemical parameter such as flow, rainfall, temperature, water chemistry etc., can describe the distribution of bacteria across the CAWS and through time. In addition to investigating the microbial dynamics between different collection sites during different weather conditions, we have further focused on the impact of the MWRD's improvement efforts. MWRD implemented disinfection systems at Calumet and O'Brien in 2016, as well as beginning operation the Calumet Tunnel and Reservoir Plan (TARP) System's Thornton Composite Reservoir (TCR) at the end of 2015. The TCR in the Calumet WRP service area was constructed to capture the CSO discharge during the rainfall events that otherwise flow into the river system. Therefore, using molecular sequencing we have attempted to interpret the bacterial taxonomic and functional dynamics both before and after disinfection and TARP TCR implementation.

16S rRNA amplicon sequencing was done on 2,706 samples collected from sediment, water, treated effluent, raw sewage and fish samples of the CAWS from 2013 to 2019. Compositional analysis of microbial communities demonstrated distinct distribution patterns across different sampling locations (biogeography) and sample types (river water, sediment, effluent, etc.). The community profiles appeared to be stable (in their diversity and composition) across these sampling years and seasons, except during wet weather events and following the disinfection/TARP implementation. Our analysis also showed that microorganisms associated with final WRP effluent (included human fecal or sewage contamination indicators like *Acinetobacter and Arcobacter)* from secondary treatment can be tracked downstream and typically showed increased abundance in proximity to the secondary treated final effluent location. For both sediment and river water samples collected across the CAWS, a significant decrease ($p < 0.05$) in alpha diversity (number of taxa within a single sample) was observed in 2016 (first year post disinfection) when compared to 2015, however, diversity recovered in 2017

and stayed stable through 2019. Taxa associated with fecal and sewage samples, such as *Acinetobacter, Arcobacter* and *Bacteroides,* reduced significantly in proportion post-disinfection/TARP (i.e., 2016–2019) when compared to the pre-disinfection phase (2013–2015). These results thus highlight the efficiency of disinfection technology that was employed at both Calumet and O'Brien as well as the Calumet TARP system which contained the number of CSOs significantly especially at the Calumet river system. This impact is particularly evident at sites immediate downstream sites of both WRPs (Calumet, Site#76; O'Brien, Site#36), which also showed a significant proportional increase in fresh-water indicators such as *Flavobacterium*, which was compounded year on year post intervention.

When focusing on the difference between wet and dry events, Calumet shows an overall trend of an increased proportion of Bacteroidetes and Firmicutes (*Arcobacter* and *Acinetobacter*) during the wet events (both with and without CSOs), which significantly decreased post-TARP (Gallagher and Wasik, 2019). These taxonomic proportional changes correlated well with quantitative fecal coliform data, while river water associated taxa such as *Synechococcus, Sediminibacterium,* and *Fluviicola* correlated negatively with increasing coliforms.

Microbial source tracking using the molecular data suggested that the majority of microbial diversity in CAWS water samples can be largely attributed to effluent, sewage, CAWS sediment, river water, and fish associated samples, with animal or human associated feces contribution being extremely low. The taxa that contributed to effluent and sewage signals were mostly bacteria that are enriched by the treatment process and sewage infrastructure, and are not considered to be human pathogens.

Shotgun metagenomic analysis was applied to 71 samples, adding taxonomic resolution and functional metabolic pathway reconstruction to the existing 16S rRNA analysis. The metagenomic data recapitulated the trends observed with 16S rRNA amplicon data, for example, demonstrating an increase in *Acinetobacter, Arcobacter,* and *Thiomonas* during wet weather events. The metabolic potential analysis demonstrated that overall functional pathways are dominated by conserved sore metabolic functions which are found in nearly all bacteria, and so overall functional pathway profiles did not significantly change across time or weather. However, 30 virulence associated genes were examined and demonstrated very low abundance (0–0.25%) and little variation over time (2013–2019).

The results from this seven-year long microbiome study have significantly augmented the District's existing analysis, and provided further evidence that the District's improvement efforts such as large scale disinfection technology and the TCR led to significant improvement in water quality, as indexed by the proportion of known pathogens, natural river water biomarkers, sources of microbial pollution, and functional pathways.

1.6 STUDY LIMITATIONS

A key step in understanding microbial community structure, dynamics, and how organisms might influence or be influenced by their surroundings is to classify DNA sequences taxonomically or phylogenetically. To date, most studies of microbial communities in systems ranging from the river systems to the human gut have depended on a single gene, the 16S small subunit ribosomal RNA (rRNA) gene. Massively parallel sequencing methods are increasingly being applied to the characterization of microbial communities based on amplification of this gene and have led to a better appreciation of extant biodiversity; however, the 16S rRNA -based techniques are known to be limited by the short read lengths obtained, limited resolution of the 16S rRNA gene among closely related species, and given the prevalence of horizontal gene transfer and the difficulty inherent in defining bacterial species. In order to overcome these biases, metagenome sequencing approaches i.e., whole genome shotgun (WGS) sequencing are commonly used to describe microbial communities, without the biases inherent to PCR amplification of a single gene. Therefore, in this study, shotgun sequencing was employed to supplement the 16S rRNA gene sequencing results. Whole genome shotgun (WGS) metagenomic approaches provide robust estimates of microbial community composition and diversity without the need to target and amplify a specific gene. Additionally, the assignment of taxonomic origin to microbiome sequences continues to be a hurdle and the confidence with which 16S rRNA gene sequences can be assigned to deep taxonomic levels such as genus can be low. And therefore, by assigning taxonomic origin to metagenomic sequences, we were able to get a more detailed sense of the community structure than by 16S rRNA gene sequencing alone. As long as sufficient reference genomes exist for identification, the metagenome performs well in describing the taxonomic composition of a sample. In addition, metagenomics offers the potential to investigate 16S rRNA gene fragments recovered in metagenomic reads without amplification. Furthermore, shotgun data can explore the metabolic potential of the resident bacteria due to its greater genomic and gene coverage and data output.

1.7 REFERENCES

Amir, Amnon, Daniel McDonald, Jose A. Navas-Molina, Evguenia Kopylova, James T. Morton, Zhenjiang Zech Xu, Eric P. Kightley, et al. 2017. “Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns.” *MSystems* 2 (2): e00191-16. https://doi.org/10.1128/mSystems.00191-16.

Anderson Marti J. 2014. “Permutational Multivariate Analysis of Variance (PERMANOVA).” *Wiley StatsRef: Statistics Reference Online*, Major Reference Works, , April. https://doi.org/10.1002/9781118445112.stat07841.

Bentzon-Tilia, Mikkel, Ina Severin, Lars H. Hansen, and Lasse Riemann. 2015. “Genomics and Ecophysiology of Heterotrophic Nitrogen-Fixing Bacteria Isolated from Estuarine Surface Water.” *MBio* 6 (4): e00929-15. https://doi.org/10.1128/mBio.00929-15.

Bokulich, Nicholas A., Sathish Subramanian, Jeremiah J. Faith, Dirk Gevers, Jeffrey I. Gordon, Rob Knight, David A. Mills, and J. Gregory Caporaso. 2013. "Quality-Filtering Vastly Improves Diversity Estimates from Illumina Amplicon Sequencing." *Nature Methods* 10 (1): 57–59. https://doi.org/10.1038/nmeth.2276.

Bowers, Robert M., Ian B. McCubbin, Anna G. Hallar, and Noah Fierer. 2012. "Seasonal Variability in Airborne Bacterial Communities at a High-Elevation Site." *Atmospheric Environment* 50 (April): 41–49. https://doi.org/10.1016/j.atmosenv.2012.01.005.

Caporaso, J. Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D. Bushman, Elizabeth K. Costello, Noah Fierer, et al. 2010. "QIIME Allows Analysis of High-Throughput Community Sequencing Data." *Nature Methods* 7 (5): 335–36. https://doi.org/10.1038/nmeth.f.303.

Chopyk, Jessica, Sarah Allard, Daniel J. Nasko, Anthony Bui, Emmanuel F. Mongodin, and Amy R. Sapkota. 2018. "Agricultural Freshwater Pond Supports Diverse and Dynamic Bacterial and Viral Populations." *Frontiers in Microbiology* 9. https://doi.org/10.3389/fmicb.2018.00792.

Das, Tapas K. 2001. "Ultraviolet Disinfection Application to a Wastewater Treatment Plant." *Clean Products and Processes* 3 (2): 69–80. https://doi.org/10.1007/s100980100108.

Eraqi, Walaa A., Marwa T. ElRakaiby, Salwa A. Megahed, Noha H. Yousef, Mostafa S. Elshahed, and Aymen S. Yassin. 2018. "The Nile River Microbiome Reveals a Remarkably Stable Community Between Wet and Dry Seasons, and Sampling Sites, in a Large Urban Metropolis (Cairo, Egypt)." *OMICS: A Journal of Integrative Biology* 22 (8): 553–64. https://doi.org/10.1089/omi.2018.0090.

Falk, N., T. Reid, A. Skoyles, A. Grgicak-Mannion, K. Drouillard, and C. G. Weisener. 2019. "Microbial Metatranscriptomic Investigations across Contaminant Gradients of the Detroit River." *Science of The Total Environment* 690 (November): 121–31. https://doi.org/10.1016/j.scitotenv.2019.06.451.

Feng, Bi-Wei, Xiao-Ran Li, Jin-Hui Wang, Zi-Ye Hu, Han Meng, Ling-Yun Xiang, and Zhe-Xue Quan. 2009. "Bacterial Diversity of Water and Sediment in the Changjiang Estuary and Coastal Area of the East China Sea." *FEMS Microbiology Ecology* 70 (2): 236–48. https://doi.org/10.1111/j.1574-6941.2009.00772.x.

Fierer, N. J.W. Leff, B.J. Adams, U. N. Nielsen, S. Thomas Bates, C.L. Lauber, S. Owens, J.A. Gilbert, D.H. Wall, J. G. Caporaso. 2012. Cross-biome metagenomic analyses of soil microbes. Proceedings of the National Academy of Sciences 109 (52) 21390-21395; DOI: 10.1073/pnas.1215210110

Fisher, Jenny C., A. Murat Eren, Hyatt C. Green, Orin C. Shanks, Hilary G. Morrison, Joseph H. Vineis, Mitchell L. Sogin, and Sandra L. McLellan. 2015. "Comparison of Sewage and Animal Fecal Microbiomes by Using Oligotyping Reveals Potential Human Fecal Indicators in Multiple Taxonomic Groups." *Applied and Environmental Microbiology* 81 (20): 7023–33. https://doi.org/10.1128/AEM.01524-15.

Fisher, Jenny C., Arturo Levican, María J. Figueras, and Sandra L. McLellan. 2014. "Population Dynamics and Ecology of Arcobacter in Sewage." *Frontiers in Microbiology* 5 (November). https://doi.org/10.3389/fmicb.2014.00525.

Flores Ribeiro, Angela, Josselin Bodilis, Lise Alonso, Sylvaine Buquet, Marc Feuilloley, Jean-Paul Dupont, and Barbara Pawlak. 2014. "Occurrence of Multi-Antibiotic Resistant Pseudomonas Spp. in Drinking Water Produced from Karstic Hydrosystems." *The Science of the Total Environment* 490 (August): 370–78. https://doi.org/10.1016/j.scitotenv.2014.05.012.

Franke, Thomas, and Uwe Deppenmeier. 2018. "Physiology and Central Carbon Metabolism of the Gut Bacterium Prevotella Copri." *Molecular Microbiology* 109 (4): 528–40. https://doi.org/10.1111/mmi.14058.

Franzosa, Eric A., Lauren J. McIver, Gholamali Rahnavard, Luke R. Thompson, Melanie Schirmer, George Weingart, Karen Schwarzberg Lipson, et al. 2018. "Species-Level Functional Profiling of Metagenomes and Metatranscriptomes." *Nature Methods* 15 (11): 962. https://doi.org/10.1038/s41592-018-0176-y.

Gallagher, D. and J. Wasik. 2019 Post Construction Monitoring Report for the Calumet Tunnel and Reservoir Plan System. Monitoring and research Department Draft Report. Metropolitan WRP District of Greater Chicago. Pending IEPA review and publication.

García-Bayona, Leonor, and Laurie E. Comstock. 2019. "Streamlined Genetic Manipulation of Diverse Bacteroides and Parabacteroides Isolates from the Human Gut Microbiota." *MBio* 10 (4): e01762-19. https://doi.org/10.1128/mBio.01762-19.

Gibbons SM, Jones E, Bearquiver A, Blackwolf F, Roundstone W, Scott N, Hooker J, Madsen R, Coleman ML, Gilbert JA. (2014) Human and Environmental Impacts on River Sediment Microbial Communities. PLoS ONE 9(5): e97435. doi:10.1371/journal.pone.0097435

Heinrich, Friederike, Alexander Eiler, and Stefan Bertilsson. 2013. "Seasonality and Environmental Control of Freshwater SAR11 (LD12) in a Temperate Lake (Lake Erken, Sweden)." *Aquatic Microbial Ecology* 70 (1): 33–44.

Hoshino, T., T. Terahara, S. Tsuneda, A. Hirata, and Y. Inamori. 2005. "Molecular Analysis of Microbial Population Transition Associated with the Start of Denitrification in a Wastewater Treatment Process." *Journal of Applied Microbiology* 99 (5): 1165–75. https://doi.org/10.1111/j.1365-2672.2005.02698.x.

Hu, Anyi, Xiaoyong Yang, Nengwang Chen, Liyuan Hou, Ying Ma, and Chang-Ping Yu. 2014. "Response of Bacterial Communities to Environmental Changes in a Mesoscale Subtropical Watershed, Southeast China." *The Science of the Total Environment* 472 (February): 746–56. https://doi.org/10.1016/j.scitotenv.2013.11.097.

Huber, Bettina, Bastian Herzog, Jörg E. Drewes, Konrad Koch, and Elisabeth Müller. 2016. "Characterization of Sulfur Oxidizing Bacteria Related to Biogenic Sulfuric Acid Corrosion in Sludge Digesters." *BMC Microbiology* 16 (July). https://doi.org/10.1186/s12866-016-0767-7.

Huttenhower, C., Gevers, D., Knight, R. et al. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486, 207–214. https://doi.org/10.1038/nature11234

Jackson, Colin R., Justin J. Millar, Jason T. Payne, and Clifford A. Ochs. 2014. "Free-Living and Particle-Associated Bacterioplankton in Large Rivers of the Mississippi River Basin Demonstrate Biogeographic Patterns." *Applied and Environmental Microbiology* 80 (23): 7186–95. https://doi.org/10.1128/AEM.01844-14.

Jezberová, Jitka, Jan Jezbera, Petr Znachor, Jiří Nedoma, Vojtěch Kasalický, and Karel Šimek. 2017. "The Limnohabitans Genus Harbors Generalistic and Opportunistic Subtypes: Evidence from Spatiotemporal Succession in a Canyon-Shaped Reservoir." *Applied and Environmental Microbiology* 83 (21). https://doi.org/10.1128/AEM.01530-17.

Jiang, Hongchen, Hailiang Dong, Gengxin Zhang, Bingsong Yu, Leah R. Chapman, and Matthew W. Fields. 2006. "Microbial Diversity in Water and Sediment of Lake Chaka, an Athalassohaline Lake in Northwestern China." *Applied and Environmental Microbiology* 72 (6): 3832–45. https://doi.org/10.1128/AEM.02869-05.

Kamika, Ilunga, Shohreh Azizi, and Memory Tekere. 2018. "Comparing Bacterial Diversity in Two Full-Scale Enhanced Biological Phosphate Removal Reactors Using 16S Amplicon Pyrosequencing." *Polish Journal of Environmental Studies* 27 (2): 709–45. https://doi.org/10.15244/pjoes/69029.

Kasalický, Vojtěch, Jan Jezbera, Martin W. Hahn, and Karel Šimek. 2013. "The Diversity of the Limnohabitans Genus, an Important Group of Freshwater Bacterioplankton, by Characterization of 35 Isolated Strains." *PLoS ONE* 8 (3). https://doi.org/10.1371/journal.pone.0058209.

Kim, Michael H., Oliver J. Hao, and Nam S. Wang. 1997. "Acinetobacter Isolates from Different Activated Sludge Processes: Characteristics and Neural Network Identification." *FEMS Microbiology Ecology* 23 (3): 217–27. https://doi.org/10.1111/j.1574-6941.1997.tb00404.x.

Klai, Nouha, Song Yan, Rajeshwar Dayal Tyagi, and Rao Y. Surampalli. 2015. "EPS Producing Microorganisms from Municipal Wastewater Activated Sludge." *Journal of Petroleum & Environmental Biotechnology* 7: 1000255.

Knights, Dan, Justin Kuczynski, Emily S. Charlson, Jesse Zaneveld, Michael C. Mozer, Ronald G. Collman, Frederic D. Bushman, Rob Knight, and Scott T. Kelley. 2011. "Bayesian Community-Wide Culture-Independent Microbial Source Tracking." *Nature Methods* 8 (9): 761–63. https://doi.org/10.1038/nmeth.1650.

Lee, Jangho, Banghyo Park, Sung-Geun Woo, Juyoun Lee, and Joonhong Park. 2014. "Prosthecobacter Algae Sp. Nov., Isolated from Activated Sludge Using Algal Metabolites." *International Journal of Systematic and Evolutionary Microbiology* 64 (Pt 2): 663–67. https://doi.org/10.1099/ijs.0.052787-0.

Lee, Philip O., Sandra L. McLellan, Linda E. Graham, and Erica B. Young. 2015. "Invasive Dreissenid Mussels and Benthic Algae in Lake Michigan: Characterizing Effects on Sediment Bacterial Communities." *FEMS Microbiology Ecology* 91 (1): 1–12. https://doi.org/10.1093/femsec/fiu001.

Leight, Andrew K., Byron C. Crump, and Raleigh R. Hood. 2018. "Assessment of Fecal Indicator Bacteria and Potential Pathogen Co-Occurrence at a Shellfish Growing Area." *Frontiers in Microbiology* 9 (March). https://doi.org/10.3389/fmicb.2018.00384.

Liu, Jiwen, Bingbing Fu, Hongmei Yang, Meixun Zhao, Biyan He, and Xiao-Hua Zhang. 2015. "Phylogenetic Shifts of Bacterioplankton Community Composition along the Pearl Estuary: The Potential Impact of Hypoxia and Nutrients." *Frontiers in Microbiology* 6. https://doi.org/10.3389/fmicb.2015.00064.

Liu, Yan, Tong Zhang, and Herbert H. P. Fang. 2005. "Microbial Community Analysis and Performance of a Phosphate-Removing Activated Sludge." *Bioresource Technology* 96 (11): 1205–14. https://doi.org/10.1016/j.biortech.2004.11.003.

Lozupone, Catherine, Manuel E Lladser, Dan Knights, Jesse Stombaugh, and Rob Knight. 2011. "UniFrac: An Effective Distance Metric for Microbial Community Comparison." *The ISME Journal* 5 (2): 169–72. https://doi.org/10.1038/ismej.2010.133.

Ma, Tao, Qian Chen, Mengyao Gui, Can Li, and Jinren Ni. 2016. "Simultaneous Denitrification and Phosphorus Removal by Agrobacterium Sp. LAD9 under Varying Oxygen Concentration." *Applied Microbiology and Biotechnology* 100 (7): 3337–46. https://doi.org/10.1007/s00253-015-7217-6.

Mandal, Siddhartha, Will Van Treuren, Richard A. White, Merete Eggesbø, Rob Knight, and Shyamal D. Peddada. 2015. "Analysis of Composition of Microbiomes: A Novel Method for Studying Microbial Composition." *Microbial Ecology in Health and Disease* 26 (May). https://doi.org/10.3402/mehd.v26.27663.

Marotz, Clarisse, Amnon Amir, Greg Humphrey, James Gaffney, Grant Gogul, and Rob Knight. 2017. "DNA Extraction for Streamlined Metagenomics of Diverse Environmental Samples." *BioTechniques* 62 (6): 290–93. https://doi.org/10.2144/000114559.

McLain, Jean E. T., Channah M. Rock, Kathleen Lohse, and James Walworth. 2011. "False-Positive Identification of Escherichia Coli in Treated Municipal Wastewater and Wastewater-Irrigated Soils." *Canadian Journal of Microbiology* 57 (10): 775–84. https://doi.org/10.1139/w11-070.

McLellan, Sandra L., Jenny C. Fisher, and Ryan J. Newton. 2015. "The Microbiome of Urban Waters." *International Microbiology: The Official Journal of the Spanish Society for Microbiology* 18 (3): 141–49. https://doi.org/10.2436/20.1501.01.244.

McLellan, S.L., S.M. Huse, S.R. Mueller-Spitz, E.N. Andreishcheva, and M.L. Sogin. 2010. "Diversity and Population Structure of Sewage Derived Microorganisms in Wastewater Treatment Plant Influent." *Environmental Microbiology* 12 (2): 378–92. https://doi.org/10.1111/j.1462-2920.2009.02075.x.

Minich, Jeremiah J., Greg Humphrey, Rodolfo A. S. Benitez, Jon Sanders, Austin Swafford, Eric E. Allen, and Rob Knight. 2018. "High-Throughput Miniaturized 16S RRNA Amplicon Library Preparation Reduces Costs While Preserving Microbiome Integrity." *MSystems* 3 (6). https://doi.org/10.1128/mSystems.00166-18.

Moon, Kira, Ilnam Kang, Suhyun Kim, Sang-Jong Kim, and Jang-Cheon Cho. 2018. "Genomic and Ecological Study of Two Distinctive Freshwater Bacteriophages Infecting a Comamonadaceae Bacterium." *Scientific Reports* 8 (1): 1–9. https://doi.org/10.1038/s41598-018-26363-y.

Mougin, Benjamin, Roger B. D. Tian, and Michel Drancourt. 2015. "Tropical Plant Extracts Modulating the Growth of Mycobacterium Ulcerans." *PLOS ONE* 10 (4): e0124626. https://doi.org/10.1371/journal.pone.0124626.

Mustakhimov, Ildar, Marina G. Kalyuzhnaya, Mary E. Lidstrom, and Ludmila Chistoserdova. 2013. "Insights into Denitrification in Methylotenera Mobilis from Denitrification Pathway and Methanol Metabolism Mutants." *Journal of Bacteriology* 195 (10): 2207–11. https://doi.org/10.1128/JB.00069-13.

Nascimento, Altina Lacerda, Adijailton Jose Souza, Pedro Avelino Maia Andrade, Fernando Dini Andreote, Aline Renée Coscione, Fernando Carvalho Oliveira, and Jussara Borges Regitano. 2018. "Sewage Sludge Microbial Structures and Relations to Their Sources, Treatments, and Chemical Attributes." *Frontiers in Microbiology* 9. https://doi.org/10.3389/fmicb.2018.01462.

Newton, Ryan J., Sandra L. McLellan, Deborah K. Dila, Joseph H. Vineis, Hilary G. Morrison, A. Murat Eren, and Mitchell L. Sogin. 2015. "Sewage Reflects the Microbiomes of Human Populations." *MBio* 6 (2). https://doi.org/10.1128/mBio.02574-14.

Payne, Jason T., Justin J. Millar, Colin R. Jackson, and Clifford A. Ochs. 2017. "Patterns of Variation in Diversity of the Mississippi River Microbiome over 1,300 Kilometers." *PLOS ONE* 12 (3): e0174890. https://doi.org/10.1371/journal.pone.0174890.

Pohlner, Marion, Leon Dlugosch, Bernd Wemheuer, Heath Mills, Bert Engelen, and Brandi Kiel Reese. 2019. "The Majority of Active Rhodobacteraceae in Marine Sediments Belong to Uncultured Genera: A Molecular Approach to Link Their Distribution to Environmental Conditions." *Frontiers in Microbiology* 10. https://doi.org/10.3389/fmicb.2019.00659.

Read, Daniel S., Hyun S. Gweon, Michael J. Bowes, Lindsay K. Newbold, Dawn Field, Mark J. Bailey, and Robert I. Griffiths. 2015. "Catchment-Scale Biogeography of Riverine Bacterioplankton." *The ISME Journal* 9 (2): 516–26. https://doi.org/10.1038/ismej.2014.166.

Sangwan N, Zarraonaindia I, Hampton-Marcel JT, Ssegane H, Eshoo TW, Rijal G, Negri MC, Gilbert JA. 2016. Differential Functional Constraints Cause Strain-Level Endemism in Polynucleobacter populations. mSystems. DOI: 10.1128/mSystems.00003-16

Saunders, Aaron M., Mads Albertsen, Jes Vollertsen, and Per H. Nielsen. 2016. "The Activated Sludge Ecosystem Contains a Core Community of Abundant Organisms." *The ISME Journal* 10 (1): 11–20. https://doi.org/10.1038/ismej.2015.117.

Savio, Domenico, Lucas Sinclair, Umer Z. Ijaz, Juraj Parajka, Georg H. Reischer, Philipp Stadler, Alfred P. Blaschke, et al. 2015. "Bacterial Diversity along a 2600 Km River Continuum." *Environmental Microbiology* 17 (12): 4994–5007. https://doi.org/10.1111/1462-2920.12886.

Segata, Nicola, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, and Curtis Huttenhower. 2012. "Metagenomic Microbial Community Profiling Using Unique Clade-Specific Marker Genes." *Nature Methods* 9 (8): 811–14. https://doi.org/10.1038/nmeth.2066.

Shchegolkova, Nataliya M., George S. Krasnov, Anastasia A. Belova, Alexey A. Dmitriev, Sergey L. Kharitonov, Kseniya M. Klimina, Nataliya V. Melnikova, and Anna V. Kudryavtseva. 2016. "Microbial Community Structure of Activated Sludge in Treatment Plants with Different Wastewater Compositions." Frontiers in Microbiology 7 (February). https://doi.org/10.3389/fmicb.2016.00090.

Shi, Peili, Yuxiu Zhang, Zhenqi Hu, Kang Ma, Hao Wang, and Tuanyao Chai. 2017. "The Response of Soil Bacterial Communities to Mining Subsidence in the West China Aeolian Sand Area." https://pubag.nal.usda.gov/catalog/5825226.

Silva-Bedoya, Lina Marcela, María Solange Sánchez-Pinzón, Gloria Ester Cadavid-Restrepo, and Claudia Ximena Moreno-Herrera. 2016. "Bacterial Community Analysis of an Industrial Wastewater Treatment Plant in Colombia with Screening for Lipid-Degrading Microorganisms." *Microbiological Research* 192 (November): 313–25. https://doi.org/10.1016/j.micres.2016.08.006.

Sommers, Pacifica, John L. Darcy, Dorota L. Porazinska, Eli M. S. Gendron, Andrew G. Fountain, Felix Zamora, Kim Vincent, et al. 2019. "Comparison of Microbial Communities in the Sediments and Water Columns of Frozen Cryoconite Holes in the McMurdo Dry Valleys, Antarctica." *Frontiers in Microbiology* 10. https://doi.org/10.3389/fmicb.2019.00065.

Sun, H. Y., Noe, J., Barber, J., Coyne, R. S. , Cassidy-Hanley, D., Clark, T. G., Findly, R. C. and Dickerson, H. W. 2009. Endosymbiotic Bacteria in the Parasitic Ciliate Ichthyophthirius multifiliis. Applied and Environmental Microbiology, Vol. 75, p. 7445-7452.

Tao, Yongzhen, Deng Liu, Xing Yan, Zhihua Zhou, Jeong K. Lee, and Chen Yang. 2012. "Network Identification and Flux Quantification of Glucose Metabolism in Rhodobacter Sphaeroides under Photoheterotrophic H2-Producing Conditions." *Journal of Bacteriology* 194 (2): 274–83. https://doi.org/10.1128/JB.05624-11.

Tejedor-Sanz, Sara, Patricia Fernández-Labrador, Steven Hart, Cesar I. Torres, and Abraham Esteve-Núñez. 2018. "Geobacter Dominates the Inner Layers of a Stratified Biofilm on a Fluidized Anode During Brewery Wastewater Treatment." *Frontiers in Microbiology* 9 (March). https://doi.org/10.3389/fmicb.2018.00378.

Thompson, Luke R., Jon G. Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J. Locey, Robert J. Prill, et al. 2017. "A Communal Catalogue Reveals Earth's Multiscale Microbial Diversity." *Nature* 551 (7681): 457–63. https://doi.org/10.1038/nature24621.

Tully, B., Graham, E. & Heidelberg, J. 2018. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data 5, 170203. https://doi.org/10.1038/sdata.2017.203

Van Rossum, Thea, Michael A. Peabody, Miguel I. Uyaguari-Diaz, Kirby I. Cronin, Michael Chan, Jared R. Slobodan, Matthew J. Nesbitt, et al. 2015. "Year-Long Metagenomic Study of River Microbiomes Across Land Use and Water Quality." *Frontiers in Microbiology* 6 (December). https://doi.org/10.3389/fmicb.2015.01405.

VandeWalle, J. L., G.W. Goetz, S.M. Huse, H. G. Morrison, M.L. Sogin, R.G. Hoffmann, K. Yan, and S.L. McLellan. 2012. "Acinetobacter, Aeromonas, and Trichococcus Populations Dominate the Microbial Community within Urban Sewer Infrastructure." *Environmental Microbiology* 14 (9): 2538–52. https://doi.org/10.1111/j.1462-2920.2012.02757.x.

Vieweg, Michael, Marie J. Kurz, Nico Trauth, Jan H. Fleckenstein, Andreas Musolff, and Christian Schmidt. 2016. "Estimating Time-Variable Aerobic Respiration in the Streambed by Combining Electrical Conductivity and Dissolved Oxygen Time Series." *Journal of Geophysical Research: Biogeosciences* 121 (8): 2199–2215. https://doi.org/10.1002/2016JG003345.

Vörösmarty, C. J., P. B. McIntyre, M. O. Gessner, D. Dudgeon, A. Prusevich, P. Green, S. Glidden, et al. 2010. "Global Threats to Human Water Security and River Biodiversity." *Nature* 467 (7315): 555–61. https://doi.org/10.1038/nature09440.

Wang, Yongming, Jun Yang, Lemian Liu, and Zheng Yu. 2015. "Quantifying the Effects of Geographical and Environmental Factors on Distribution of Stream Bacterioplankton within Nature Reserves of Fujian, China." *Environmental Science and Pollution Research International* 22 (14): 11010–21. https://doi.org/10.1007/s11356-015-4308-y.

Wang, Yu, Hua-Fang Sheng, Yan He, Jin-Ya Wu, Yun-Xia Jiang, Nora Fung-Yee Tam, and Hong-Wei Zhou. 2012. "Comparison of the Levels of Bacterial Diversity in Freshwater, Intertidal Wetland, and Marine Sediments by Using Millions of Illumina Tags." *Applied and Environmental Microbiology* 78 (23): 8264–71. https://doi.org/10.1128/AEM.01821-12.

Wright, J., V. Kirchner, W. Bernard, N. Ulrich, C. McLimans, M. F. Campa, T. Hazen, et al. 2017. "Bacterial Community Dynamics in Dichloromethane-Contaminated Groundwater Undergoing Natural Attenuation." *Frontiers in Microbiology* 8: 2300–2300. https://doi.org/10.3389/fmicb.2017.02300.

Zeglin, Lydia H. 2015. "Stream Microbial Diversity in Response to Environmental Changes: Review and Synthesis of Existing Research." *Frontiers in Microbiology* 6: 454. https://doi.org/10.3389/fmicb.2015.00454.

Zhang, Qian, Xia He, and Tao Yan. 2015. "Differential Decay of Wastewater Bacteria and Change of Microbial Communities in Beach Sand and Seawater Microcosms." *Environmental Science & Technology* 49 (14): 8531–40. https://doi.org/10.1021/acs.est.5b01879.

Zhou, Aijuan, Jiaguang Zhang, Kaili Wen, Zhihong Liu, Guoying Wang, Wenzong Liu, Aijie Wang, and Xiuping Yue. 2016. "What Could the Entire Cornstover Contribute to the Enhancement of Waste Activated Sludge Acidification? Performance Assessment and Microbial Community Analysis." *Biotechnology for Biofuels* 9 (November). https://doi.org/10.1186/s13068-016-0659-y.

Zimmermann, Rosalie E., Olivier Ribolzi, Alain Pierret, Sayaphet Rattanavong, Matthew T. Robinson, Paul N. Newton, Viengmon Davong, Yves Auda, Jakob Zopfi, and David A. B. Dance. 2018. "Rivers as Carriers and Potential Sentinels for Burkholderia Pseudomallei in Laos." *Scientific Reports* 8 (1): 1–7. https://doi.org/10.1038/s41598-018-26684-y.

2 CHICAGO AREA WATERWAYS SYSTEM FECAL INDICATOR BACTERIA MODEL DEVELOPMENT

2.1 INTRODUCTION

The main objective of this task is to explore the applicability of using a data-driven modeling platform for predicting fecal indicator bacteria concentrations in the CAWS. We explored two approaches 1) classical statistical methods and 2) the Chicago Area Waterways System Fecal Indicator Bacteria (CAWS-FIB) model which is designed to predict the FIB concentrations at any point along the CAWS using machine learning (ML, the subfield of computer science that allows computers to learn without being explicitly programmed) (Samuel, 1959). ML is suited for predicting response variables (e.g., FIBs) that require high dimensional/multi-feature predictor variables and commonly used in cases where patterns exist between response and predictor variables, but functional relationships are very difficult to pin down mathematically. ML algorithms including artificial neural networks (ANNs) and gradient boosting machine (GBM) and classical statistical approaches such as multiple linear regress with adaptive least absolute shrinkage and selection operator (MLR-AL) and partial least squares regression (PLSR) were explored and compared using water quality and other relevant data associated with seven sampling sites (36, 56, 57, 73, 76, 99, and 100) along the CAWS. Testing data for each of these sites were divided into two groups: 2013–2015 (pre-disinfection/pre-TARP period) and 2016–2018 (post-disinfection/TARP implementation period). This task was achieved through 1) data management and streamlining from multiple sources including the Metropolitan WRP District (MWRD) of Greater Chicago, Illinois State Water Survey (ISWS), U.S. Geological Survey (USGS), and NOAA - National Centers for Environmental Information (NCEI) and 2) implementation of each of the aforementioned ML and traditional statistical algorithms in Python, an open-source and widely used software package for data science and machine learning applications.

As will be discussed, there was not enough data to train the models. However, the model architecture has been completed and can be used for future application if fecal coliform sampling frequency increases. The details of the methods, results of the model training and testing performance evaluation, model functionalities, and the applicability of the data-driven modeling platform for predicting FIB density in the CAWS will be discussed in the following sections.

2.2 MATERIALS AND METHODS

2.2.1 Water Quality Sampling Sites Used for Model Development

Seven sites (36, 56, 57, 73, 76, 99, and 100) were used for training and testing the model both in the pre-disinfection/pre-TARP (2013–2015) and post-disinfection/TARP implementation (2016–2018) periods. These seven sites, in addition to having relatively more data, are located in different sections of the CAWS and represent both contact and non-contact uses (Figure 36). Thus, they were chosen to be used for training and testing each of the candidate models in the

hope of developing a robust data-driven model that can be used to predict FIB density at any point along the CAWS given a suite of important predictors.

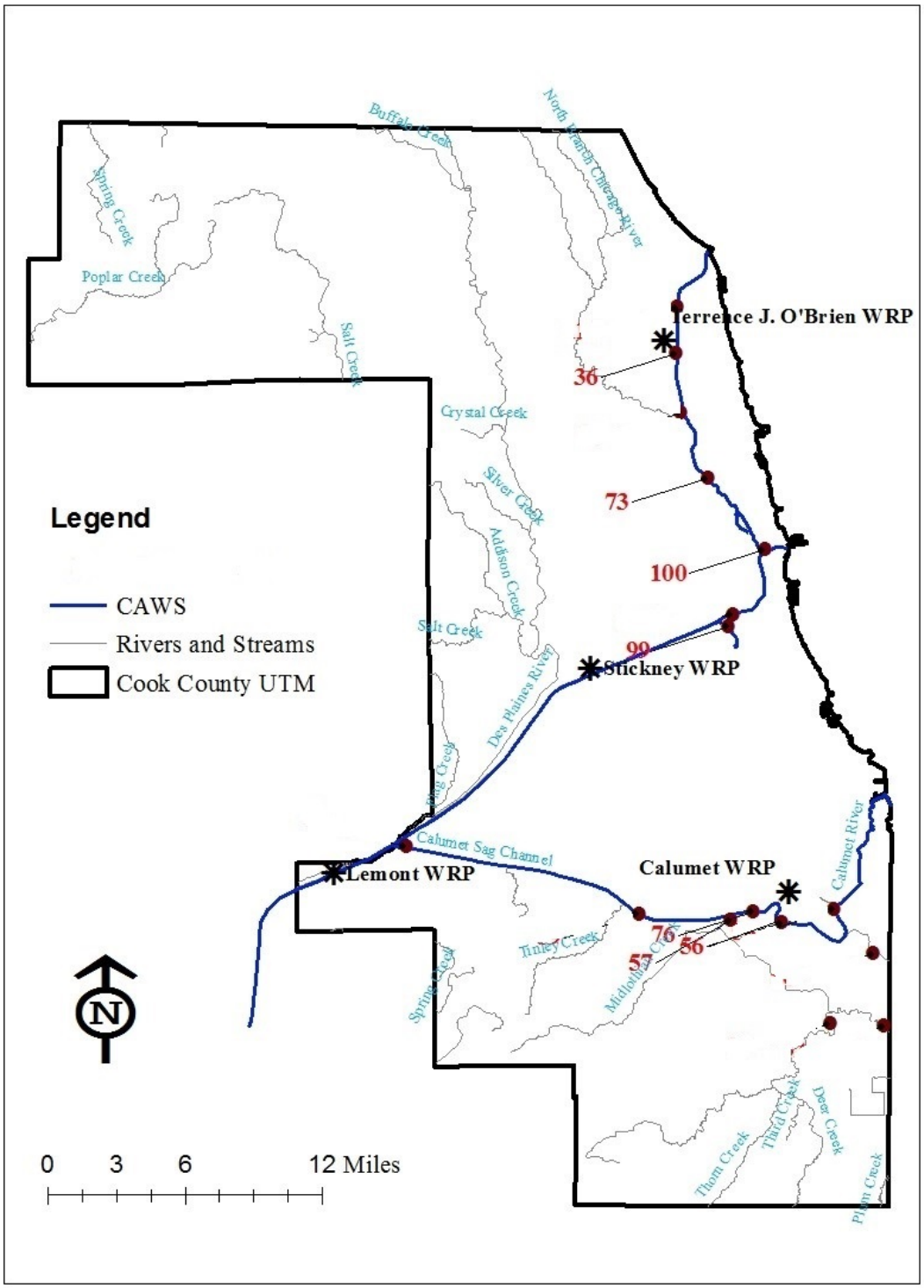


FIGURE 36 The water quality sampling sites where fecal coliform data were analyzed. Sites 36, 56, 57, 73, 76, 99, and 100 were used for model development.

2.2.2 The CAWS-FIB Conceptual Modeling Framework

The main components of the CAWS-FIB model and their connections are shown in Figure 37. The first component compiles (and where applicable, transforms) a variety of relevant input data from multiple sources. The second component is focused on the highly iterative process of developing a data-driven (ML or traditional statistics-based) model that best describes the underlying complex mathematical relationship between FIB densities and environmental variables. The third component summarizes predicted FIB densities (fecal coliform in this case) in the water column at a specified location within the CAWS, probability of exceedance based on certain threshold determined by U.S. Environmental Protection Agency (USEPA) regulatory limit and a decision value, model performance metrics, and list of explanatory variables ranked in order of importance.

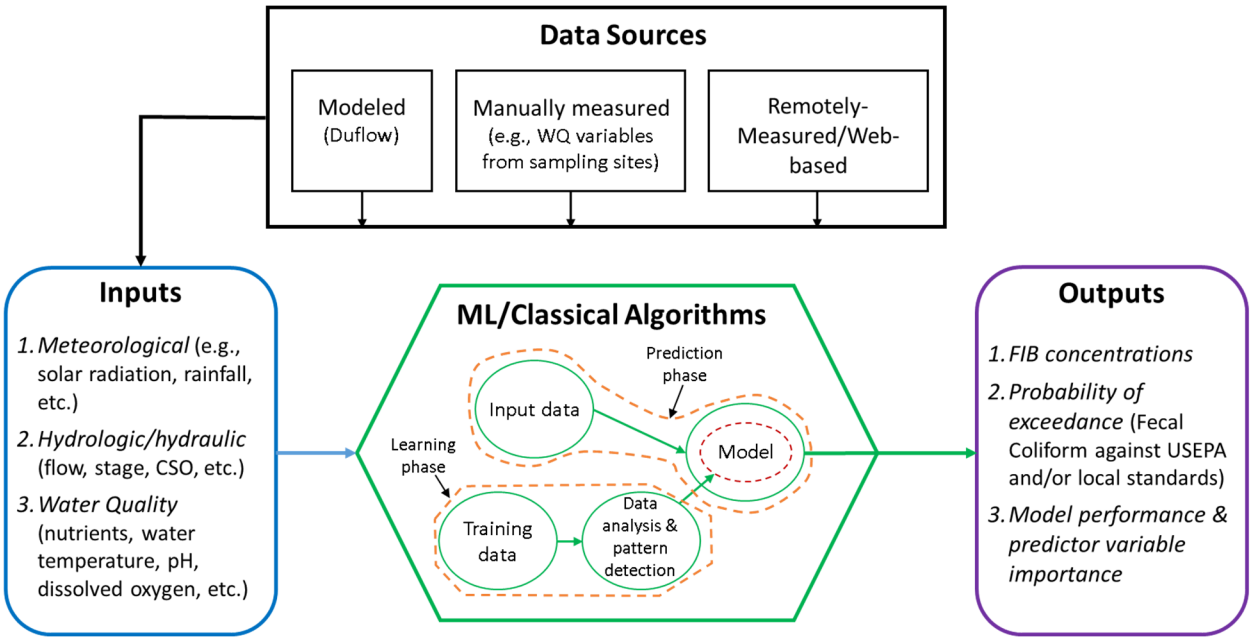


FIGURE 37 Schematic of the CAWS-FIB modeling process.

2.2.2.1 Data and Data Sources

Data for the CAWS-FIB model include daily FIB concentrations and multiple environmental variables (Figure 37) from various sources including the MWRD of Greater Chicago, ISWS, USGS, and NOAA-NCEI. Environmental variables included three major categories: meteorological (e.g., solar radiation, precipitation, etc.), hydrologic and hydraulic (e.g., flow, stage, combined sewer overflows, etc.), and water quality (e.g., pH and concentrations of nutrients) data. The model can take environmental variables that come from frequent (hourly to daily) and one-time manual measurements. Manually-measured environmental variables include those that were measured (e.g., pH, dissolved oxygen

concentration, water temperature, etc.) at the sampling points during times when river water samples for FIB measurements were collected.

High frequency environmental variables were summarized over 1-, 2-, 6-, 12-, 24-, 48-, 72-, 96-, and 120-hr. time windows or lagged times, a technique used by previous studies (e.g., (Jones *et al.*, 2013, Brooks *et al.*, 2016) and shown to improve the accuracy of regression models in predicting FIB levels (Cyterski *et al.*, 2012). Summary statistics over the chosen time windows or lagged times included min, max, mean, range, sum, and standard deviation. The choice of which statistics to apply for an environmental variable was based on the insights from related studies (e.g., Jones et al.,2013; Brooks et al., 2016) and knowledge of the CAWS ecosystems. For instance, combined sewer overflows (CSO), which contain about 90% stormwater and 10% untreated sewage, entering the CAWs is considered a major source of micro-pollutants. Table B.1 shows the list of the environmental variables and the corresponding summary statistics used at each of the sampling points (Sites 36, 56, 57, 73, 76, 99, and 100) over the indicated time windows. Widely used data transformation techniques including logarithmic and square root transformations were applied to the data as necessary. Determining which transformation technique is appropriate for an environmental variable was based on the works of Ge & Frick, (2007) and Frick *et al.* (2008). R and Python scripts were developed to download, pre-process, and summarize datasets for each site from 2013–2018.

2.2.3 Model Development

2.2.3.1 Overview of the Algorithms

The CAWS-FIB model was developed and implemented in Python 3.7.3 using mostly the scikit-learn package (Pedregosa *et al.*, 2011). Each of the algorithms being explored is described in the following subsections.

ANNs. An artificial neural network (ANN) is a popular ML algorithm designed to mimic the functionality of a human neurological system. ANNs have been applied to numerous domains including business, health and medicine, manufacturing, and engineering (Paliwal and Kumar, 2009). Vijayashanthar et al. (2018) summarized the multitude of works that have been done in the applications of ANNs to water resources and environmental engineering ranging from rainfall-runoff modeling/forecasting to drinking water quality. Brion and Lingireddy (2003) have shown that ANNs can be used for predicting peak microbial concentrations, sorting land use associated fecal pollution sources and relative ages of runoff, and selecting and studying surrogate parameters. The ANNs used in the CAWS-FIB model are multilayer-perceptron networks with one hidden layer consisting of 100 neurons. The stochastic gradient method is used to minimize the least squares function.

GBM. The Gradient Boosting Machine method (GBM method (Friedman, 2001) used for the CAWS-FIB model belongs to the ensemble group of ML algorithms and is a variant of the random forests method (Breiman, 2001). GBM has been shown to perform well in predicting recreational water quality advisories. In a comparison of 14 regression and machine learning

methods, GBM was identified as the most accurate method for predicting FIBs (Brooks *et al.*, 2016). GBM uses decision or regression trees rather than linear equations (Friedman, 2001). Each decision or regression tree is composed of virtual branches and nodes and controlled by a set of decision rules. For instance, "if pH is greater than 7.0 go to the left branch, otherwise go right." At the end of any branch is a "node," which contains a predictive value for the response variable. Under GBM, each regression tree is called a weak or base-learner. The ensemble of base-learners are constructed sequentially to improve the performance of the model by fitting the subsequent regression trees to the residual error after the previous trees have all been fit (Cyterski et al., 2013; Natekin & Knoll, 2013). Main strengths of GBM are its robustness against overfitting of the training data and ability to handle non-linear relationships between the response and explanatory variables, but its drawback lies in its nature as being of a "black box," i.e., the model is difficult to inspect graphically or pin down mathematically (Cyterski et al., 2013). Detailed discussion of the GBM algorithm can be found in Friedman (2001), Hastie et al. (2001) and Natekin & Knoll (2013). During model development, the number of regression trees or base learners used ranged from 7000 and 10000 (consistent with similar past research (e.g., Jones et al., 2013; Cyterski et al., 2013)), while the model learning rate and loss function were set to 0.01 and ls (least squares), respectively. The final choice of the number of regression trees was based on the trade-off between predictive accuracy and computational time requirement.

MLR-AL. In addition to ANNs and GBM, multiple linear regression (MLR) method was evaluated for its potential to model FIB concentration in the CAWS. As the most common method of linear analysis, MLR describes the relationship between a continuous response or dependent variable and two or more explanatory or independent variables. To develop an MLR-based model for this study, the adaptive least absolute shrinkage and selection operator (adaptive LASSO) (Zou, 2006) was used for selecting explanatory variables and estimating their coefficients. The LASSO is a regression analysis method that can simultaneously perform relevant variable selection and estimation of coefficients of the variable being selected (Tibshirani, 1996). The adaptive LASSO is an improvement of the LASSO and capable of correctly selecting the right explanatory variables as well as accurately estimating them due to its so-called "oracle properties" (Zuo, 2006). Brooks et al. (2016) showed that an MLR method with adaptive LASSO is one of the most accurate methods in predicting recreational water quality advisories and their work was the primary basis in developing an MLR-based model for CAWS-FIB model.

PLSR. Another linear analysis method that was tested is the partial least squares regression (PLSR) (Hou et al., 2006; Brooks et al., 2013). As a regression technique, PLSR improves a major limitation of MLR using ordinary least squares, which is overfitting in the presence of collinear or correlated explanatory variables. It inherently takes into account the effect of collinearity in explanatory variable selection. Overfitting occurs when a model works very well in fitting with historical data, but performs poorly in predicting future values. PLSR is suited for constructing predictive regression model that involves many explanatory variables (Wold et al., 2011). An important feature of PLSR method is its use of mutually orthogonal components derived from the decomposition of the explanatory variables as covariates in a regression model. These orthogonal components are chosen in such a way that they are related to the response variable. Choosing how many components to include in the model is a requirement

in using the PLSR. The predictive residual sum of squares (PRESS) statistic was used to select the number of model components following Brooks et al. (2013). PLSR is one of the algorithms used in the VirtualBeach model version 3, a model developed by the US EPA for predicting fecal indicator bacteria concentration on recreational beaches (Cyterski et al., 2013).

2.2.3.2 Model Training and Testing

The first step in the model estimation was to subdivide the preprocessed (cleansed/transformed) data into two sets: training and testing with 85%-15% split. For each sampling site, the training set (85% of the entire dataset) was used for determining or learning the model parameters and assessing the initial model performance, while the testing set (the remaining 15% of the dataset not used in model training) was used to quantify a final, unbiased estimate of the predictive performance of the model. The training and testing sequence, being an iterative process, was conducted several times to get the estimate of the model's true error rate.

The main focus of the training phase is to avoid overfitting. Overfitting occurs when the approximated function or model can only define the relationship of the explanatory variables and a response variable on a particular set of data (e.g., training dataset). In other words, the model "memorizes" the specific relationship between the explanatory (e.g., environmental) variables and the response variable (fecal coliform) of the training dataset only, instead of the underlying general structure representing the entire environmental and fecal coliform variable space or distribution. As a result, the approximated function performs very well during the training phase, but performs poorly in the testing phase. The primary causes of overfitting are using insufficient and/or noisy data in training the model and having a number of model parameters that is equal to or greater than the number of observations. A number of measures were taken to address overfitting including using most of the dataset (85%) for training, utilizing the "shuffle" function in Python and cross-validation (CV), and feature or dimensionality reduction. The Python's "shuffle" function randomizes the entire dataset so that a random set of explanatory variables and fecal coliform value pair is chosen at each iteration in the training phase. CV is a widely used technique for evaluating the predictive capability of models for data that they have not seen before. The usual steps in applying CV are: 1) partitioning the training dataset into *k* different subsets or folds, 2) using *k*-1 subsets for training *k* models and testing on the remaining subset, and 3) taking the average of each of the measured performance of the *k* models (Garreta and Moncecchi, 2013). In developing the CAWS-FIB model, a 5-fold CV was used in the model training phase, meaning that 4/5 subsets of the training dataset were used to develop each set of explanatory variable coefficients and using these coefficients to predict the remaining 1/5 subset of the training dataset. Lastly, feature or dimensionality reduction was conducted. As shown in Table B.1, there are a multitude of features or explanatory variables used in model development. The features derived from manually and more frequently (time-lagged) measured environmental variables (Table B.1) ranged from 119–198, while the number of site observations ranged from 25–38. Thus, after the 119–182 features were ranked by relative importance based on preliminary model estimation, only the top 15 most relevant feature or variables were chosen in the final model development, resulting into a data space of 15 (columns) × 25–28 (rows) dimensions. Consequently, model training and testing were conducted using three trials: 1) a 15-feature model, 2) a 10-feature model, and 3) a 5-feature model.

2.2.3.3 Model Prediction

The model with the best overall predictive performance based on predefined metrics was chosen to perform prediction and other computations on a hypothetical dataset designed to evaluate model functionality. The hypothetical dataset covers variations of the 119–198 features/explanatory variables used for model training and testing. For each site, these variations include the mean of the explanatory variables (MEV), 1.1 × MEV, 0.9 × MEV, MEV + one standard deviation of the MEV (MEV + SD1), 1.2 × MEV, and 0.8 × MEV.

2.2.3.4 Model Capability and Performance Metrics

In addition to predicting fecal indicator bactria density given a set of relevant features or environmental variables, the model is capable of estimating the probability of exceedance (POE) similar to the VirtualBeach model (Cyterski et al., 2013). Basically, the POE is the probability (%) that a predicted fecal coliform density will exceed a threshold number. The threshold number is a function of the regulatory limit (RL) and a decision value (DV). A DV is basically used as the basis for determining whether or not to issue a water quality advisory on a portion of the CAWS used for contact recreation. While RL is fixed as set by law or proclamation, DVs can be set lower, higher, or equal to the RL depending on which value will optimize model performance (i.e., balancing between sensitivity, specificity, and/or overall accuracy, which are defined below) based on the plot of model fits vs. actual observations. In these model tests, the CAWS limit for fecal coliform of 200 CFU/100 mL was used as the RL and DV, respectively.

Model performance metrics include coefficient of determination (R^2), mean squared error (MSE), root MSE (RMSE), specificity, sensitivity, and accuracy. Accuracy, sensitivity, and specificity are computed as (true positives + true negatives)/number of total observations, true positives/(true positives + false negatives), and true negatives/(true negatives + false positives), respectively. Predicted fecal coliform values are categorized as false positives (FP), true positives (TP), false negatives (FN), and true negatives based on the quadrant they fall into in the observed vs. predicted scatter plot (Figure 38). The choice of DV is arbitrary and the final choice should be based on optimizing balance between sensitivity and specificity. Higher sensitivity means that people are rarely exposed to high fecal coliform concentrations, while high specificity corresponds to minimizing the chances of putting up a water quality advisory when the actual fecal coliform density is below the threshold. Values of R^2 close to 1 (normal range is between 0 and 1), lower values of MSE and RMSE, particularly close to 0, and higher values of accuracy, especially close to 1 are indicative of an accurate predictive model.

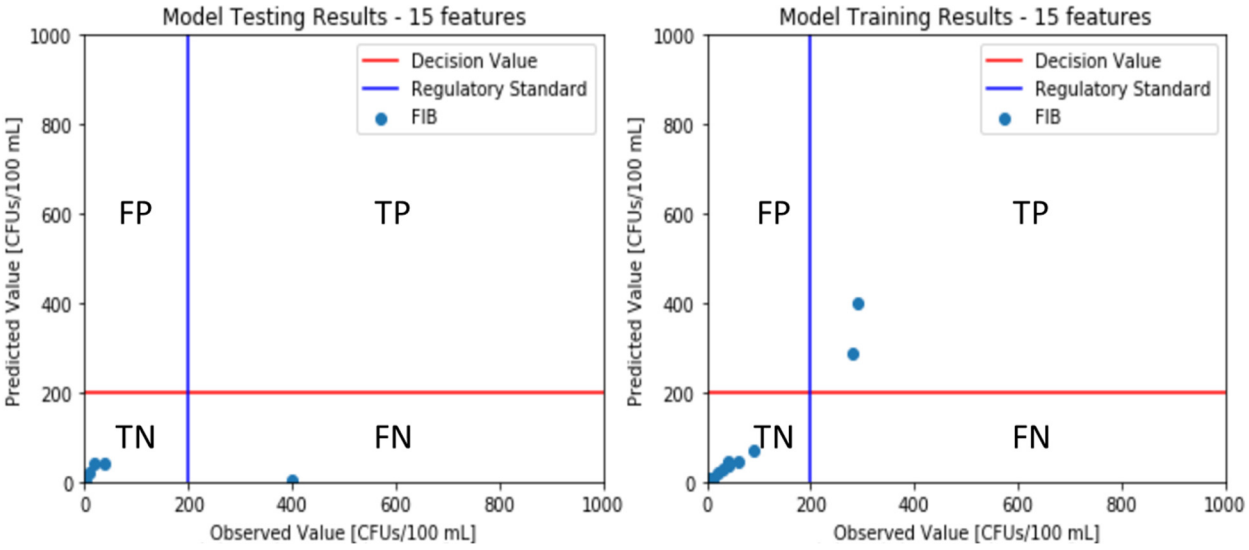


FIGURE 38 A sample a scatter plot of observed and predicted fecal values during model training and testing showing false positives (FP), true positives (TP), true negatives (TN), and false negatives (FN).

2.3 RESULTS AND DISCUSSION

2.3.1 Model Training and Testing

In general, a 15-feature model showed relatively better model performance across the four algorithms (ANNs, GBM, MLR-AL, and PLSR) being tested. Thus, for the purpose of brevity, only the performance metrics of a 15-feature model for each of the four algorithms are summarized and only R^2, RMSE, and accuracy, three of the most important metrics of accurate model training and testing performance, are shown. The values of the R^2, RMSE, and accuracy for each of a 15-feature model averaged over the seven water quality sampling sites used for model testing are shown in Tables 17, 18, and 19, respectively. The values of R^2, RMSE, and accuracy for each of a 15-feature model by site are shown in Tables B.2, B.3, and B.4, respectively. The performance metrics shown represent both the model training and testing phases in both the pre-disinfection/pre-TARP implementation (2013–2015) and post-disinfection/TARP implementation (2016–2018) periods.

Both the ML-based and traditional statistics-based models showed better R^2 values in the training phase compared to the testing phase, regardless of whether it is during the pre-disinfection period/pre-TARP implementation period (2013–2015) or the post-disinfection/TARP-implementation period (2016–2018) (Table 2-1). In the training phase, ML-based models consistently outperformed traditional statistics-based models with GBM-based model consistently showing an $R^2 = 1.00$. However, in the testing phase none of the models showed acceptable R^2 values. All the R^2 values are outside of the normal (0–1.0) range. These results indicate overfitting, because the models performed very well using historical or training data, but performed poorly in the testing phase.

TABLE 17 Mean values of the coefficient of determination (R^2) during testing and training phases based on a 15-feature model conducted over seven water quality sampling sites grouped by periods (2013–2015 and 2016–2018). Values inside the parentheses are standard deviation. (R^2) values close to 1.0 are indicative of accurate models.

	2013–2015		2016–2018	
Model	Training	Testing	Training	Testing
ANNs	0.97 (0.02)	-61.33 (97.95)	0.96 (0.06)	-0.56 (1.36)
GBM	1.00 (0.00)	-1.21 (2.86)	1.00 (0.00)	-0.40 (1.66)
MLR-AL	0.52 (0.21)	6.93 (16.88)	0.68 (0.14)	-0.05 (0.49)
PLSR	0.76 (0.24)	-87.11 (186.75)	0.86 (0.15)	-487.88 (1130.09)

The RMSE values showed that models with ANNs and MLR-AL algorithms outperformed models with either the GBM and PLSR algorithm (Table 10). In the both the training and testing phases, models with ANNs and MLR-AL algorithms have average RMSE values of 1.08 or lower. The model with GBM algorithm performed well in the training phase with an average RMSE = 0.00, a value that is very rarely achieved in practice, but jumped to approximately 30000 and 37000 in the testing phase on the 2013–2015 and 2016–2018 datasets, respectively. The model with PLSR algorithm performed the worst with average RMSE values higher than 12000 in both the training and testing phases.

Like the R^2 and RMSE values, the mean values of accuracy (Table 11), a measure of the model's ability to classify whether a predicted value is above (true positive) or below (true negative) the regulatory standard and decision value (DV), consistently showed better model performance in training than in testing phase. With the exception of PLSR results for the 2016–2018 dataset, the difference in the average accuracy between model training and testing ranged from 0.09 (PLSR-based model on 2013–2015 dataset) to 0.49 (ANNs-based model on 2013–2015 dataset).

TABLE 18 Mean values of the root-mean-square error (RMSE) during testing and training phases based on a 15-feature model conducted over seven water quality sampling sites grouped by periods (2013–2015 and 2016–2018). Values inside the parentheses are standard deviation. RMSE values close to 0 are indicative of accurate predictive models.

	2013–2015		2016–2018	
Model	Training	Testing	Training	Testing
ANNs	0.28 (0.11)	9.47 (10.99)	0.18 (0.15)	1.08 (0.36)
GBM	0.00 (0.00)	29898.54 (65876.54)	0.00 (0.00)	36669.22 (88449.57)
MLR-AL	1.06 (0.17)	2.07 (0.81)	0.61 (0.21)	1.04 (0.50)
PLSR	13238.49 (15058.66)	44215.42 (64834.57)	124143.91 (326154)	83127.89 (214468)

TABLE 19 Mean values of the accuracy performance metric during testing and training phases based on a 15-feature model conducted over seven water quality sampling sites grouped by periods (2013–2015 and 2016–2018). Values inside the parentheses are standard deviation. Accuracy = (true positives + true negatives)/number of total observations. Accuracy values close to 1 are indicative of accurate classification models.

	2013–2015		2016–2018	
Model	Training	Testing	Training	Testing
ANNs	0.95 (0.05)	0.46 (0.23)	0.93 (0.06)	0.73 (0.13)
GBM	1.00 (0.00)	0.79 (0.12)	0.94 (0.15)	0.83 (0.24)
MLR-AL	0.84 (0.10)	0.59 (0.14)	0.85 (0.08)	0.66 (0.23)
PLSR	0.68 (0.10)	0.59 (0.20)	0.76 (0.17)	0.77 (0.19)

Sample plots showing long lists of explanatory variables for fecal coliform ranked from the most to the least important is shown in Figure B.1. This is one of the outputs in the initial step of model development. Due to the small size of the dataset, it was found that using varying random configurations of the testing and training sets produced large differences in the variable importance due to the large amount of variables compared to the total number of data points per variable. Therefore, reduction of the dimensionality of the feature space before using each of the four models to find variable importance was necessary. Initially, each model was run 30 times over the 30 differing random configurations of the test and training data use and the median of the variable importance was used to determine the most relevant explanatory features (Figure B.1). The final model development was then conducted by running each model on the 15, 10, and 5 most relevant explanatory variables for each site. Figures 39, 40, 41, and 42 show the top 15 most relevant explanatory variables for ANNs, GBM, MLR-AL, and PLSR-based models, respectively, based on all the seven water quality sites using for model development in both the 2013–2015 and 2016–2018 datasets. The top 15 most explanatory variables across the seven sites are dominated by more frequently measured or time-lagged/transformed environmental (solar radiation, rainfall, air temperature, dissolved oxygen, water temperature, and specific conductance) and hydraulic (stage and discharge) variables and CSOs. However, manually-measured parameters (whose data were collected at the same time as the fecal coliform samples) including NO_3, water temperature (TwMan), dissolved oxygen (DOMan), and pH have also shown to be important as they were at the top three of the list in more than three instances, most notably using the GBM algorithm (Figure 40).

There is the potential for key explanatory variables to vary between sites due to site specific differences in local hydrology geography, and surrounding land use. Overall, however, the most important explanatory variables for individual sites were generally inconsistent by model method and between the two time periods, indicating there was insufficient data for the model to identify key environmental factors at local scales.

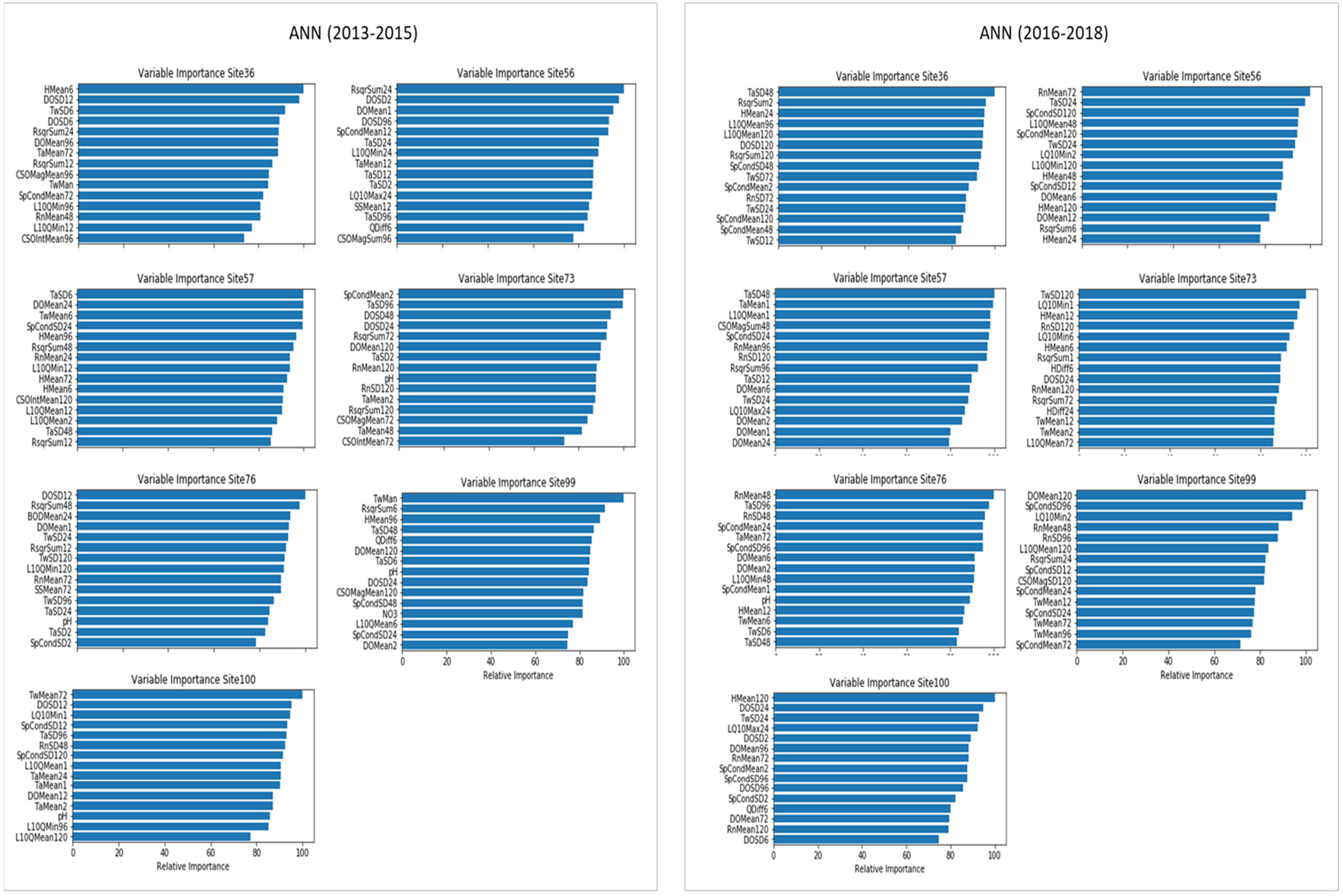


FIGURE 39 The 15 most important explanatory variables for all the seven sites using the ANNs-based model. Abbreviations for model explanatory variables can be found in Table B.1. Explanatory variables, in general, are named by an abbreviation for the environmental variable, followed by the transformation used (if any, i.e., mean and standard deviation (SD)), and the time-lag (1–120 hours). For instance, standard deviation of solar radiation for the last 96 hours and mean of specific conductance for the last 120 hours are symbolized as RnSD96 and SpCondMean120, respectively. The only exception are the Log-transformed variables such as discharge (Q) where the name starts with "L" for logarithmic transformation and the logarithm base (10) is included. For example, the maximum value of the logarithm to the base 10 of Q for the last 24 hours is named as $LQ_{10}Max_{24}$.

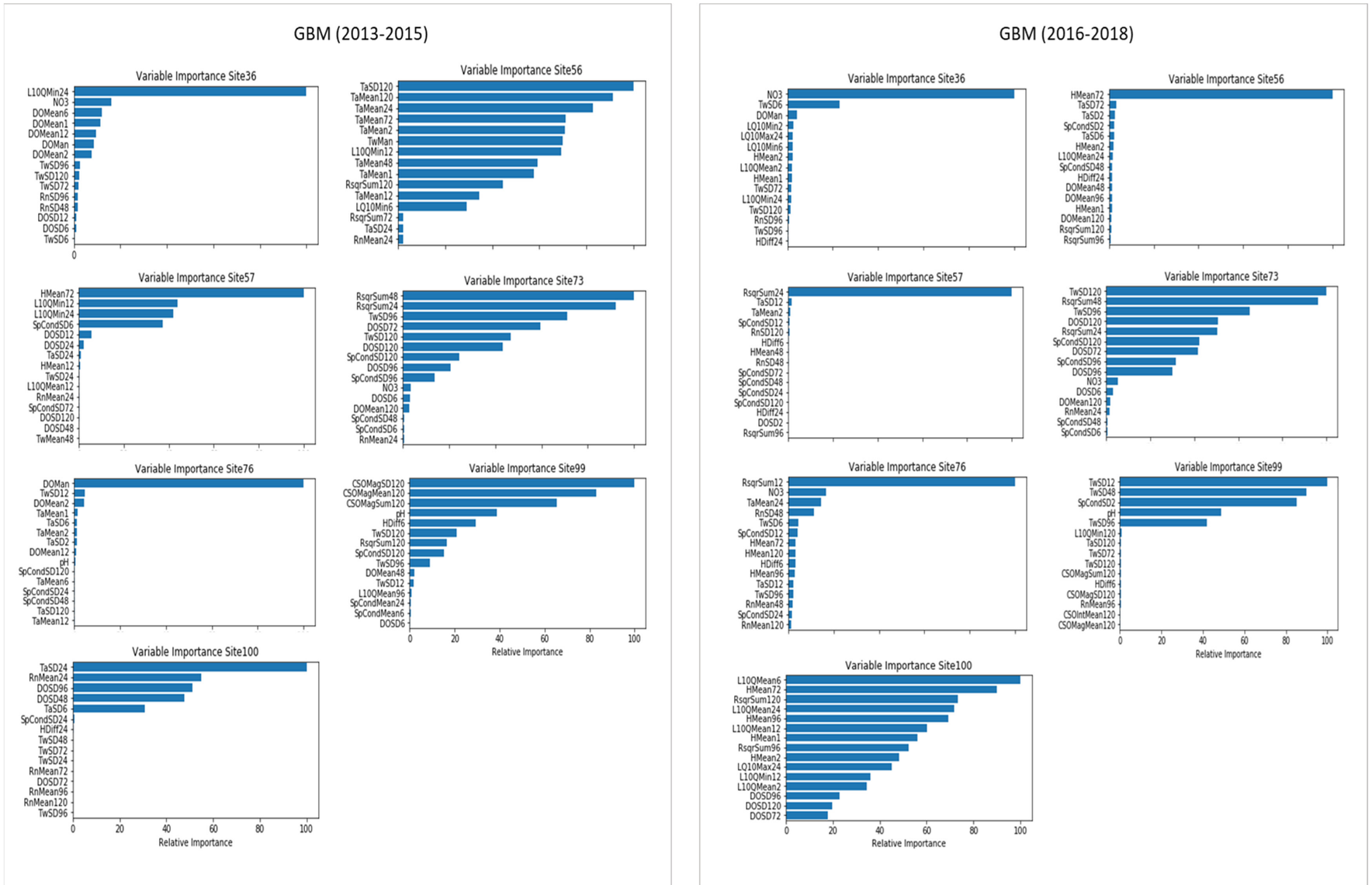
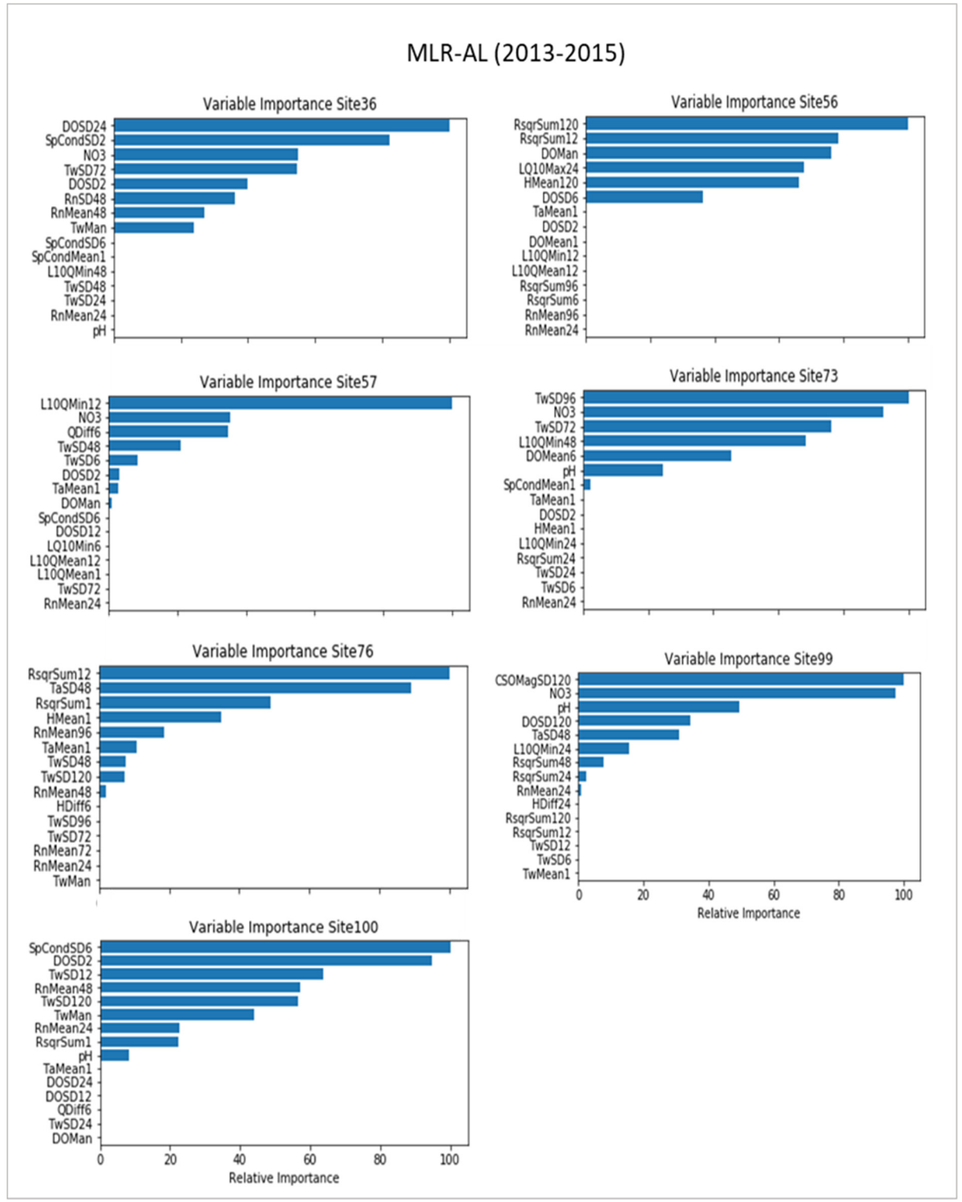


FIGURE 40 The 15 most important explanatory variables for all the seven sites using GBM-based model. Abbreviations for model explanatory variables can be found in Table B.1 Explanatory variables, in general, are named by an abbreviation for the environmental variable, followed by the transformation used (if any, i.e., mean and standard deviation (SD)), and the time-lag (1-120 hours). For instance, standard deviation of solar radiation for the last 96 hours and mean of specific conductance for the last 120 hours are symbolized as RnSD96 and SpCondMean120, respectively. The only exception are the Log-transformed variables such as discharge (Q) where the name starts with "L" for logarithmic transformation and the logarithm base (10) is included. For example, the maximum value of the logarithm to the base 10 of Q for the last 24 hours is named as $LQ_{10}Max_{24}$.



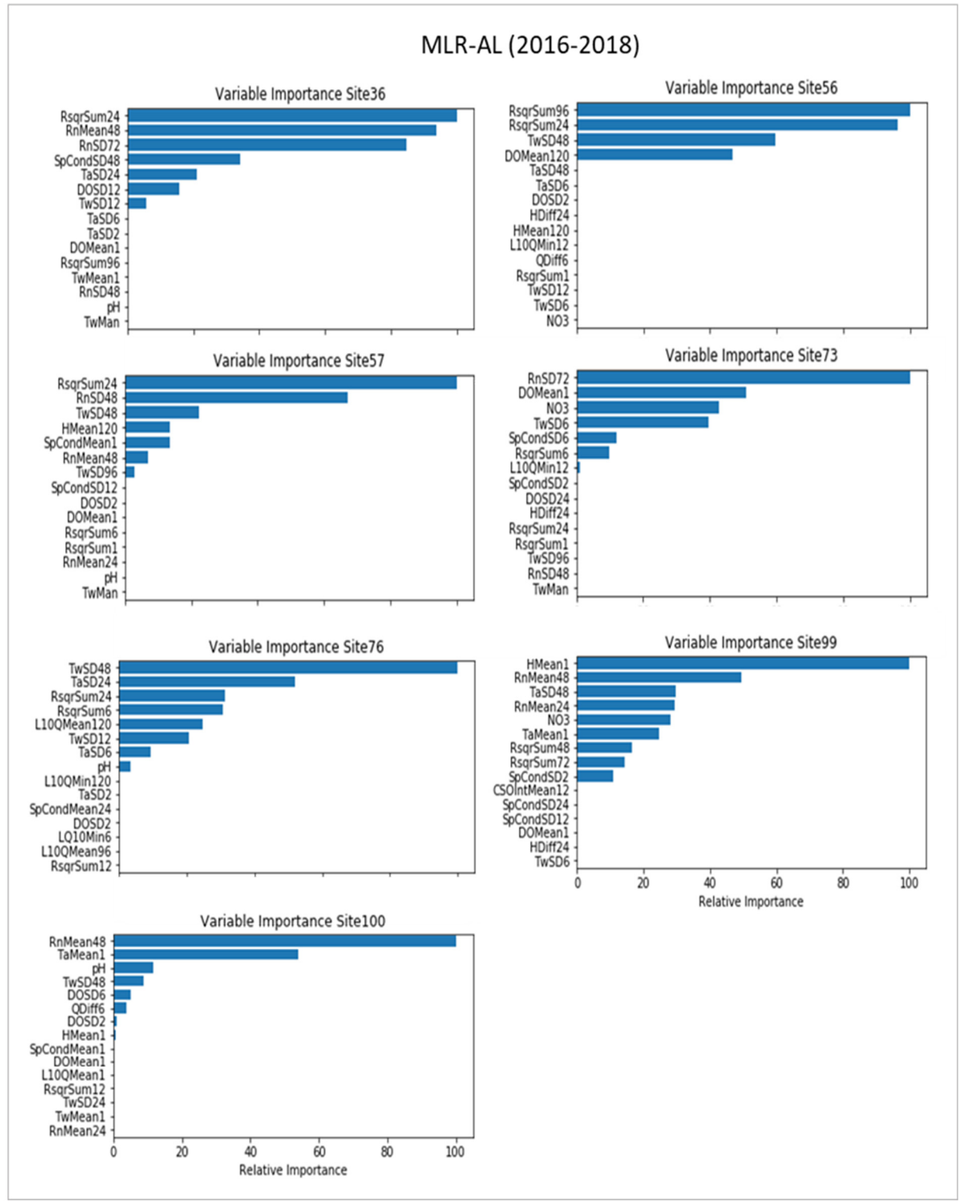
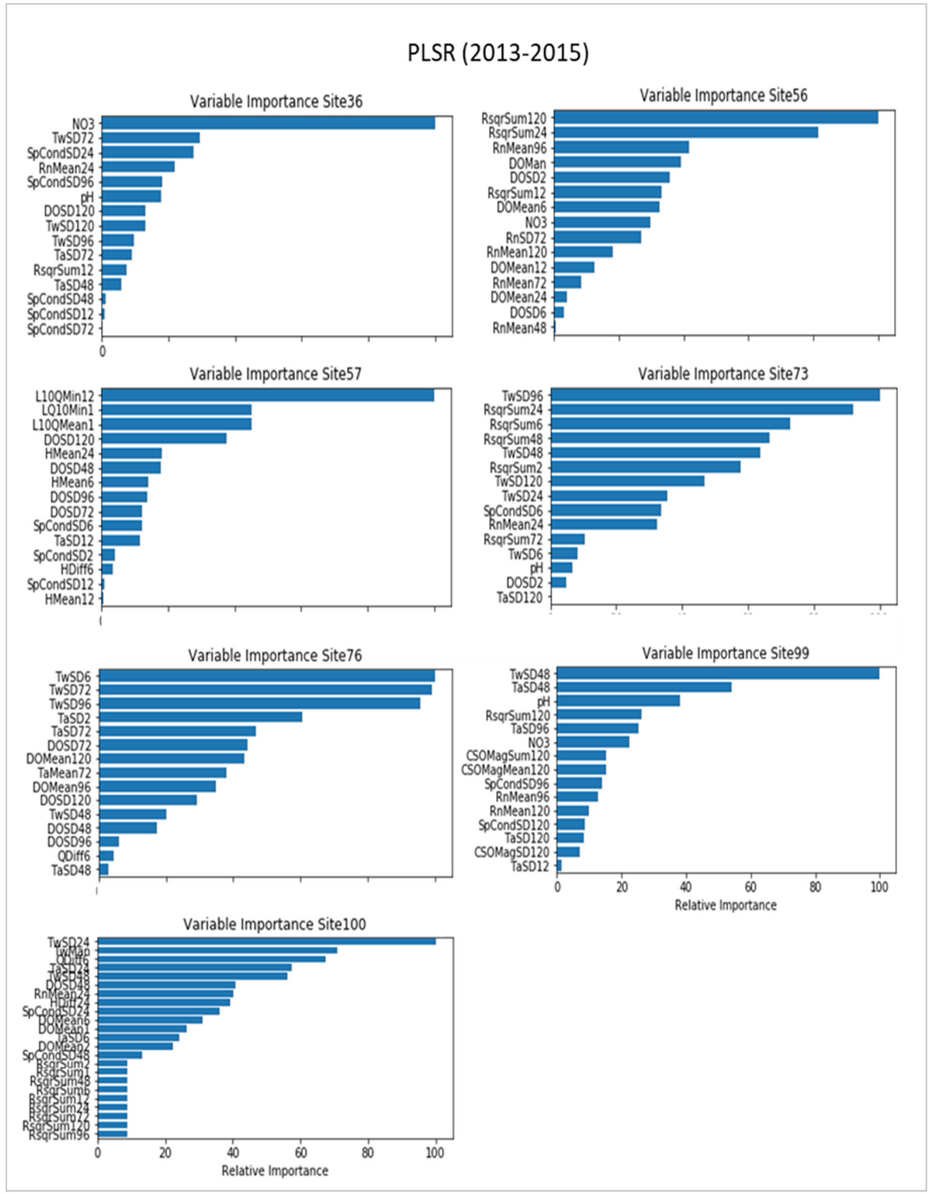


FIGURE 41 The 15 most important explanatory variables for the seven sites using MLR-AL-based model. Abbreviations for model explanatory variables can be found in Table B.1. Explanatory variables, in general, are named by an abbreviation for the environmental variable, followed by the transformation used (if any, i.e., mean and standard deviation (SD)), and the time-lag (1–120 hours). For instance, standard deviation of solar radiation for the last 96 hours and mean of specific conductance for the last 120 hours are symbolized as RnSD96 and SpCondMean120, respectively. The only exception are the Log-transformed variables such as discharge (Q) where the name starts with "L" for logarithmic transformation and the logarithm base (10) is included. For example, the maximum value of the logarithm to the base 10 of Q for the last 24 hours is named as $LQ_{10}Max_{24}$.



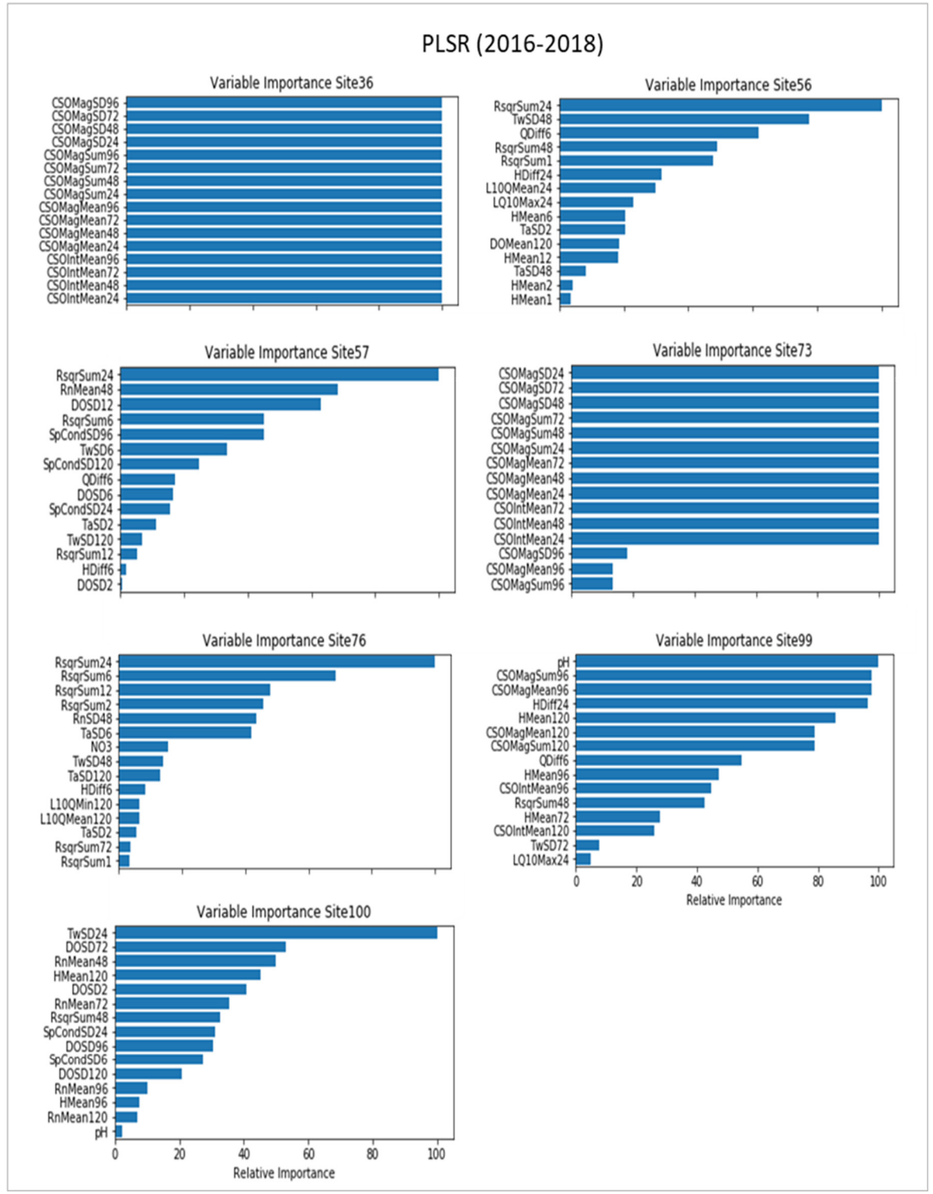


FIGURE 42 The 15 most important explanatory variables for the seven sites using PLSR-based model. Abbreviations for model explanatory variables can be found in Table B.1. Explanatory variables, in general, are named by an abbreviation for the environmental variable, followed by the transformation used (if any, i.e., mean and standard deviation (SD)), and the time-lag (1–120 hours). For instance, standard deviation of solar radiation for the last 96 hours and mean of specific conductance for the last 120 hours are symbolized as RnSD96 and SpCondMean120, respectively. The only exception are the Log-transformed variables such as discharge (Q) where the name starts with "L" for logarithmic transformation and the logarithm base (10) is included. For example, the maximum value of the logarithm to the base 10 of Q for the last 24 hours is named as $LQ_{10}Max_{24}$.

2.3.2 Model Prediction

Table 20 shows the predicted fecal coliform concentrations using the hypothetically generated dataset and probability of exceedance (POE) values based on the RL and DV of 200 CFUs/100 mL for site 76. Predicted fecal coliform concentrations results for the sites 36, 56, 57, 73, 99, and 100 can be found in Tables B.5, B.6, B.7, B.8, B.9, and B.10, respectively.

Table 20 depicted the general picture of the predicted fecal coliform density values for each of the four models. Results of model tests using seven sites showed that models with GBM and PLSR algorithms consistently overpredicts fecal coliform density values, and in many instances produced very large values shown as "inf" in the table, which stands for infinity. This observation is true regardless of the input or predictors used (e.g., MEV, 1.1 × MEV, 0.9 × MEV). Conversely, an MLR-AL-based model tends to predict low fecal coliofrm density values relative to the predictions of the other three models. The ANN algorithm is the only model that produces values that are, with a few exceptions, reasonably within the range of the fecal coliofrm density values used for training and testing. The ANN algorithm was relatively robust in predicting the fecal coliform density in the CAWS despite of the limited sampling data available for model training and testing.

The probability of exceedance (POE) assignment to fecal coliform value that is higher than both the regulatory standard and DV values is consistently close or equal to 100% across all the four models (Table 20 and Tables B.5–B.10), especially for relatively large fecal coliform values. Conversely, the POE assignment for predicted fecal coliform values lower than the regulatory standard and DV values seem to be inconsistent, particularly for the classical statistical approaches. The expectation is that POE should be 0 or close to 0 when the predicted fecal coliform value is 0 or close to 0. However, a predicted value of less than 1 CFU/100 mL from an MLR-AL-based was assigned a POE of approximately 47% (Table B.7). In the same manner, a predicted fecal coliform of 0 CFU/100 mL from site 99 was assigned a POE of approximately 49% (Table B.9). This inconsistency in the POE assignment could be primarily attributed to the inability of models based on traditional statistical methods to properly handle interdependent or non-orthogonal features. This limitation is one of the primary areas of continuous improvement for future CAWS-FIB model development. Another limitation is that the data used for model training and testing has very limited volume (26–38 data points) with a very large number of features (119–182) and very large dynamic value range. It has to be noted though that these predictions are based on a hypothetical dataset whose primary purpose is to showcase the basic model functionalities, not the absolute values of the predictions, and may not completely represent the distribution of the dataset from which the models were trained and tested. The only way to evaluate the accuracy of the absolute predicted values and their associated POE values is to fix the model limitation for classical statistical methods and use actual/measured data for prediction accuracy evaluation purposes only; i.e., measured data that have not be used for either model training and/or testing phase(s).

TABLE 20 Model predicted fecal coliform concentration (CFUs/100 mL) and probability of exceedance (POE, %) based on the regulatory standard of 200 CFUs/100 mL and decision value (DV) of 200 CFUs/100 mL for site 76. MEV = mean of the observed values for the 119–192 explanatory variables (MEV) for a given site for each of the two periods (i.e., 2013–2015 and 2016–2018); 1.1 × MEV; 0.90 × MEV; MEV + one standard deviation (SD1); 1.2 × MEV; 0.8 × MEV.

				2013–2015		2016–2018	
Model	Input	Regulatory Standard [CFU/100 mL]	DV [CFU/100 mL]	FIB_predicted [CFU/100 mL]	POE [%]	FIB_predicted [CFU/100 mL]	POE [%]
ANNs	MEV	200	200	3.03E+09	100	16.27	0.00
GBM				inf	100	1.03E+51	100
MLR-AL				2.94E-05	0.00	49.28	1.90
PLSR				inf	100	2.97E+93	100
ANNs	1.1 × MEV	200	200	1.02E+10	100	35.19	0.00
GBM				inf	100	1.03E+34	100
MLR-AL				5.65E-06	0.02	46.78	1.75
PLSR				inf	100	3.00E+72	100
ANNs	0.9 × MEV	200	200	2.39E+09	100	7.54	0.00
GBM				inf	100	4.10E+243	100
MLR-AL				1.53E-04	0.00	240.24	71.02
PLSR				inf	100	2.3144E+177	100
ANNs	MEV + SD1	200	200	111337.06	100	1.73	0.00
GBM				inf	100	8.99859E+36	100
MLR-AL				8.57E-10	3.33	48.02	1.82
PLSR				inf	100	9.4334E+82	100
ANNs	1.2 × MEV	200	200	4.50E+10	100	236.81	98.00
GBM				inf	100	7.73488E+62	100
MLR-AL				1.07E-06	0.17	50.58	1.98
PLSR				inf	100	9.3483E+103	100
ANNs	0.8 × MEV	200	200	3.02E+09	100	803.89	100
GBM				inf	100	2.22673E+27	100
MLR-AL				8.03E-04	0.00	45.58	1.68
PLSR				inf	100	9.51928E+61	100

2.4 CONCLUSIONS AND RECOMMENDATIONS

Model training and testing indicated that overfitting was a problem for the CAWS-FIB models. Regardless of algorithm used (ML or classical statistics-based) the models performed well when predicting fecal coliform density during training, but performed poorly during testing. The overfitting is likely related to the low, monthly, sampling frequency for fecal coliform which limits the amount of data available for model training and testing. Overfitting occurred despite the implementation of all the widely used techniques for addressing overfitting from the training and testing dataset management perspective such as randomization of the dataset using Python's "shuffle" function and cross-validation. Under the current water quality sampling scheme of collecting 1–3 grab samples per month from March to November, quantifying/simulating the CAWS' microbial water quality may be best conducted using processed based models as they do not require larger volume of data as their data-driven model counterparts for calibration/validation.

The modeling approaches employed in this study fell short of consistently demonstrating acceptable predictive capabilities. However, in addition to producing an operational model with multiple functionalities, the overall model development effort provided information that would be critical for the success of future ML model development. For example, it provided a short list of the most relevant explanatory variables (including the associated transformations and lagged times) and highlighted the potential of the ANNs as the future algorithm of choice for the CAWS-FIB model as it outperformed the other three algorithms across most of the performance metrics. The key explanatory variables were dominated by environmental (solar radiation, rainfall, air temperature, dissolved oxygen, water temperature, and specific conductance) and hydraulic (stage and discharge) variables although NO_3, water temperature (TwMan), dissolved oxygen (DOMan), and pH were also in the top three most important explanatory variables in more than three algorithms.

Improving the current CAWS-FIB model should start with collecting fecal coliform samples more frequently. The general rule-of-thumb for ML-based model development is for the number of observations (data points) per feature or explanatory variable be, at minimum, 30 times the number of features. This means that for a 15-feature model, a total of 450 measured FIB density samples per site per year must be collected, which translates to 50 data points per water quality sampling site per month from March to October. However, this level of sampling intensity is not likely to be realistic.

Increasing sampling frequency would also incorporate more FIB samples collected during periods of rainfall and CSO (gravity or pumped) discharges. CSOs are one of the most important factors controlling the FIB concentrations, yet under the current dataset CSO-related features or explanatory variables (either as intensity or magnitude) are dominated by 0 values when summarized hourly for the last 120 hours prior to the FIB density sampling time. This indicates that most water quality samples from which the FIB density data were analyzed were oftentimes collected 120 hours after CSO discharges had already occurred, although it should be noted that 2017–2019 post CSO samples were collected within 12 hours of CSO events. Moving forward, the sampling scheme should incorporate event-based sampling into the current sampling

scheme, and the former must be conducted, if at all possible, within 24–120 hours after the CSO discharges have occurred.

In addition, the CAWS-FIB model should be "re-trained" and "re-tested" on an annual-basis, i.e., the model to be used for the current year should be trained and tested using data from the year before, thereby taking into account important, incremental changes in the system such as land use/land cover and structural changes in the hydraulic system (e.g.,TARP). This suggested re-training and re-testing scheme is an improvement of the current model, which assumed that the CAWS' LULC, hydraulic structures, etc. remained exactly the same in each of the two three-year periods (2013–2015 and 2016–2018).

On the model architecture side, the inability of the model to avoid selecting features that are interdependent or non-orthogonal affected the accuracy of the POE values, especially for the traditional statistical models. Additionally, once the dimensionality reduction process is completed, the final selection of features must be iterative, instead of a fixed selection of 5-, 10-, and 15-feature models. In other words, once the number of important features is reduced from say 119 to 20, then the iterative elimination process used to identify the ≤20 most relevant features to be used for the final model training and testing will commence. The first elimination process starts with 19 variables, leaving one different factor out of the set each time an iteration is performed. After all the iterations are completed, the 19-feature set with the best performance metric values will be used to proceed to the 18-variable iteration step. The process is repeated until such time when the performance of the (n-1) feature model becomes inferior relative to the n-feature model. For instance, when a 12-feature model registers an $R^2 = 0.8$, while a previously evaluated 13-feature registered an $R^2 = 0.86$, then the iteration stops, and the identified 13-most relevant features will be used for the final model development.

2.5 REFERENCES

Breiman L (2001) Random forests. *Machine learning* **45**: 5–32.

Brion, G. M., & Lingireddy, S. (2003). Artificial neural network modelling: a summary of successful applications relative to microbial water quality. *Water science and technology*, *47*(3), 235–240.

Brooks, W. R., Fienen, M. N., & Corsi, S. R. (2013). Partial least squares for efficient models of fecal indicator bacteria on Great Lakes beaches. *Journal of environmental management*, *114*, 470–475.

Brooks W, Corsi S, Fienen M & Carvin R (2016) Predicting recreational water quality advisories: A comparison of statistical methods. *Environmental Modelling & Software* **76**: 81–94.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA & Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Meth* **13**: 581-583.

Cyterski M, Zhang S, White E, Molina M, Wolfe K, Parmar R & Zepp R (2012) Temporal synchronization analysis for improving regression modeling of fecal indicator bacteria levels. *Water, Air, & Soil Pollution* **223**: 4841–4851.

Frick WE, Ge Z & Zepp RG (2008) Nowcasting and forecasting concentrations of biological contaminants at beaches: a feasibility and case study. *Environmental science & technology* **42**: 4818–4824.

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232.

Garreta, R., & Moncecchi, G. (2013). *Learning scikit-learn: machine learning in python*. Packt Publishing Ltd., Berlin, Heidelberg.

Ge Z & Frick WE (2007) Some statistical issues related to multiple linear regression modeling of beach bacteria concentrations. *Environmental research* **103**: 358–364.

Hastie, T, Tibshirani, R, & Friedman, J. (2001). The elements of statistical learning. Springer, New York.

Hou, D., Rabinovici, S. J., & Boehm, A. B. (2006). Enterococci predictions from partial least squares regression models in conjunction with a single-sample standard improve the efficacy of beach management advisories.

Jones RM, Liu L & Dorevitch S (2013) Hydrometeorological variables predict fecal indicator bacteria densities in freshwater: data-driven methods for variable selection. *Environmental monitoring and assessment* **185**: 2355–2366.

Natekin A & Knoll A (2013) Gradient boosting machines, a tutorial. *Frontiers in neurorobotics* **7**: 21.

NOAA (2012) North Central Forecasting System. p.^pp.

NOAA (2015) Great Lakes Coastal Forecasting System. p.^pp.

Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert systems with applications*, *36*(1), 2–17.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R & Dubourg V (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**: 2825–2830.

Samuel AL (1959) Some studies in machine learning using the game of checkers. *IBM Journal of research and development* **3**: 210-229.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

USGS (2014a) The National Water Information System. p.^pp.

USGS (2014b) Environmental Data Discovery and Transformation. p.^pp.

Vijayashanthar, V., Qiao, J., Zhu, Z., Entwistle, P., & Yu, G. (2018). Modeling fecal indicator bacteria in urban waterways using artificial neural networks. *Journal of Environmental Engineering*, *144*(6), 05018003.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, *58*(2), 109-130.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429

APPENDIX A:
DNA EXTRACTION FROM ENVIRONMENTAL SAMPLES AND 16S rRNA GENE AMPLIFICATION

A.1 PROTOCOL # 1: DNA EXTRACTION FROM ENVIRONMENTAL SAMPLES

1. The first step of the genomic DNA (gDNA) isolation/extraction is to add sample material to the desired wells of the 96-well PowerBead® DNA Plate(s), Garnet.
2. Remove nuclease and nucleotide contamination from work surfaces and instruments prior to starting using an appropriate solution, such as RNase AWAY™ (Thermo Scientific™ catalogue: 700511), followed by wiping with 70% to 100% molecular biology grade ethanol to remove additional contaminants. This is the way that all surface and instrument, other than the KingFisher™ Flex (follow manufacturer documentation), cleaning steps are carried out for EMP KingFisher™ HTP gDNA extractions by the Knight Lab at University of California San Diego.
3. Remove the PowerBead® DNA Plate (Bead Plate) from the QIAGEN® MagAttract® PowerSoil® DNA KF Kit (384), and centrifuge for 1 minute at 2500 × g to pellet the garnet beads prior to sample addition.
4. Remove the Square Well Mat from the Bead Plate and set aside in a sterile location. The Square Well Mat will be put back on to the Bead Plate after the samples have been added.
5. Add samples to Bead Plate:
    a. Sediment Material: 0.1 to 0.25 grams per well
    b. Liquid Material: 250 µl or less per well
6. Clean all work surfaces and instruments with RNase AWAY™ reagent (Thermo Scientific™ catalogue: 700511), wipe dry, and repeat with 70%-100% ethanol, wipe dry. Turn on a water bath to 65°C.
7. In a sterile reservoir, add 400 µl RNase A Solution to 75 ml of PowerBead® Solution (Bead Solution) for every 96-well plate that will be processed.
    a. 400 µl RNase A Solution (25 mg/ml)
    b. 75 ml PowerBead® Solution
    c. PowerBead® Solution contains guanidinium thiocyanate (CAS: 593-84-0, less than 10% w/w); handle this reagent with care, and dispose of as hazardous chemical waste in accordance with all institutional and local regulations.
    d. RNase A Solution is also a hazardous chemical mixture (ribonuclease, CAS: 9001-99-4, less than 10% w/w), and should be disposed of properly.
    e. RNase A Solution is stable for approximately 1 year at room temperature, 25°C. For longer storage, it is recommended that you store the RNase A Solution at 2°–8°C. The Knight Lab at UC San Diego currently uses this solution at room temperature.

8. Add 750 µl of Bead Solution/RNase A Solution to each well of the Bead Plate(s). 750 µl Bead Solution/RNase A Solution
9. Check the bottle(s) of SL Solution, Lysis buffer. If precipitate is visible, heat at 60°C until dissolved.
    a. SL Solution contains SDS (CAS: 151-21-3, less than 10% concentration w/w), which can precipitate if cold. Heating at 60°C will dissolve the SDS; SL Solution can be used while still warm.
    b. 60°C Water Bath
10. Add 60 µl of SL Solution to each well. Secure the Square Well Mat (retained during sample addition to Bead Plate) tightly to the plate.
11. 60 µl SL Solution
12. Ensure that there is a complete seal of every well in order to prevent sample cross-contamination and/or loss. It is often necessary to use both gloved hands and a plate-sealing roller.
13. Place sealed Bead Plate(s) in 65°C water bath for 10 minutes. DO NOT SUBMERGE THE PLATE(S).
    a. 65°C Water Bath
    b. 00:10:00 Water Bath
    c. During incubation, or prior to starting, fill an ice container that is large enough to accommodate the Bead Plate(s) with enough ice to surround the Bead Plate(s).
14. Remove excess water from the Bead Plate(s), and make sure that all wells are still fully sealed. Place the Bead Plate(s) between two Adapter Plates (QIAGEN® catalogue: 11990) and securely fasten to a 96-well Plate Shaker (such as, QIAGEN® TissueLyser® II; QIAGEN® catalogue: 85300). Most Plate Shakers are designed to process two plates at once. If this is the case, it is important to balance the Plate Shaker. If working with two Bead Plates, simply attach each Bead Plate to a station on the Plate Shaker. If you only have one Bead Plate to affix to the Plate Shaker, attach the sample-containing Bead Plate to one station, and a spare/empty PowerBead® DNA Plate to the second station as a balance.
15. Shake at speed 20 Hz for 20 minutes. It is important to make sure that the Adapter Plates, holding the Bead Plates, are properly situated in the Plate Shaker and tightly fastened. No parts should rub against the Plate Shaker during operation if attached properly.
16. Centrifuge Bead Plate(s) at room temperature for 6 minutes at 3220 × g (or 4500 × g depending on centrifuge). While centrifuging, or during Bead Plate shaking, add 450 µl IR Solution (Inhibitor Removal Technology® Solution) to each well of a clean/empty Collection Plate (1 ml) (provided in kit), and cover with clear Sealing Tape (provided in kit).
    a. 00:06:00 Centrifugation
    b. 450 µl IR Solution to clean Collection Plate

17. Sufficiently clean the work surface and a multichannel pipette that is capable of transferring volumes up to 1000 µl. The tips used with the multichannel pipette must be able to fit in the round wells of the Collection Plate. The Knight Lab uses Rainin™ tips (catalogue: RT-1000F).
18. Remove the lysate-containing Bead Plate(s) from the centrifuge, and carefully remove and discard the Square Well Mat. If working with more than one Bead Plate, only uncover and work with one at a time.
19. Remove the Sealing Tape from the IR Solution-containing Collection Plate.
    a. Transfer 640 µl, or less, lysate from each well of the Bead Plate to the Collection Plate, and mix gently by pipetting up and down 4 times.
    b. 640 µl Bead Plate lysate to IR-Collection Plate
    c. The transferred lysate may contain some particulate matter.
20. Apply a new Sealing Tape to the lysate/IR-containing Collection Plate (repeat process if working with a second Bead Plate).
    a. Incubate Collection Plate(s) at 4°C for 10 minutes.
    b. 4°C Incubation
    c. 00:10:00 Incubation
21. Centrifuge lysate/IR Collection Plate(s) at 3220 × g for 6 minutes.
22. Remove Sealing Tape from lysate/IR Collection Plate(s).
23. Transfer entire volume of supernatant (~ 850 µl), avoiding pellet, to a new/sterile 1 ml Collection Plate (Collection Plate #2). Discard the used Collection Plate(s). Some pellet material will likely be transferred to the new Collection Plate(s) #2.
    a. 850 µl supernatant
24. Apply new Sealing Tape to Collection Plate(s) #2, and centrifuge at 3220 × g for 6 minutes.
25. Remove the Sealing Tape from Collection Plate(s) #2, and transfer 450 µl of supernatant to a clean KingFisher™ Deep Well 96 Plate. Transfer the remaining supernatant, 400 µl, to a second KingFisher™ Deep Well 96 Plate. Discard Collection Plate(s) #2.
    a. 450 µl supernatant
    b. 400 µl remaining supernatant
    c. This is an appropriate place to stop. If stopping, seal the KingFisher™ Deep Well 96 Plates with plate sealing foil, not Sealing Tape from kit, and store at 4°C overnight. Do not store longer than 1 day.
    d. ClearMag® Reagent Aliquoting

26. For each 96 well plate processed, aliquot 500 µl ClearMag® Wash Solution to each well of 3 clean KingFisher™ Deep Well 96 Plates, and 100 µl Solution EB, Elution buffer, to each well of 1 clean KingFisher™ 96 KF Microtiter (200 µl) Plate.
    a. 500 µl ClearMag® Wash Solution
    b. 100 µl Solution EB, Elution buffer
    c. 65 µl Solution EB is used for low biomass sample types.
27. For each 96 well plate processed, suspend 2 ml ClearMag® Zorb Reagent in 45 ml of ClearMag® Binding Solution in a clean reservoir. Pipette up and down thoroughly to evenly disperse the magnetic beads in solution.
    a. 2 ml ClearMag® Zorb Reagent
    b. 45 ml ClearMag® Binding Solution
    c. The beads will settle quickly, mix thoroughly right before addition to sample.
28. For each 96 well plate processed, add 47 ml of ClearMag® Binding Solution to a separate, clean, reservoir.
    a. 47 ml ClearMag® Binding Solution
29. Add 470 µl ClearMag® Zorb Reagent/ClearMag® Binding Solution to each well of one sample lysate containing KingFisher™ Deep Well 96 Plate.
    a. 470 µl ClearMag® Zorb Reagent/ClearMag® Binding Solution
30. To the remaining KingFisher™ Deep Well 96 Plate(s) containing lysate, add 470 µl ClearMag® Binding Solution to each well.
    a. 470 µl ClearMag® Binding Solution
31. Initiate the 'KF_Flex_MoBio_PowerMag_Soil_DNA' program on the KingFisher™ Flex™ robot. Ensure that the protocol is set to utilize both sample lysate aliquots. Depending on the version of BindIt™ Software operating on the KingFisher™ Flex™ Purification System, the "KF_Flex_MoBio_PowerMag_Soil_DNA" script may need to be downloaded and transferred to the KingFisher™ Flex™ machine. Currently, the robotic script for the MagAttract® PowerSoil® DNA KF Kit can be found by visiting the following site, opening the SDS/Protocols tab, and selecting the "KingFisher Flex" option under the "Robotic Scripts" header: https://mobio.com/products/dna-isolation/soil/powermag-soil-dna-isolation-kit.html. If using two aliquots of the sample lysate for purification, as in this EMP protocol, add an additional sample binding step after the first (before the first wash step) in the BindIt™ Software, and transfer the modified protocol to the robotic platform.
32. Follow the onscreen prompts to properly load the KingFisher™ Flex™. The loading order should be: the tip comb, elution plate, ClearMag® Wash Solution filled plates, lysate with ClearMag® Binding Solution plate (Bind 2), and lysate containing ClearMag® Binding Solution/ClearMag® Zorb Reagent plate (Bind1). Running the KingFisher™ Flex™ Purification System.

33. The selected KingFisher™ Flex™ program will execute itself once the final plate is added and "Start" is pressed. The program takes approximately 65 minutes to complete, and requires no user intervention.
34. When the KingFisher™ Flex™ program finishes, remove the gDNA containing elution plate, and seal this with an appropriate storage seal (not Sealing Tape from kit).
35. Follow the onscreen prompts to cycle through each station on the KingFisher™ Flex™ deck. Dispose of all liquids from plates as hazardous chemical waste (the gDNA elution plate should already be removed and appropriately sealed), and discard the emptied plates. The gDNA is now ready for downstream applications.

A.2 PROTOCOL # 2: 16S rRNA GENE AMPLIFICATION

1. Amplify samples in triplicate, meaning each sample will be amplified in 3 replicate 25-µL PCR reactions.
2. Pool triplicate PCR reactions for each sample into a single volume (75 µL). Do not combine amplicons from different samples at this point.
3. Run amplicons from each sample on an agarose gel. Expected band size for 515F–806R is ~300–350 bp. Low-biomass samples may yield faint or no visible bands; alternative methods such as a Bioanalyzer could be used to verify presence of PCR product.
4. Quantify amplicons with Quant-iT PicoGreen dsDNA Assay Kit (ThermoFisher/Invitrogen cat. no. P11496; follow manufacturer's instructions).
5. Combine an equal amount of amplicon from each sample (240 ng) into a single, sterile tube. Higher amounts can be used if the final pool will be gel-isolated or when working with low-biomass samples. Note: When working with multiple plates of samples, it is typical to produce a single tube of amplicons for each plate of samples.
6. Clean amplicon pool using MoBio UltraClean PCR Clean-Up Kit (cat. no. 12500; follow manufacturer's instructions). If working with more than 96 samples, the pool may need to be split evenly for cleaning and then recombined. Optional: If spurious bands were present on gel (in step 3), one-half of the final pool can be run on a gel and then gel extracted to select only the target bands.
7. Measure concentration and A260/A280 ratio of final pool that has been cleaned. For best results the A260/A280 ratio should be between 1.8–2.0.
8. Send an aliquot for sequencing along with sequencing primers listed below.

   16S sequencing primers
   Read 1 sequencing primer

   Field descriptions (space-delimited):

   Forward primer pad

Forward primer linker
Forward primer

TATGGTAATT GT GTGYCAGCMGCCGCGGTAA
Read 2 sequencing primer

Field descriptions (space-delimited):

Reverse primer pad
Reverse primer linker
Reverse primer

AGTCAGCCAG CC GGACTACNVGGGTWTCTAAT
Index sequencing primer

AATGATACGGCGACCACCGAGATCTACACGCT

Note: The 5′ adapter sequence/index sequencing primer has an extra GCT at its 3′ end compared to Illumina's usual index primer sequences. These bases were added to the 3′ end of the Illumina 5′ adapter sequence to increase the Tm for read 1 during sequencing.

APPENDIX B: CAWS-FIB VARIABLES

TABLE B.1 Variables and the corresponding summary statistics used in predicting FIB at each sampling point. Variables recorded at an hourly (or more frequent) time intervals are summarized over nine time windows or lag times.

Variable and Statistics	Description and Units	Time Windows/lags (hrs) 1	2	6	12	24	48	72	96	120
R_n	Net solar radiation, MJ m^{-2} hr^{-1}									
Mean		✓	✓	✓	✓	✓	✓	✓	✓	✓
Std. Dev.		✓	✓	✓	✓	✓	✓	✓	✓	
T_a	Air temperature, °C									
Mean		✓	✓	✓	✓	✓	✓	✓	✓	✓
Std. Dev.			✓	✓	✓	✓	✓	✓	✓	✓
T_w	Water temperature, °C									
Mean		✓	✓	✓	✓	✓	✓	✓	✓	✓
Std. Dev.				✓	✓	✓	✓	✓	✓	✓
R_{sqr}	Square root of rainfall measured									
Sum	in mm	✓	✓	✓	✓	✓	✓	✓	✓	✓
Q	Discharge, m^3 s^{-1}									
Difference				✓						
$log_{10}Q$	Logarithm of discharge									
Mean	measured in m^3 s^{-1}	✓	✓	✓	✓	✓	✓	✓	✓	✓
Min		✓	✓	✓	✓	✓	✓	✓	✓	✓
Max						✓				
H	Stage, m									
Mean		✓	✓	✓	✓	✓	✓	✓	✓	✓
Difference				✓		✓				
CSO_int	Intensity of the combined sewer									
Mean	overflows (CSOs), gph	✓	✓	✓	✓	✓	✓	✓	✓	✓
Difference			✓	✓		✓				
log10 CSO_int	Logarithm of the intensity of									
Mean	CSOs measured in gph	✓	✓	✓	✓	✓	✓	✓	✓	✓
Min		✓	✓	✓	✓	✓	✓	✓	✓	✓
Max						✓				
CSO_mag	Magnitude of CSO, gal									
Sum		✓	✓	✓	✓	✓	✓	✓	✓	✓
Mean		✓	✓	✓	✓	✓	✓	✓	✓	✓
Std. Dev.			✓	✓	✓	✓	✓	✓	✓	✓
log10 CSO_mag	Logarithm of the magnitude of									
Mean	CSO measured in gal	✓	✓	✓	✓	✓	✓	✓	✓	✓
Min		✓	✓	✓	✓	✓	✓	✓	✓	✓
Max						✓				
log10 Turb	Logarithm of turbidity, NTU	Manual								
log10 SS	Logarithm of suspended solids measured in mg L^{-1}	Manual								
pH	potential of Hydrogen	Manual								
TOC	Total organic carbon, mg L^{-1}									
DO	Dissolved oxygen, mg L^{-1}									
TDS	Total dissolved solids, mg L^{-1}	Manual								

TABLE B.1 (Cont.)

Variable and Statistics	Description and Units	Time Windows/lags (hrs) 1	2	6	12	24	48	72	96	120
TP	Total phosphorus, mg L^{-1}	Manual								
TKN	Total Kjeldahl Nitrogen, mg L^{-1}	Manual								
Chl	Chlorophyll, µg L^{-1}	Manual								
Cl	Chlorine, mg L^{-1}	Manual								
NO3	Nitrate+Nitrite nitrogen, mg L^{-1}	Manual								

TABLE B.2 Values of the coefficient of determination (R^2) for each of the seven water quality sampling sites during training and testing phases based on a 15-feature model. Numbers inside the parentheses represent R^2 values for the testing phase.

	Coefficient of Determination (R^2)			
2013–2015	ANN[a]	GBM[b]	MLR-AL[c]	PLSR[d]
Site36	0.95(-277.90)	1.00(-0.24)	0.66(-1.23)	0.88(-1.62)
Site56	0.99(-2.03)	1.00(-1.10)	0.33(0.00)	0.71(-58.02)
Site57	0.97(-32.07)	1.00(-7.57)	0.24(-1.00)	0.37(-4.06)
Site73	0.97(-28.39)	1.00(-0.46)	0.52(0.11)	0.95(-11.19)
Site76	0.97(-14.70)	1.00(-0.97)	0.41(-1.14)	0.52(-508.07)
Site99	1.00(-67.52)	1.00(0.70)	0.79(-0.06)	0.96(0.41)
Site100	0.94(-6.67)	1.00(0.20)	0.72(-45.18)	0.94(-27.23)
2016–2018				
Site36	0.99(-0.07)	1.00(-1.59)	0.78(0.10)	1.00(0.00)
Site56	0.98(-3.33)	1.00(0.30)	0.59(0.02)	0.78(-19.70)
Site57	0.83(0.34)	1.00(0.59)	0.55(0.60)	0.85(-3033.06)
Site73	1.00(-0.65)	1.00(-3.62)	0.90(-3.62)	1.00(-2.87)
Site76	0.98(0.30)	1.00(0.80)	0.51(-0.11)	0.93(-0.68)
Site99	0.99(0.61)	1.00(-0.25)	0.74(-0.92)	0.88(-359.05)
Site100	0.97(-1.11)	1.00(0.95)	0.70(0.30)	0.57(0.21)

[a] ANN = Artificial Neural Network;

[b] GBM = Gradient Boosting Machine;

[c] MLR-AL = Multiple Linear Regression with adaptive LASSO;

[d] PLSR = Partial Least Squares Regression.

TABLE B.3 Values of the root-mean-square error (RMSE) for each of the seven water quality sampling sites during training and testing phases based on a 15-feature model. Numbers inside the parentheses represent RMSE values for the testing phase.

	Root-mean-square Error (RMSE)			
2013–2015	ANN[a]	GBM[b]	MLR-AL[c]	PLSR[d]
Site36	0.35(31.03)	0.00(7784.00)	0.86(2.90)	2489.57(13095.47)
Site56	0.15(1.29)	0.00(778.96)	0.96(0.88)	624.81(735.60)
Site57	0.25(7.54)	0.00(6636.93)	1.26(1.62)	3007.99(5408.22)
Site73	0.31(2.04)	0.00(1504.63)	1.10(1.37)	3738.44(14969.79)
Site76	0.34(2.88)	0.00(13382.67)	1.30(2.09)	16409.21(20955.85)
Site99	0.14(17.36)	0.00(291.76)	0.98(2.73)	39683.01(180540.26)
Site100	0.42(4.13)	0.00(178910.81)	0.95(2.90)	26716.41(73802.78)
2016–2018				
Site36	0.13(1.08)	0.00(12965.15)	0.49(1.60)	141.53(110.65)
Site56	0.07(0.82)	0.00(15.73)	0.34(0.37)	33.56(59.37)
Site57	0.49(0.82)	0.00(2772.91)	0.78(0.67)	1486.87(4941.01)
Site73	0.09(1.50)	0.00(3005.43)	0.46(1.05)	767.26(353.53)
Site76	0.09(0.58)	0.00(55.24)	0.52(0.61)	26.36(78.05)
Site99	0.12(1.28)	0.00(236995.34)	0.91(1.54)	863788.14(569454.77)
Site100	0.29(1.47)	0.00(874.72)	0.78(1.47)	2763.66(6897.85)

[a] ANN = Artificial Neural Network;

[b] GBM = Gradient Boosting Machine;

[c] MLR-AL = Multiple Linear Regression with adaptive LASSO;

[d] PLSR = Partial Least Squares Regression.

TABLE B.4 Values of the accuracy for each of the seven water quality sampling sites during training and testing phases based on a 15-feature model. Numbers inside the parentheses represent RMSE values for the testing phase.

	Accuracy			
2013–2015	ANN[a]	GBM[b]	MLR-AL[c]	PLSR[d]
Site36	1.00 (0.60)	1.00 (0.80)	1.00 (0.40)	0.90 (1.00)
Site56	1.00 (0.57)	1.00 (1.00)	0.86 (0.71)	0.64 (0.57)
Site57	0.93 (0.63)	1.00 (0.75)	0.67 (0.50)	0.67 (0.50)
Site73	0.95 (0.40)	1.00 (0.80)	0.90 (0.80)	0.60 (0.40)
Site76	0.97 (0.63)	1.00 (0.75)	0.83 (0.50)	0.66 (0.63)
Site99	0.95 (0.40)	1.00 (0.80)	0.85 (0.60)	0.70 (0.60)
Site100	0.85 (0.00)	1.00 (0.60)	0.80 (0.60)	0.60 (0.40)
2016–2018				
Site36	0.85 (0.67)	0.6 (1.00)	0.70 (0.50)	0.75(0.83)
Site56	1.00 (0.86)	1.00 (1.00)	0.93 (1.00)	0.67 (1.00)
Site57	0.85 (0.71)	1.00 (1.00)	0.95 (0.40)	0.89 (0.71)
Site73	0.95 (0.83)	1.00 (0.60)	0.80 (0.50)	0.95 (0.67)
Site76	0.89 (0.86)	1.00 (1.00)	0.85 (0.86)	0.85 (1.00)
Site99	1.00 (0.83)	1.00 (0.40)	0.90 (0.50)	0.45 (0.67)
Site100	0.95 (0.33)	1.00 (0.83)	0.86 (0.83)	0.76 (0.50)

[a] ANN = Artificial Neural Network;

[b] GBM = Gradient Boosting Machine;

[c] MLR-AL = Multiple Linear Regression with adaptive LASSO;

[d] PLSR = Partial Least Squares Regression.

TABLE B.5 Model predicted fecal coliform concentration (CFUs/100 mL) and probability of exceedance (POE, %) based on the regulatory limit (RL) of 200 CFUs/100 mL and decision value (DV) of 200 CFUs/100 mL for site 36. MEV = mean of the observed values for the 119–192 explanatory variables (MEV) for a given site for each of the two periods (i.e., 2013–2015 and 2016–2018); 1.1 × MEV; 0.90 × MEV; MEV + one standard deviation (SD1); 1.2 × MEV; 0.8 × MEV.

				2013–2015		2016–2018	
Input	Model	Regulatory Standard [CFU/100 mL]	DC [CFU/100 mL]	FIB_predicted [CFU/100 mL]	POE [%]	FIB_predicted [CFU/100 mL]	POE [%]
MEV	ANN	200	200	29987.02	100.00	1.73	48.88
	GBM			inf	100.00	inf	100.00
	MLR-AL			7.32E+12	100.00	0.01	25.07
	PLSR			inf	100.00	inf	100.00
1.1 × MEV	ANN	200	200	67564.26	100.00	6.05	48.91
	GBM			inf	100.00	inf	100.00
	MLR-AL			6.44E+13	100.00	0.03	15.77
	PLSR			inf	100.00	inf	100.00
0.9 × MEV	ANN	200	200	16313.37	99.21	20.62	48.99
	GBM			inf	100.00	inf	100.00
	MLR-AL			8.31E+11	100.00	0.00	33.42
	PLSR			inf	100.00	inf	100.00
MEV + SD1	ANN	200	200	6.24E+17	100.00	2.06	48.89
	GBM			inf	100.00	inf	100.00
	MLR-AL			8.85E+34	100.00	0.01	20.80
	PLSR			inf	100.00	inf	100.00
1.2 × MEV	ANN	200	200	163849.14	100.00	2.28	48.89
	GBM			inf	100.00	inf	100.00
	MLR-AL			5.67E+14	100.00	0.00	28.61
	PLSR			inf	100.00	inf	100.00
0.8 × MEV	ANN	200	200	11673.80	95.71	21.95	49.00
	GBM			inf	100.00	inf	100.00
	MLR-AL			9.44E+10	100.00	0.05	10.20
	PLSR			inf	100.00	inf	100.00

TABLE B.6 Model predicted fecal coliform concentration (CFUs/100 mL) and probability of exceedance (POE, %) based on the regulatory limit (RL) of 200 CFUs/100 mL and decision value (DV) of 200 CFUs/100 mL for site 56. MEV = mean of the observed values for the 119–192 explanatory variables (MEV) for a given site for each of the two periods (i.e., 2013–2015 and 2016–2018); 1.1 × MEV; 0.90 × MEV; MEV + one standard deviation (SD1); 1.2 × MEV; 0.8 × MEV.

				2013–2015		2016–2018	
Input	Model	Regulatory Standard [CFU/100 mL]	DC [CFU/100 mL]	FIB_predicted [CFU/100 mL]	POE [%]	FIB_predicted [CFU/100 mL]	POE [%]
MEV	ANN	200	200	1.05E+20	100.00	16.24	48.85
	GBM			2.78E+87	100.00	8387896.39	100.00
	MLR-AL			253.71	68.40	1022.86	100.00
	PLSR			3.32964E+65	100.00	2.20E+63	100.00
1.1 × MEV	ANN	200	200	1.91E+19	100.00	3.49	48.16
	GBM			2.17E+41	100.00	2.84E+09	100.00
	MLR-AL			153.67	26.84	474.15	100.00
	PLSR			inf	100.00	0.00	0.00
0.9 × MEV	ANN	200	200	1.45322E+17	100.00	24.68	49.10
	GBM			inf	100.00	3.08E+11	100.00
	MLR-AL			44950.77	100.00	126017.92	100.00
	PLSR			2.62E-280	3.38	2.50E+100	100.00
MEV + SD1	ANN	200	200	1.45E+17	100.00	4.36	48.52
	GBM			inf	100.00	615335.83	100.00
	MLR-AL			44950.77	100.00	696.41	100.00
	PLSR			2.62E-280	3.38	3.30E+28	100.00
1.2 × MEV	ANN	200	200	2.45E+20	100.00	71.56	49.32
	GBM			1.56E+189	100.00	9.99E+252	100.00
	MLR-AL			327.05	82.96	1502.32	100.00
	PLSR			7.12E-231	0.00	1.47E+98	100.00
0.8 × MEV	ANN	200	200	6.75E+18	100.00	3.50	48.08
	GBM			3.64E+34	100.00	3.24E+09	100.00
	MLR-AL			118.65	8.50	439.07	100.00
	PLSR			inf	100.00	0.00	0.00

TABLE B.7 Model predicted fecal coliform concentration (CFUs/100 mL) and probability of exceedance (POE, %) based on the regulatory limit (RL) of 200 CFUs/100 mL and decision value (DV) of 200 CFUs/100 mL for site 57. MEV = mean of the observed values for the 119–192 explanatory variables (MEV) for a given site for each of the two periods (i.e., 2013–2015 and 2016–2018); 1.1 × MEV; 0.90 × MEV; MEV + one standard deviation (SD1); 1.2 × MEV; 0.8 × MEV.

				2013–2015		2016–2018	
Input	Model	Regulatory Standard [CFU/100 mL]	DC [CFU/100 mL]	FIB_predicted [CFU/100 mL]	POE [%]	FIB_predicted [CFU/100 mL]	POE [%]
MEV	ANN	200	200	9.80E+17	100.00	8.28	46.41
	GBM			inf	100.00	3.08E+40	100.00
	MLR-AL			155.28	40.35	0.00	47.54
	PLSR			inf	100.00	inf	100.00
1.1 × MEV	ANN	200	200	1.81E+17	100.00	25.45	46.73
	GBM			inf	100.00	6.33E+79	100.00
	MLR-AL			162.43	37.94	0.00	47.54
	PLSR			inf	100.00	inf	100.00
0.9 × MEV	ANN	200	200	2.45E+09	100.00	3.31	46.32
	GBM			inf	100.00	inf	100.00
	MLR-AL			675.65	85.34	0.00	47.54
	PLSR			inf	100.00	inf	100.00
MEV + SD1	ANN	200	200	4.18E+17	100.00	11.72	46.48
	GBM			inf	100.00	7.02E+80	100.00
	MLR-AL			159.20	39.36	0.00	47.54
	PLSR			inf	100.00	inf	100.00
1.2 × MEV	ANN	200	200	2.50E+18	100.00	8.84	46.42
	GBM			inf	100.00	8.35E+58	100.00
	MLR-AL			150.21	40.95	0.00	47.54
	PLSR			inf	100.00	inf	100.00
0.8 × MEV	ANN	200	200	6.98E+16	100.00	81.88	47.79
	GBM			inf	100.00	inf	100.00
	MLR-AL			168.71	37.32	0.00	47.54
	PLSR			inf	100.00	inf	100.00

TABLE B.8 Model predicted fecal coliform concentration (CFUs/100 mL) and probability of exceedance (POE, %) based on the regulatory limit (RL) of 200 CFUs/100 mL and decision value (DV) of 200 CFUs/100 mL for site 73. MEV = mean of the observed values for the 119–192 explanatory variables (MEV) for a given site for each of the two periods (i.e., 2013–2015 and 2016–2018); 1.1 × MEV; 0.90 × MEV; MEV + one standard deviation (SD1); 1.2 × MEV; 0.8 × MEV.

				2013–2015		2016–2018	
Input	Model	Regulatory Standard [CFU/100 mL]	DC [CFU/100 mL]	FIB_predicted [CFU/100 mL]	POE [%]	FIB_predicted [CFU/100 mL]	POE [%]
MEV	ANN	200	200	9.54	49.3059	20.70	3.73E-12
	GBM			inf	100	inf	100.00
	MLR-AL			3.89E+24	100	7.03E-04	38.55
	PLSR			inf	100	inf	100.00
1.1 × MEV	ANN	200	200	154.98	49.8359	11.17	2.89E-30
	GBM			inf	100	inf	100.00
	MLR-AL			4.25E+20	100	5.51E-03	33.19
	PLSR			inf	100	inf	100.00
0.9 × MEV	ANN	200	200	222.12	50.0806	8318.77	100.00
	GBM			inf	100	inf	100.00
	MLR-AL			1.70E+30	100	1.51E-03	42.21
	PLSR			inf	100	inf	100.00
MEV + SD1	ANN	200	200	5.92	49.2927	12.02	1.18E-19
	GBM			inf	100	inf	100.00
	MLR-AL			4.06E+22	100	1.97E-03	36.23
	PLSR			inf	100	inf	100.00
1.2 × MEV	ANN	200	200	18.43	49.3383	44.71	2.62E-06
	GBM			inf	100	inf	100.00
	MLR-AL			3.72E+26	100	2.51E-04	40.34
	PLSR			inf	100	inf	100.00
0.8 × MEV	ANN	200	200	9035.05	79.0211	15.20	2.03E-46
	GBM			inf	100	inf	100.00
	MLR-AL			4.44E+18	100	1.54E-02	29.11
	PLSR			inf	100	inf	100.00

TABLE B.9 Model predicted fecal coliform concentration (CFUs/100 mL) and probability of exceedance (POE, %) based on the regulatory limit (RL) of 200 CFUs/100 mL and decision value (DV) of 200 CFUs/100 mL for site 99. MEV = mean of the observed values for the 119–192 explanatory variables (MEV) for a given site for each of the two periods (i.e., 2013–2015 and 2016–2018); 1.1 × MEV; 0.90 × MEV; MEV + one standard deviation (SD1); 1.2 × MEV; 0.8 × MEV.

				2013–2015		2016–2018	
Prediction	Model	Regulatory Standard [CFU/100 mL]	DC [CFU/100 mL]	FIB_predicted [CFU/100 mL]	POE [%]	FIB_predicted [CFU/100 mL]	POE [%]
I	ANN	200	200	324.40	50.28	6.89	50.00
	GBM			inf	100.00	4.15E+15	100
	MLR-AL			inf	100.00	206856.94	100.00
	PLSR			inf	100.00	inf	100
II	ANN	200	200	86.11	49.74	18.35	50.00
	GBM			inf	100.00	inf	100
	MLR-AL			inf	100.00	54882.60	100.00
	PLSR			inf	100.00	0.00E+00	49.81928731
III	ANN	200	200	1.31E+10	100.00	6.16	50.00
	GBM			inf	100.00	inf	100
	MLR-AL			inf	100.00	4885115938.59	100.00
	PLSR			inf	100.00	inf	100
IV	ANN	200	200	155.07	49.90	8.74	50.00
	GBM			inf	100.00	6.36577E+60	100
	MLR-AL			inf	100.00	106549.74	100.00
	PLSR			inf	100.00	inf	100
V	ANN	200	200	965.13	51.72	7.28	50.00
	GBM			inf	100.00	1.3655E+207	100
	MLR-AL			inf	100.00	401594.52	100.00
	PLSR			0	47.74	inf	100
VI	ANN	200	200	38.48	49.64	42.20	50
	GBM			inf	100.00	inf	100
	MLR-AL			inf	100.00	28269.43	100.00
	PLSR			inf	100.00	0	49.77959613

TABLE B.10 Model predicted fecal coliform concentration (CFUs/100 mL) and probability of exceedance (POE, %) based on the regulatory limit (RL) of 200 CFUs/100 mL and decision value (DV) of 200 CFUs/100 mL for site 100. MEV = mean of the observed values for the 119–192 explanatory variables (MEV) for a given site for each of the two periods (i.e., 2013–2015 and 2016–2018); 1.1 × MEV; 0.90 × MEV; MEV + one standard deviation (SD1); 1.2 × MEV; 0.8 × MEV.

				2013–2015		2016–2018	
Prediction	Model	Regulatory Standard [CFU/100 mL]	DC [CFU/100 mL]	FIB_predicted [CFU/100 mL]	POE [%]	FIB_predicted [CFU/100 mL]	POE [%]
I	ANN	200	200	19.30	30.43	28.41	39.92
	GBM			4.95E-92	0.00	5.39E+284	100.00
	MLR-AL			3.15E-13	33.66	0.00	49.94
	PLSR			inf	100	inf	100
II	ANN	200	200	137.21	39.52	3.04	38.46
	GBM			9.84E+102	100.00	2.13E+150	100.00
	MLR-AL			1.73E-10	26.43	0.00	49.91
	PLSR			inf	100	inf	100
III	ANN	200	200	747.90	72.16	128.84	45.78
	GBM			inf	100.00	inf	100.00
	MLR-AL			5.21E-22	41.48	0.00	49.97
	PLSR			inf	100.00	inf	100.00
IV	ANN	200	200	15.68	26.37	5.89	38.63
	GBM			2.76E+74	100.00	9.86E+185	100.00
	MLR-AL			7.38E-12	30.49	0.00	49.93
	PLSR			inf	100	inf	100
V	ANN	200	200	157.35	45.96	156.91	47.44
	GBM			4.77E-78	0.00	inf	100.00
	MLR-AL			1.34E-14	36.15	0.00	49.95
	PLSR			inf	100.00	inf	100.00
VI	ANN	200	200	1342.64	100.00	3.34	38.48
	GBM			6.38E+142	100.00	1.91E+68	100.00
	MLR-AL			4.07E-09	21.26	0.00	49.88
	PLSR			inf	100.00	inf	100.00

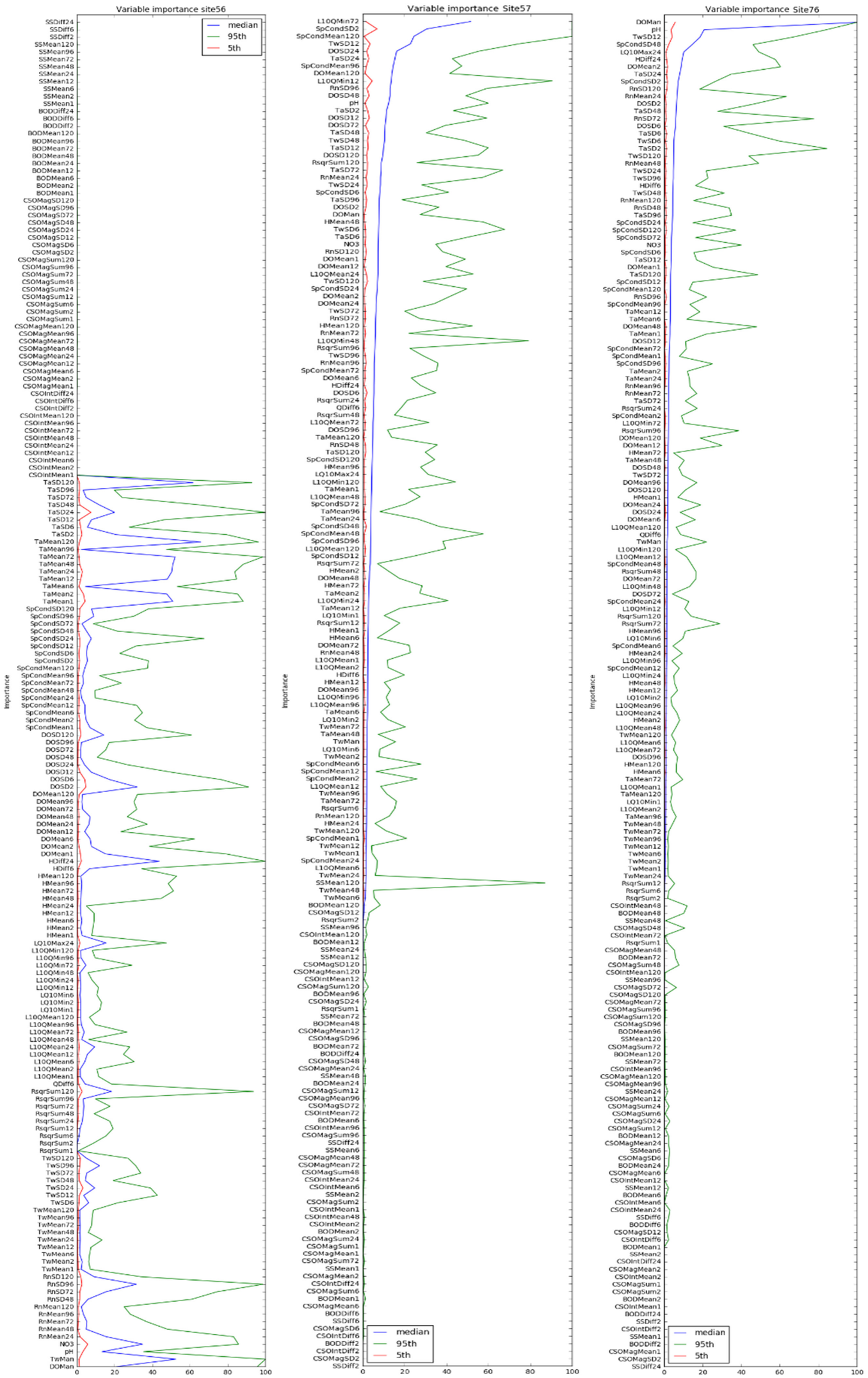


FIGURE B.1 Sample plots of the long list of explanatory variables for Fecal from the most to the least relevant variable used for the dimensionality reduction step for each site.

This page intentionally blank

APPENDIX C: CAWS DUFLOW MODEL FINAL REPORT

By Naila Mahdi
Feburuary 2020

C.1 DUFLOW MODELING

The hydraulic modeling for the years 2014–2018 was developed and completed using the calibrated and verified base hydraulic model for 2013. The base model was built on earlier work of Dr. Charles Melching (CAWS DuFlow model). The 2007–2008 CAWS DuFlow model was provided by the MWRD. The original model network used the Chicago Sanitary and Ship Canal (CSSC) at Romeoville, IL (USGS gauging station number 05536995) as a downstream boundary condition. However, the construction at the U.S. Army Corps of Engineers of an electric fish barrier in 2006 led to cessation of streamflow monitoring at the gauge. Therefore, the gauge on CSSC near Lemont, IL (05536890) was used as the new downstream boundary in this project (Figure 13).

Accordingly, the DuFlow network was modified to represent the new boundary conditions. The hydraulic model accounted for the CAWS system's major inflows and outflows. Major inflows are influenced by control structures, pumping stations, tributaries, and CSO discharges. Boundary conditions for model verification and use were set using data collected by the USGS and the MWRDGC at the three lake front control structures, the Lockport Controlling Works and the major and minor tributary flows. The control structures include Wilmette Pumping Station, Chicago River Controlling Works, and T.J. O'Brien Lock and Dam. Pumping stations include the Racine Avenue Pumping Station, NB pumping station and 125th pumping stations. Major tributaries include the North Branch Chicago River (NBCR) at Albany Ave. (USGS removed this in April 2018 so the site at N. Pulaski was considered for 2018 period), Little Calumet (LC), Grand Calumet (USGS ceased recording discharge at this site in 2016 so stage data was estimated for the years 2016, 2017, and 2018 using previous records), as well as Calumet, Stickney, and O'Brien WRP effluents. While minor tributaries include Tinley, Midlothian, Mill, Navajo, Natalie, and East and West Stony creeks as long as 44 representative CSO discharge points.

Local temporal relationships were established using data recorded immediately before and after the missing data to determine missing stage and flow data. It is necessary to estimate the inflows from ungaged tributary watersheds. Flows on Midlothian Creek were used to estimate flows on ungaged tributaries on an area-ratio basis (discharge for the ungagged tributary watersheds were estimated using area ratio of the total area to the gaged drainage area for Midlothian and Tinley Creeks (Alp, E., and Melching, C.S. 2006). The drainage area ratios for the ungaged tributaries compared to the Midlothian Creek drainage area are listed in Table C.1.

TABLE C.1 Drainage area ratios for the ungaged tributaries compared to the Midlothian Creek drainage area.

Stream Ungaged	Ratio with Midlothian*
Mill Creek West	0.55
Stony Creek West	1.086
Cal-Sag Watershed East	0.246
Navajo Creek	0.137
Stony Creek East	0.486
Ungaged Des Plaines Watershed	0.703
Calumet Union Ditch	1.168
Cal-Sag Watershed West	0.991

* The gaged Midlothian Creek drainage area is 12.6 mi2, but these ratios are computed to the total Midlothian Creek drainage area of 20 mi2. The total flow for both Midlothian and Tinley Creeks was determined by area ratio of the total drainage area to the gaged drainage area, 12.6 mi2 and 11.2 mi2 for Midlothian and Tinley Creeks, respectively.

Upstream and downstream boundary conditions included the North Shore Channel at Wilmette (05536101), Chicago River Main Stem at Columbus Drive (05536123), Calumet River at the T.J. O'Brien Lock and Dam (05536358), Racine Avenue Pump Station (RAPS), Little Calumet River at South Holland (05536290), and CSSC near Lemont (05536890). Stage (H) and discharge data (Q) were provided by MWRD, as was CSO event data. The approximately 200 annual CSO events during 2014–2018 across CAWS were represented by a system of 44 discharge points for each respective year. Flow at the locks was estimated by aggregating discharge hourly data due to navigation, blockages, leakages, and discretionary diversions. All stage data was referenced to the City of Chicago Datum of 579.48 ft.

C.2 MODEL VALIDATION

The stage data on CSSC near Lemont was used as the only boundary condition. Therefore, the flow data at this station was used to validate the accuracy of the model. Model performance was evaluated using the Nash-Sutcliffe efficiency ($-\infty \leq NSE \leq 1$), the linear regression coefficient of determination ($0 \leq R^2 \leq 1$), and the percent bias ($-100 \leq PBIAS \leq 100$). The optimal value is 1 for NSE and R^2 and 0 for PBIAS. A positive or negative PBIAS is indicative of model under- or overprediction (Moriasi 2007).

C.3 RESULTS

C.3.1 Validation of DuFlow Streamflow Simulations

Figure C.1, Figure C.2, Figure C.3, Figure C.4 and Figure C.5 visually compare streamflow observations (USGS data) and DuFlow simulations for a gage on the CSSC near Lemont, IL for the years 2014, 2015, 2016, 2017, 2018 respectively. The graphs and model performance metrics are indicative of DuFlow's ability to capture both the magnitude and sequence of flows at this gaged station. To illustrate, in Figure C.5, for the modeling year 2018, the coefficient of determination of 0.83 (R^2=0.83) indicates that the model captured 83% of the hourly streamflow variability. A PBIAS of 0.005% indicates that, on average, the model perfectly estimates hourly streamflow. A NSE of 0.80 indicates that the model performance is considered efficient.

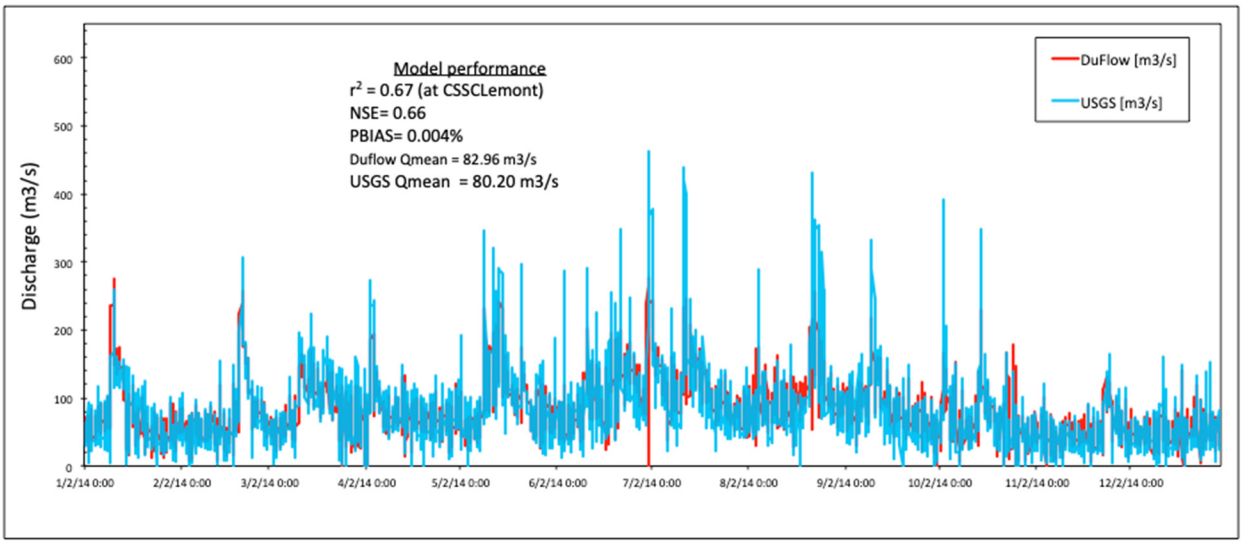


FIGURE C.1 Observed (USGS) and simulated (DuFlow) streamflow for the year 2014 on CSSC near Lemont (USGS gaging station 05536890). The simulation results are at an hourly time step. The bottom two graphs are magnifications of the two shaded regions in the top graph.

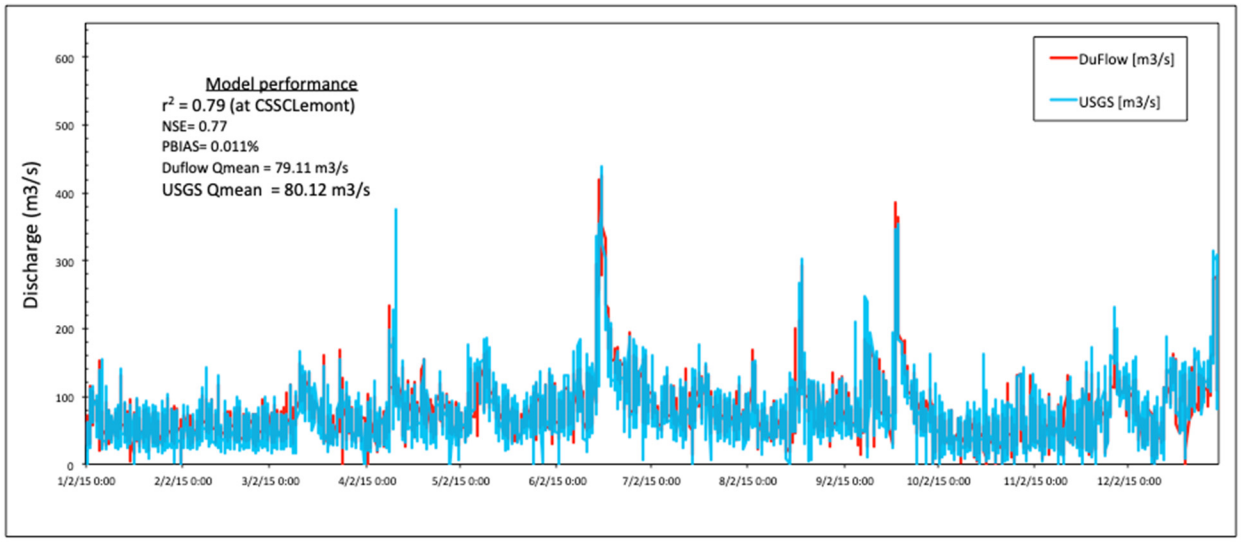


FIGURE C.2 Observed (USGS) and simulated (DuFlow) streamflow for the year 2015 on CSSC near Lemont (USGS gaging station 05536890). The simulation results are at an hourly time step. The bottom two graphs are magnifications of the two shaded regions in the top graph.

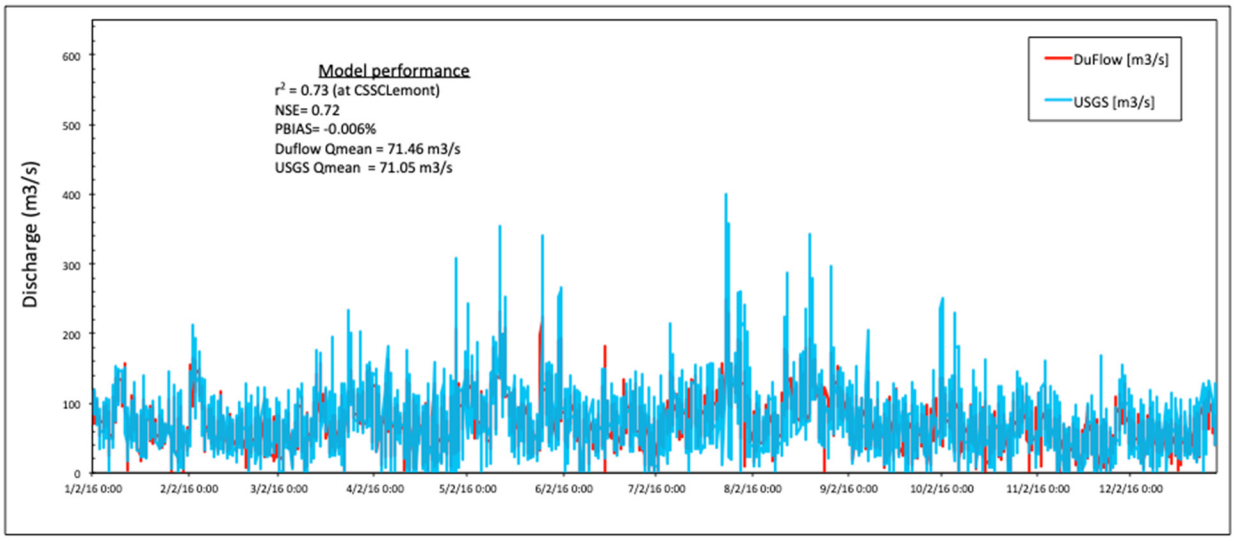


FIGURE C.3 Observed (USGS) and simulated (DuFlow) streamflow for the year 2016 on CSSC near Lemont (USGS gaging station 05536890). The simulation results are at an hourly time step. The bottom two graphs are magnifications of the two shaded regions in the top graph.

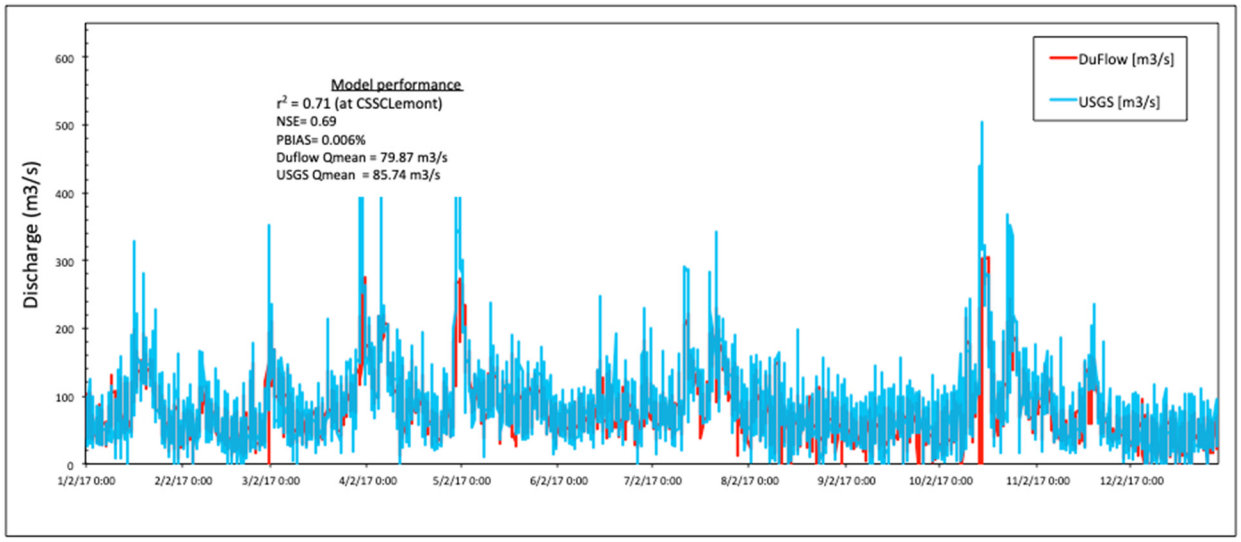


FIGURE C.4 Observed (USGS) and simulated (DuFlow) streamflow for the year 2017 on CSSC near Lemont (USGS gaging station 05536890). The simulation results are at an hourly time step. The bottom two graphs are magnifications of the two shaded regions in the top graph.

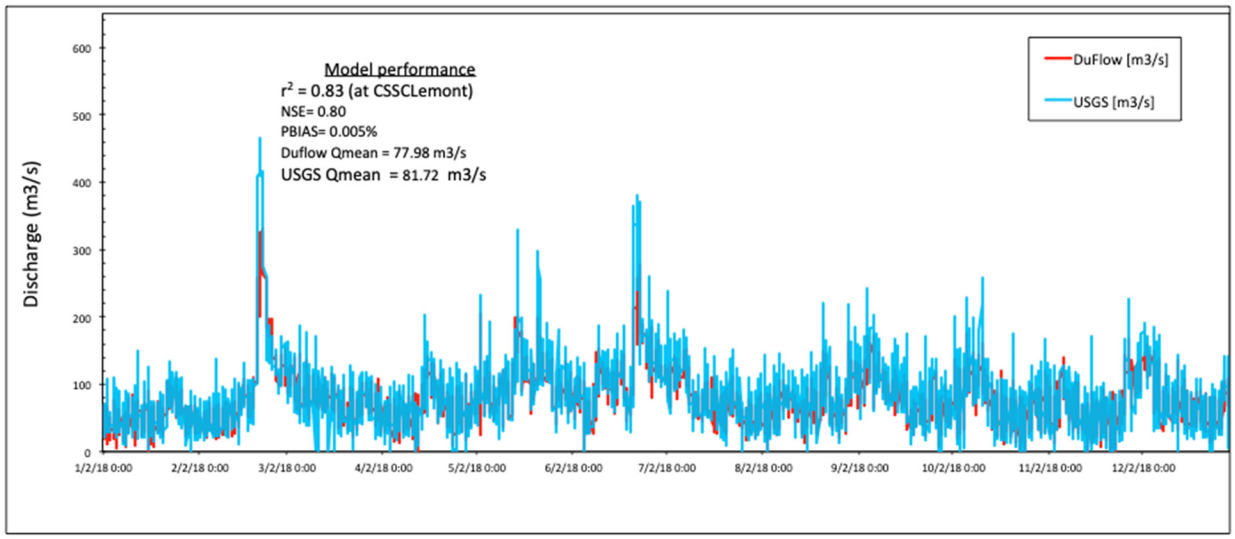


FIGURE C.5 Observed (USGS) and simulated (DuFlow) streamflow for the year 2018 on CSSC near Lemont (USGS gaging station 05536890). The simulation results are at an hourly time step. The bottom two graphs are magnifications of the two shaded regions in the top graph.

C.4 REFERENCES

Alp, E., and Melching, C. S. 2006. "Calibration of a model for simulation of water quality during unsteady flow in the Chicago Waterway System and application to evaluate use attainability analysis remedial actions." Technical Rep. No. 18, Institute of Urban Environmental Risk Management, Marquette Univ., Milwaukee.

This page intentionally blank

APPENDIX D: SAMPLE COLLECTION AND PROCESSING PROTOCOLS

D.1 WATER AND SEDIMENT COLLECTION AND PROCESSING PROTOCOL

Sampling days (2nd, 3rd and 4th Monday of each month):

- 250 mL and 1000mL sterile Nalgene Polypropylene Copolymer (PPCO) bottles for water sample collection.
- 1 quart glass bottles for sediment collection for each sample.
- Whirl Pak bags for each sample.
- Sterile trays for each sample.
- Sterile scoops for each sample.
- Chain of Custody forms for both water and sediment samples. (See the attached electronic copies)

Filtration equipment:

- Vacuum Filter Manifold
- Sterile Filter base and funnel
- Sterile Millipore Polyvinylidene difluoride (PVDF), 0.22µm pore size, 47 mm diameter membrane filters. (will change to sterile Mixed Cellulose Ester (MCE), 0.22µm pore size, 47 mm diameter membrane from 4/14/14)
- Forceps

Procedure to follow on the sampling day:

For water/outfall samples:
QC/Blank:

- Transfer one sterile filter aseptically to the 50 mL tube as a filter blank with each batch of waterways samples filtered. (Please note that this could be the source of contamination as we used the same forceps to handle both waterways sample and a blank filter. We are going to keep separate forceps for blank in 2014.)
- Equipment blank sample will be received in 250mL bottle on each sampling day. Filter **200**mL sample and transfer to the tube, label and freeze it.

- Trip blank bottle with 100mL sterile milliq water will be sent with sampling crew on each sampling event. When received in the lab **100** mL will be filtered and filtered will be saved in the freezer.

- Receive 1000 mL water sample for each sampling location, check the labels, COCs.

- Filter **200** mL of water from each sample in duplicate using sterile membrane filter. Aseptically remove filter and put in sterile 50 mL blue cap tube. Label the tube with the corresponding LIMS #, sample location and date and freeze at -80°C immediately.

- Record turbidity, conductivity and pH of all water samples collected.

- Every Monday you will receive Calumet WRP and every Tuesday O'Brien WRP outfall sample in 1000 mL bottle, check COCs and filter **200** mL of sample in duplicate and freeze the filters in 50 mL tube at -80°C.

For Wet/Dry Weather samples:

- Filter **200** mL samples in duplicate and freeze the filters in 50 mL sterile tube at -80°C. **If the sample is very turbid and having problem filtering, then just filter the sample for 30 minutes and then remove the filter and freeze it. Record the sample volume filtered on the tube.**

For sediment samples:

- Receive 1 Quart bottle and 1 Whirl Pak bag of sediment sample for each location, check the labels and COCs.

- Receive all the samples in the LIMS. If there is no sample collected immediately cancel the sample from the LIMS.

- Aliquot approximately 1/3 of the sample from the quart bottle into the 1 plastic half quart bottles. Attach the LIMS label for **Hg** analysis for that sample.

- Attach all other LIMS labels (for Solids, Lachat (NH3/TKN), Metals) to 1 quart glass bottle with rubber band and deliver both bottles to ALD- Log in area along with COCs.

- Make copies of the COCs.

- Take out approximately 20 wet gram of sediment sample from the Whirl Pak bag into a sterile 50 mL blue cap tube, label it with corresponding LIMS #, sample location and date and freeze it immediately at -80°C immediately.

- Use the remaining sample from Whirl Pak bag to carry out bacterial analysis by Colilert method.

- Next day autoclave the whirl pak bag and discard them.

For Raw Sewage samples: (from July 2014)

- Last Monday of every month you will receive raw sewage sample from Calumet WRP and second Tuesday of each month O'Brien WRP Raw sample in 1000 mL bottle, check COCs and filter **30** mL of sample in duplicate and freeze the filters in 50 mL tube at -80°C.

D.2 COLLECTION AND SUBMISSION OF FISH SAMPLES

PURPOSE: The purpose of this protocol is to describe the Argonne National Laboratory recommended procedures for collecting and transporting fish sample collection to Analytical Microbiology Laboratory for metagenomic testing. These samples include whole fish and/or surface swabs. These samples are obtained from CAWS.

RESPONSIBILITIES: It is the responsibility of Analytical Microbiology Laboratory personnel to advise Aquatic Ecology and Water Quality section staff collecting and transporting samples; as to these recommended procedures. Any deviations from this protocol should be reported on the chain of custody sheet.

SUPPLIES:

- Sterile 15ml falcon tube
- Sterile cotton tipped swabs with wooden or plastic shafts
- Sterile Whirl-pak bag
- Gloves
- Chain of Custody Sheets (COC)

PROCEDURE:

1. Use of personal protective equipment and precautions is recommended when collecting any fish specimen.

2. For fish mucous sampling,

   2.1 Fish should be caught and handled with clean gloves.

   2.2 A dry sterile cotton swab should be swabbed against the fish for 10 seconds rotating the cotton swab over both sides of the fish including swabbing their mouths.

2.3 The sterile swab should be immediately placed into a sterile 15ml falcon tube. No buffer is needed to preserve the sample.

2.4 If the wooden shaft is longer than the tube, break it off by bending it against the side of the tube.

2.5 Cap the tube and then properly label with fish identification and collection date.

2.6 Place the tubes in a sterile whirl-pak and seal properly to avoid any contamination.

2.7 Place the whir-pak bag in a plastic bag. All tubes should be securely packed and sealed to prevent leakage that may occur when kept in a cooler filled with ice.

2.8 Ensure that the tubes are not in direct contact with ice or water. Ice must be contained in a separate strong, plastic bag. Use generous quantities of ice in the cooler.

3. Check and ensure the COC information for the following:

   3.1 Sample Collection ID, Date is identical to the sample tubes.

   3.2 Sample collector name and initial

4. Keep COC related paperwork dry and separate from fish sample specimens.

5. Use a new glove for each location and each fish swab collection.

6. Samples transported to Analytical Microbiology Laboratory will be immediately checked and will be kept frozen at -80°C freezer until picked up by the Argonne Laboratory staff.