ANL/EVS-18/6

Phase II Interim Report: Chicago Area Waterway System Microbiome Research

April 2013 – December 2017

Environmental Science Division

About Argonne National Laboratory
Argonne is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC under contract DE-AC02-06CH11357. The Laboratory's main facility is outside Chicago, at 9700 South Cass Avenue, Argonne, Illinois 60439. For information about Argonne and its pioneering science and technology programs, see www.anl.gov.

DOCUMENT AVAILABILITY

Online Access: U.S. Department of Energy (DOE) reports produced after 1991 and a growing number of pre-1991 documents are available free via DOE's SciTech Connect (http://www.osti.gov/scitech/).

Reports not in digital format may be purchased by the public from the National Technical Information Service (NTIS):
U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
www.ntis.gov
Phone: (800) 553-NTIS (6847) or (703) 605-6000
Fax: (703) 605-6900
Email: **orders@ntis.gov**

Reports not in digital format are available to DOE and DOE contractors from:
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062

Disclaimer
This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor UChicago Argonne, LLC, nor any of their employees or officers, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of document authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, Argonne National Laboratory, or UChicago Argonne, LLC.

ANL/EVS-18/6

Phase II Interim Report: Chicago Area Waterway System Microbiome Research

April 2013 – December 2017

Principal Investigators
M. Cristina Negri[1]
Jack A. Gilbert[2]
Mark Grippo[1]

Research Team
Anukriti Sharma[3]
Melissa Dsouza[3]
Jules F. Cacho[1]
Patty Campbell[1]

[1] Environmental Science Division, Argonne National Laboratory
[2] Group Leader for Microbial Ecology, Argonne National Laboratory
[3] University of Chicago

Prepared for
Metropolitan Water Reclamation District of Greater Chicago

December 2018

CONTENTS

NON-TECHNICAL SUMMARY ix

EXECUTIVE SUMMARY xxi

1 INTRODUCTION 1

2 16S RRNA-BASED ANALYSIS OF RIVER WATER AND SEDIMENT SAMPLES COLLECTED POST DISINFECTION AT CALUMET AND O'BRIEN WATER RECLAMATION PLANTS 3

2.1 Introduction 3
2.2 Materials and Methods 4
2.2.1 Assessing Microbial Community Structure in CAWS Samples Over Five Years Using 16S rRNA Amplicon Gene Sequencing 4
2.2.2 Amplicon Based Microbial Community Sequencing Analysis 6
2.2.3 16S rRNA Gene Sequence Analyses 7
2.2.4 Statistical Analyses 7
2.2.5 Assessing Microbial Community Structure and Function Across the CAWS Using Shotgun Metagenomic Sequence Data 7
2.3 Results 8
2.3.1 Alpha- and Beta- Diversity Comparison of the CAWS Samples from 2013-2017 8
2.3.2 Compositional Variation Among River Water and Sediment Samples Collected Pre- and Post-Disinfection 13
2.3.3 Determining the Sources of Microbial Community in the CAWS 20
2.3.4 Bioball Experiment 22
2.3.5 Taxonomic and Functional Annotations of Metagenomic Samples 25
2.4 Conclusions 30
2.5 References 32

3 CHICAGO AREA WATERWAY SYSTEM FECAL INDICATOR BACTERIA MODEL DEVELOPMENT 39

3.1 Introduction 39
3.2 Materials and Methods 40
3.2.1 Selected Sampling Sites Used for Model Development 40
3.2.2 The CAWS-FIB Conceptual Modeling Framework 40
3.2.3 GBM Model Development 43
3.3 Results 47
3.3.1 Model Training and Testing 47
3.3.2 Model Prediction 51
3.4 Limitations of the Current Modeling Work and Future Work 52
3.5 References 53

APPENDIX A 55

FIGURES

1 CAWS and Water Reclamation Plant Sampling Locations xi

2 The Sewage and Wastewater Microbes Decrease Significantly Post-Disinfection Compared to Prior Condition without Disinfection xiii

3 Geometric Mean Fecal Coliform Concentrations Observed March-November in Final Effluent, Upstream and Downstream of O'Brien WRP, North Shore Channel and North Branch Chicago River Locations xv

4 Geometric Mean Fecal Coliform Concentrations Observed March-November in Final Effluent, Upstream and Downstream of Calumet WRP, Grand Calumet, Little Calumet, and Cal-Sag Channel River Locations xv

5 CAWS Microbial Community Sources Using Earth Microbiome Project Database xvii

1 The Alpha and Beta Diversity Analyses of CAWS Samples Collected from 2013-2017 9

2 Alpha Diversity Analyses for Sediment and Water Samples Collected at Different Sites Over a Period of Five Years 10

3 Shannon Alpha Diversity Indices of Sewage, Effluent, River Water and Sediment Samples from Calumet and O'Brien WRPs Over a Period of Five Years 12

4 Geometric Mean Fecal Coliform Concentrations Observed March-November in Final Effluent, Upstream and Downstream of O'Brien WRP, North Shore Channel and North Branch Chicago River Locations 14

5 Geometric Mean Fecal Coliform Concentrations Observed March-November in Final Effluent, Upstream and Downstream of Calumet WRP, Grand Calumet, Little Calumet, and Cal-Sag Channel River Locations 14

6 Non-Parametric Two Group Tests Done Between Pre- and Post-Disinfection River Water and Sediment Samples Downstream of the Two Water Reclamation Plants i.e. Calumet WRP and O'Brien WRP 17

7 Non-Parametric Two Group Tests Done between the Two Disinfection Years i.e. 2016 and 2017 for River Water and Sediment Samples Downstream of the Calumet WRP and O'Brien WRP 21

FIGURES (Cont.)

8 SourceTracker 2.0 Analysis of Water Column Samples by Sampling Site for Years 2013-2017 Using a Curated Database for Calumet, O'Brien, and Main..... 22

9 Heatmap Showing Distribution of 10,000 Sequence Reads Assigned to the 10 Most Abundant Bacterial Genera by Sample..... 24

10 Heatmap Showing Distribution of 1000 Sequence Reads Assigned to the 10 Most Abundant Bacterial Genera by Sample..... 24

11 Heatmap Showing the Relative Abundance of the 30 Most Variable Bacterial Genera Distributed Across the Three Upstream and Three Downstream Sites from the Calumet WRP..... 27

12 Two-Group Tests Performed to Identify Statistically Differential Bacterial Genera between Sites, Upstream and Downstream of the Calumet WRP..... 28

13 Heatmap Showing Relative Abundance of the SEED Subsystems Annotated from the 24 Shotgun Samples..... 29

14 The 12 Sampling Sites to Be Used for Model Development..... 41

15 Schematic of the CAWS FIB Modeling Process..... 42

16 A Schematic of a Scatter Plot of Observed and Predicted Fecal Values During Model Training and Testing Showing False Positives, True Positives, True Negatives, and False Negatives..... 46

17 Plots of the 15 Most Relevant Explanatory Variables for Site 56, Site 57, and Site 76..... 48

18 Plots of Deviance for Site 56, Site 57, and Site 76 between Training and Testing of Models that Include 15 Features, 10 Features, and 5 Features..... 49

A.1 A Sample Plot 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..... 59

A.2 Predicted vs. Observed Plot of Fecal Values for Site 56 with Regulatory Limit of 200 CFU/100 mL and Decision Value of 200 CFU/100 mL in Both Model Training and Testing and Using 15-, 10- and 5-Most Relevant Explanatory Variables..... 60

FIGURES (Cont.)

A.3 Predicted vs. Observed Plot of Fecal Values for Site 57 with Regulatory Limit of 200 CFU/100 mL and Decision Value of 200 CFU/100 mL in Both Model Training and Testing and Using 15-, 10- and 5-Most Relevant Explanatory Variables 61

A.4 Predicted vs. Observed Plot of Fecal Values for Site 76 with Regulatory Limit of 200 CFU/100 mL and Decision Value of 200 CFU/100 mL in Both Model Training and Testing and Using 15-, 10- and 5-Most Relevant Explanatory Variables 62

TABLES

1 Total Number of Samples Collected per Sample Type from 2013 to 2017..... 5

2 Details of Location for Each Site for the Two Water Reclamation Plants 5

3 Summary of Sediment and Water Column Samples by Sites on the CAWS from 2013-2017 6

4 List of Significantly Differential Genera between the Pre- and Post-Disinfection Period across CAWS Water Samples Collected at Downstream of the Calumet WRP 15

5 List of Significantly Differential Genera between the Pre- and Post-Disinfection Period across CAWS Sediment Samples Collected Downstream of the Calumet WRP 15

6 List of Significantly Differential Genera between the Pre- and Post-Disinfection Period across CAWS Water Samples Collected at Downstream of the O'Brien WRP 16

7 List of Significantly Differential Genera between the Pre- and Post-Disinfection Period across CAWS Sediment Samples Collected at Downstream of the O'Brien WRP 16

8 Differentially Abundant Genera Between O'Brien WRP and Calumet WRP Downstream Water Samples Before Disinfection 19

9 Differentially Abundant Genera Post-Disinfection 20

10 Summary of Sequence Depth per Sample 23

TABLES (Cont.)

11 Summary of Samples Chosen with Reference to Calumet WRP for Deep Metagenome Sequencing 25

12 Summary of Sampling Sites with Fecal Data During the Pre-Disinfection Period 40

13 Values of the Model Training and Testing Performance Metrics for Site 56 50

14 Values of the Model Training and Testing Performance Metrics for Site 57 50

15 Values of the Model Training and Testing Performance Metrics for Site 76 50

16 Model Predicted Fecal Coliform Concentration and Probability of Exceedance Based on the Regulatory Limit of 200 CFUs/100 mL and Decision Value of 200 CFUs/100 mL for Site 56 51

17 Model Predicted Fecal Coliform Density and Probability of Exceedance Based on the Regulatory Limit of 200 CFUs/100 mL and Decision Value of 200 CFUs/100 mL for Site 57 51

18 Model Predicted Fecal Coliform Density and Probability of Exceedance Based on the Regulatory Limit of 200 CFUs/100 mL and Decision Value of 200 CFUs/100 mL for Site 76 52

A.1 Variables and the Corresponding Summary Statistics Used in Predicting FIB at Each Sampling Point 57

A.2 General Description of the Hypothetical Dataset Used in Model Prediction to Showcase Model Predictive Functionality and Other Computations 58

This page intentionally left blank.

NON-TECHNICAL SUMMARY

This report summarizes the first five years (2013-2017) of a novel, seven-year exploration of the Chicago Area Waterway System (CAWS) microbiome documenting the microbial community changes with the Metropolitan Water Reclamation District of Greater Chicago's (MWRD) significant improvement efforts (disinfection and storm water reservoir control management). The study, conducted by Argonne National Laboratory, examines CAWS microbial communities (microbiome) prior to and following disinfection treatment of secondary treated effluents at the O'Brien (UV) and Calumet (chlorine and dechlorination) Water Reclamation Plants (WRPs) and the phased implementation of the Tunnel and Reservoir Plan (TARP). The Thornton Composite Reservoir (TCR) was completed in 2015. It 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 overflow into CAWS in rainy weather. In addition to the effects of TARP and disinfection, we present an analysis of potential sources of the different type of microbes found in the CAWS obtaining highly detailed gene based information (16S rRNA amplicon sequencing) to characterize microbial community abundance and variability as a function of location, season, and environmental conditions.

The results of sequencing data from samples collected from 2013 to 2017 indicate thus far:

- The CAWS has greater than twenty thousand species of microbes in the water and sediment.

- Compared to the pre-disinfection period (2013-2015), the final effluent from both the Calumet and O'Brien WRPs and river water samples immediately downstream of the WRPs demonstrated a significant decrease in microbial taxa that are generally associated with sewage.

- Fecal coliform bacteria levels at sites downstream of the T.J. O'Brien WRP and the Calumet WRP showed reduction in the post-disinfection period (2016-2017) compared to pre-disinfection period (2013 to 2015).

- Microbes that have ability to cause disease appeared at low levels across all the samples upstream and downstream of WRPs.

- Sources of microbial diversity across all river water samples can be largely attributed to effluent, sewage, CAWS sediment, freshwater, and fish associated samples. North, Main, and Calumet, have a unique compilation of potential sources that best explain the microbial signatures in those regions. The contribution made by human fecal matter across all water column samples was extremely low. However, there remains a large proportion of bacterial taxonomic diversity in the CAWS that cannot be reliably attributed to a 'source'.

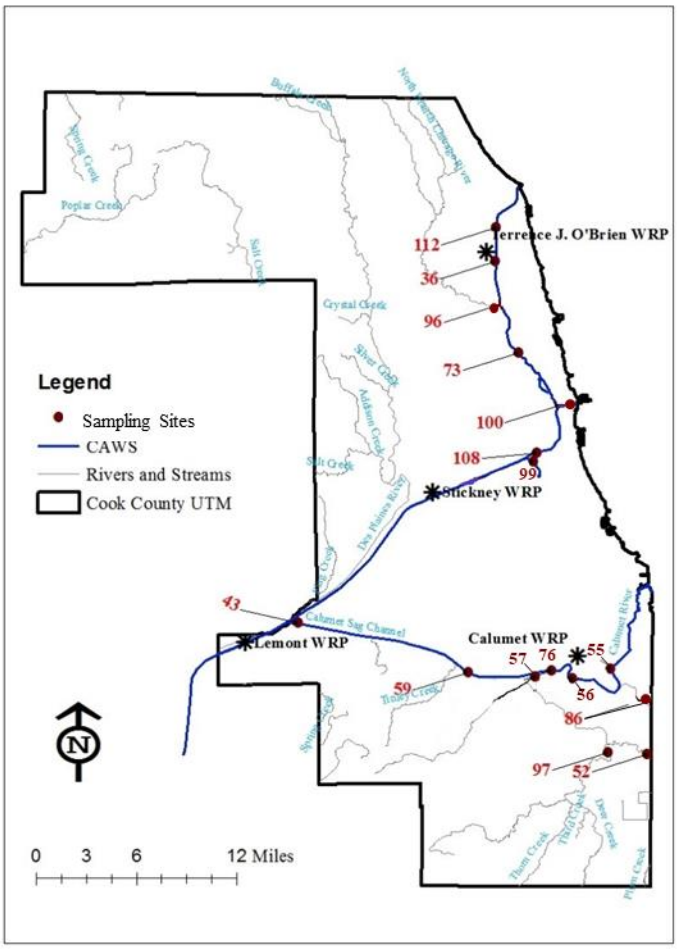
The report also describes the preliminary testing of the CAWS-Fecal Indicator Bacteria (FIB) model. The CAWS-FIB model uses Gradient Boosting Machine (GBM), a machine learning algorithm, to relate FIB concentration to weather, water chemistry, and hydrology related variables. The model can be used to predict FIB concentrations at any point along the CAWS for a given set of predetermined environmental variables. The CAWS-FIB model is being tested to assess the influence of physical and chemical water quality parameters, effluent from WRPs, direct storm water runoff, and combined sewer overflows (CSOs) on the microbial communities in the CAWS. Preliminary results indicate that overfitting (relates to complex river parameters corresponding closely to particular set of data, and may therefore fail to fit additional data or predict future observations) was a problem for the CAWS-FIB model, although predicted fecal concentrations were within the range of concentrations observed during 2013-2015, suggesting the model produced reasonable estimates. Model accuracy may improve after additional sites are incorporated into the model training and testing.

INTRODUCTION TO MICROBIAL METAGENOMICS APPROACHES

Traditional laboratory-culture methods such as FIB counts and select pathogen Polymerase Chain Reaction (PCR)-based methods have 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. Microbiome gene sequencing can supersede, for qualitative analyses, typical culture-based methods that currently only detect approximately 8% of known microbes. Metagenomics-based sequencing can capture all genes present in a microbial community giving insight into the functional potential of microbes present in that sample, including their ability to cause disease, something that plate counts do not provide.

This novel gene study seeks to provide the following information, 1) a description of the diversity of the CAWS microbial community and the impact of MWRD's improvement efforts (disinfection/TARP) on bacterial diversity in final disinfected effluent (effluent) and river water samples downstream of the Calumet and O'Brien WRPs, 2) and functional characteritics of the CAWS microbial communites, 3) the potential sources of microbes at different points in the CAWS. 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).

We extensively sampled CAWS river water and sediment as well as treated effluent discharged from two WRPs over the course of three years (2013-2015) prior to and two years (2016-2017) following the implementation of new disinfection processes and the phased implementation of TARP (Fig. 1). We used gene-based sequencing to determine the particular microbial species present in the CAWS (i.e., "Who is there?"). A total of 2,077 samples collected from WRPs and CAWS river water and sediment were characterized using high-throughput sequencing.



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 Water Reclamation Plant (*) Sampling Locations. Tributaries feeding to CAWS are in bold and upstream sites are in italic fonts.

RESULTS

Diversity of the CAWS Microbial Community

As described in the Phase I report, metagenomic analyses of CAWS samples from the pre-disinfection period (2013-2015), microbial communities showed a distinct distribution pattern across different sampling locations (biogeography), in which the main differentiator was the sample type (river water, sediment, effluent, etc.). These communities appear to be stable (in their diversity and composition) across these sampling years and sampling seasons. Our analysis also showed that microorganisms associated with final WRP effluent (included human fecal or sewage contamination indicators like *Bifidobacterium* and *Acinetobacter)* from secondary-treatment can be tracked downstream and typically showed increased abundance in proximity to the treated final effluent discharge location.

Whilst microbial composition was relatively stable across sample types during the pre-disinfection period from 2013 to 2015, for both river water and sediment samples collected across the CAWS, we observed a significant decrease ($p < 0.05$) in alpha diversity (number of taxa within a single sample) in 2016 (first year post disinfection) when compared to 2015 followed by a significant increase ($p < 0.05$) in 2017 when compared to 2016. Similar patterns were found for river water samples collected immediately downstream of the Calumet WRP (Site 76) and in sediment samples immediately downstream of O'Brien WRP (Site 36).

However, for the final treated effluent and influent raw sewage, we observed a different pattern. At Calumet WRP, there was significant increase in microbial diversity in 2016 effluent samples compared to 2015, and then a significant decrease in 2017 compared to 2016. At O'Brien WRP, there was a significant decrease in microbial diversity from 2016 to 2017 in influent (raw sewage) samples only. These results are interesting since sewage samples represent the influent wastewater flow coming into the WRP, and would not be affected by changes in the disinfection process. Therefore, the significant reduction in microbial diversity in 2017 suggests a compositional variation in the incoming sewage samples between 2016 and 2017. However, the precise cause of this variation upstream is difficult to attribute to a specific causal factor (for instance water chemistry, rainfall). It is also uncertain whether these trends in diversity described for river water, sediment, effluent, and sewage samples will continue or whether they are simply examples of natural inter-annual variation.

To better understand the implications of the microbial diversity trends described above, we used the compositional analyses of the 16S (small subunit of a microbe genetic material used as the standard for classification and identification) data to characterize the microbial taxa changes in pre disinfection, post-disinfection, and the phased TARP completion. This was mainly for understanding whether changes in microbial taxa made sense in relation to the disinfection process and TARP improvement efforts. At O'Brien WRP, the downstream North Shore channel river water samples demonstrated a significant decrease in the abundance of genera such as Acinetobacter and Claocibacterium, which are known sewage indicators, and an increase in the abundance of the genus, Hydrogenophaga which is known for its association with waste-water treatment plants and its role in biodegradation (Fig. 2). The sediment samples from the waterway by O'Brien WRP area also showed differential microbial signatures during pre- and post-disinfection period. Genera like *Anaerolinea, Bifidobacterium, Devosia,* and *Paracoccus* significantly reduced post-disinfection whereas we observed a significant increase in the abundance of genera such as *Dechloromonas* and *Rhodoplanes*. The genera *Devosia* and *Paracoccus* are also known to be enriched in sludge across wastewater treatment plants (Fan et al. 2017). *Dechloromonas* increased post-disinfection, and members of this genus are known to be a part of the bacterial community in wastewater treatment plants and significantly correlate with improved performance of wastewater treatment (Yang et al. 2011). Hence, *Dechloromonas* might have been introduced downstream due to disinfection process taking place at the WRPs. These results emphasize the impact of the disinfection in greatly reducing sewage and human fecal indicators in the CAWS.

The TCR storage in 2016 and 2017 provided crucial protection by capturing all the first flush of Calumet WRP sewage and stormwater from combined sewers that previously (prior to 2016) flowed into Cal-Sag Channel. Together with the disinfection, at Calumet WRP, the

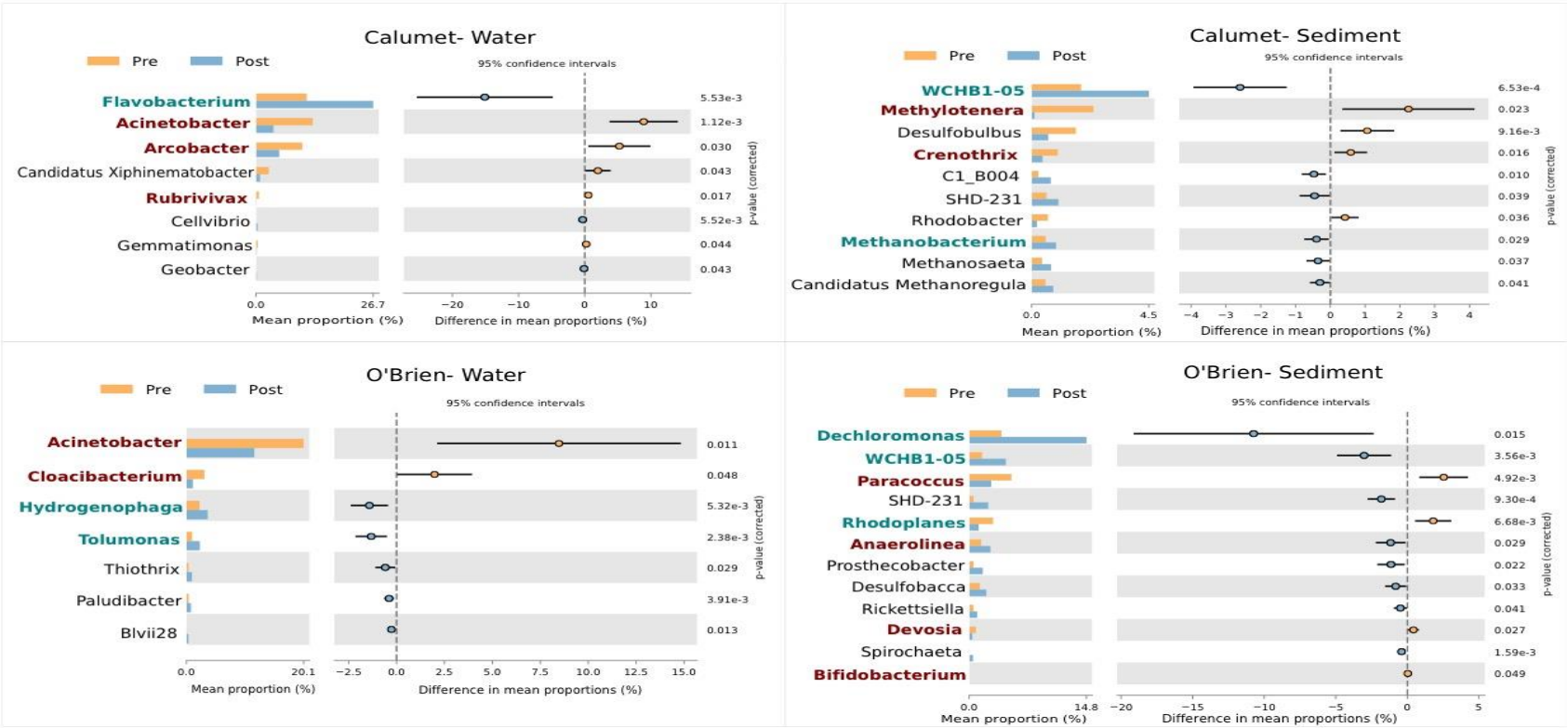


FIGURE 2 The Sewage and Wastewater Microbes Decrease Significantly Post-Disinfection Compared to Prior (pre) Condition without Disinfection. The freshwater microbes increase significantly post-disinfection. Few of the indicators for sewage (e.g. *Acinetobacter, Cloacibacterium*) and human fecal material (e.g. *Arcobacter, Bifidobacterium*) have been highlighted using red color. Green colored labels represent the bacterial genera associated with fresh water (e.g. *Flavobacterium*) and also the ones which are known to be introduced by wastewater treatment plants (e.g. *Hydrogenophaga, WCHB1, Dechloromonas*). This figure shows list of statistically differential bacterial genera with Benjamini-Hochberg FDR corrected p-values (< 0.05) labelled for each taxon. For all four different sample types, in each figure there are two sub-panels : i) Mean proportion (%) which stands for average relative abundance/proportion of the taxa in the data, ii) Difference in mean proportions (%) stands for the percentage increase or decrease of the specific taxa in one of the groups over the other group compared (which in this case are: Pre and Post disinfection).

year 2017 was characterized by a reduction of the genus *Bacteroides* and the phylum *TM7*, order *Streptophyta*, and MLE1-12, when compared to 2016 ($p_{FDR} < 0.05$). *Bacteroides* is also a known marker of sewage pollution and has been used as a tracer of ecosystem. Similarly, the sediment samples downstream of the Calumet WRP showed significant differences between the two disinfection years i.e. 2016 and 2017. We identified a set of 16S exact sequence variants (ESV- is now used replacing operational taxonomic units in marker-gene data analysis) belonging to genera *Sediminibacterium*, *Epulopiscium*, family *Lachnospiraceae*, and order *Clostridales* in the year 2017 when compared to 2016 (Fig. 2). *Lachnospiraceae* and *Clostrdiales* are used as human fecal indicators for the influent sewage samples (McLellan et al. 2013). *Dehalococcoides* also showed a significant increase in the year 2017. *Dehalococcoides* are strictly anaerobic bacteria which are capable of metabolizing water pollutants (such as chlorinated ethenes, polychlorinated biphenyls) produced during water disinfection processes such as reductive dechlorination (Islam, Edwards, and Mahadevan 2010). This is interesting because Calumet WRP's disinfection process is based on chlorination and dechlorination.

Significant reduction in sewage indicators and increase in fresh water indicators can be attributed to both disinfection as well as phased TARP completion.

Genomic Data and Culture Based Fecal Indicator Bacteria (FIB) Monitoring Method

The USEPA's bacteria standards for *E. coli* detection are 126 CFU / 100 mL as geometric mean, and 410 CFU / 100mL as a statistical threshold value. *E. coli* is the EPA-recommended indicator of fecal pollution, however fecal coliform (FC) continues to be widely used for bacteria quality monitoring. In Illinois, FC is tested per the National Pollutant Discharge Elimination System (NPDES) permit requirement for disinfected effluent to meet the monthly geometric mean of 200 CFU/ 100 ml and 10 % of sample to be less than 400 CFU/100 ml. The traditional plate count monitoring for FC bacteria levels in effluent at sites downstream of the O'Brien WRP (Fig. 3) and the Calumet WRP (Fig. 4) showed higher concentrations in the pre-disinfection period (2013 to 2015) compared to the post-disinfection period (2016-2017). There was significant FC bacteria reduction in the final effluent and immediately downstream of both WRPs, suggesting that disinfection was effective at reducing FC bacteria levels meeting the WRPs permit compliance required levels (Fig. 3 and 4). However, similar decreases in FC concentrations were not seen in river waters upstream of the WRPs and in tributaries. The FC concentration was found to be above the water quality standard in select tributaries and further downstream river locations.

We wanted to compare the FIB amplicon sequence data with the culture based FC and/or *E. coli* abundance data that is required for regulatory monitoring. We found that FIB was an uncommon microbial subgroup in pre-effluent and river samples. This suggested that amplicon analysis might need a greater depth of sequencing to detect and quantify *E. coli* and fecal coliforms at the concentrations and relative abundance found in environmental samples. Therefore, to test the amplicon sequencing method sensitivity and detection limit, we conducted a spike recovery experiment using a Bioball® spiked with a known concentration of *E. coli* to determine if we could detect *E. coli* abundances of ~100 cells per 100mL of water using

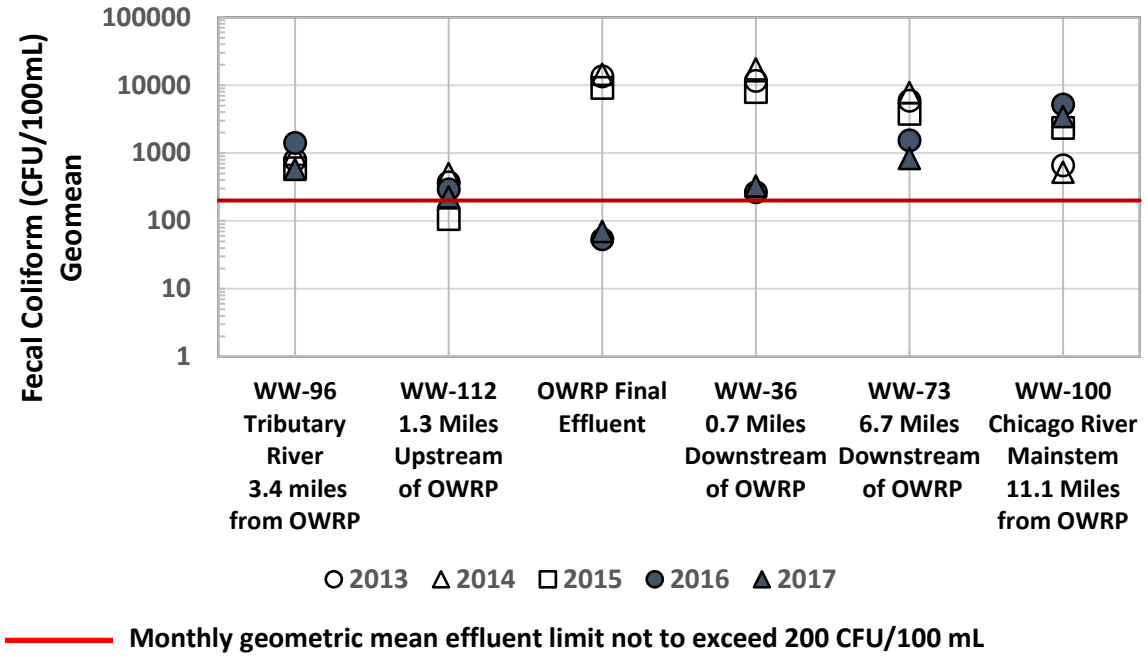


FIGURE 3 Geometric Mean Fecal Coliform Concentrations Observed March-November (2013-2017) in Final Effluent (UV treated), Upstream and Downstream of O'Brien WRP (OWRP), North Shore Channel and North Branch Chicago River Locations

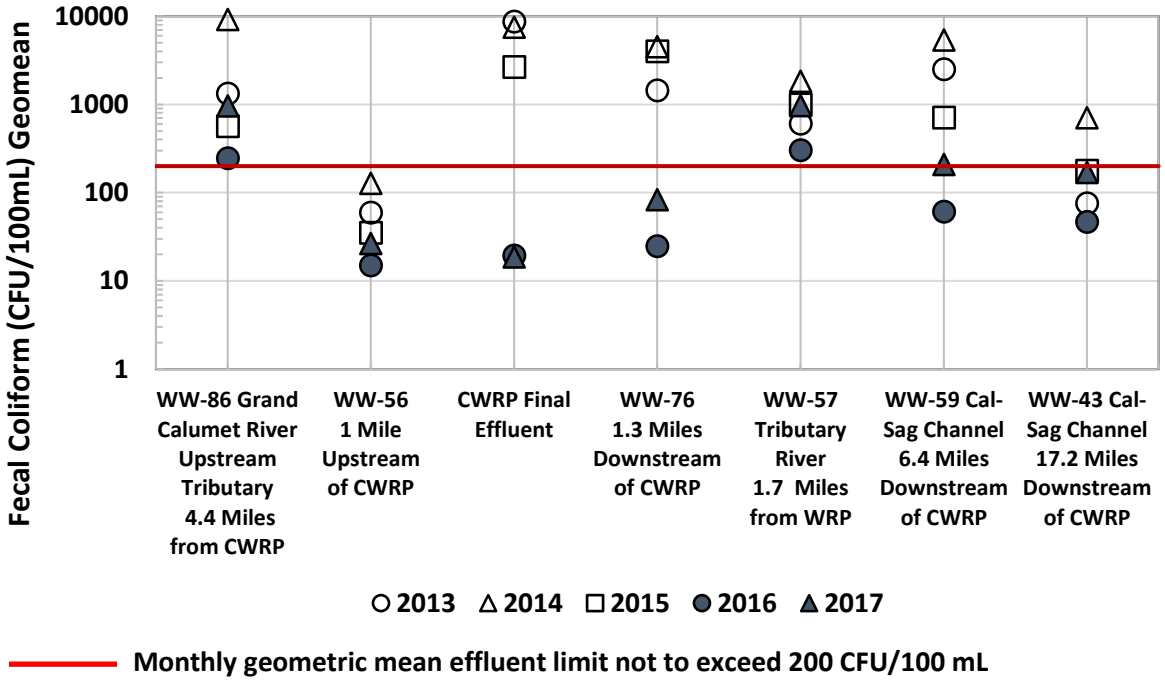


FIGURE 4 Geometric Mean Fecal Coliform Concentrations Observed March-November (2013-2017) in Final Effluent (chlorinated/dechlorinated treated), Upstream and Downstream of Calumet WRP (CWRP), Grand Calumet, Little Calumet, and Cal-Sag Channel River Locations

16S rRNA amplicon detection techniques. Spiked sample types included: (i) an O'Brien WRP secondary treated effluent sample, (ii) a CAWS water column sample, (iii) a 0.22 μm filtered CAWS water column sample, and (iv) a Phosphate buffered saline (PBS) control sample. The Bioball® experiment demonstrated that as few as 100 *E. coli* cells can be reliably detected across all four spiked, sample types at a sequencing depth as low as 1000 sequences per sample.

Thus, the apparent lack of *E. coli* sequences in CAWS 16S rRNA amplicon datasets is a result of a true low abundance of *E. coli* and not inadequate sequencing or detection potential.

Functional Attributes of CAWS Microbial Communities. Most FIBs are not pathogenic. Therefore, it is of interest to use shotgun sequencing to determine the relative contribution of disease causing bacteria to the CAWS microbial community. Selected samples were analyzed by shotgun metagenome sequencing, which provides us with information on the gene functions associated with these microorganisms (i.e., "What are they capable of doing?"). Using this information, we are able to determine key functional community characteristics of interest by comparing sequence structure against a database of known sequences.

We sequenced 24 samples which were collected from 10 upstream and 14 downstream sites at Calumet WRP in 2013-2015 using whole community metagenome sequencing by Illumina HiSeq platform, which were then annotated. Overall, these samples contained an average of 99% bacteria, 0.6% eukaryotes, 0.2% archaea, 0.1% viruses, and the remainder 0.006% of sequences were unclassified. Shotgun metagenomic results thus far revealed that abundant genera in downstream sites were taxa typically found in post-wastewater disinfection water column and sediment samples. Interestingly, multiple species from these genera have been used for the degradation and remediation of organic contaminants. Microbes that have ability to cause disease appeared to exist at low levels across the select samples upstream and downstream of WRPs. The results demonstrated a significant impact of WRPs on downstream CAWS in detoxification and improving the health of the CAWS ecosystem.

Sources of microbial organisms at different points in the CAWS. We used a Bayesian statistical tool, SourceTracker 2.0 to determine the potential sources of microbes associated with each sample by sampling location and sampling period (Mustakhimov et al. 2013). For this analysis, a curated database was built using CAWS samples (effluent, sewage, sediment, fish gut and mucus) and additional 100,000 samples from the Earth Microbiome Project (EMP) 2017 release version. The sources from the EMP database included- animal feces, fresh water, soil, and stream sediment. We then attempted to match the unknown genetic material in the CAWS samples to the known sources in the database (Fig. 5).

The results thus far indicated that the sources of microbial diversity across all river water samples can be largely attributed to effluent, sewage, CAWS sediment, freshwater, and fish associated samples. The three CAWS regions i.e. North, 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

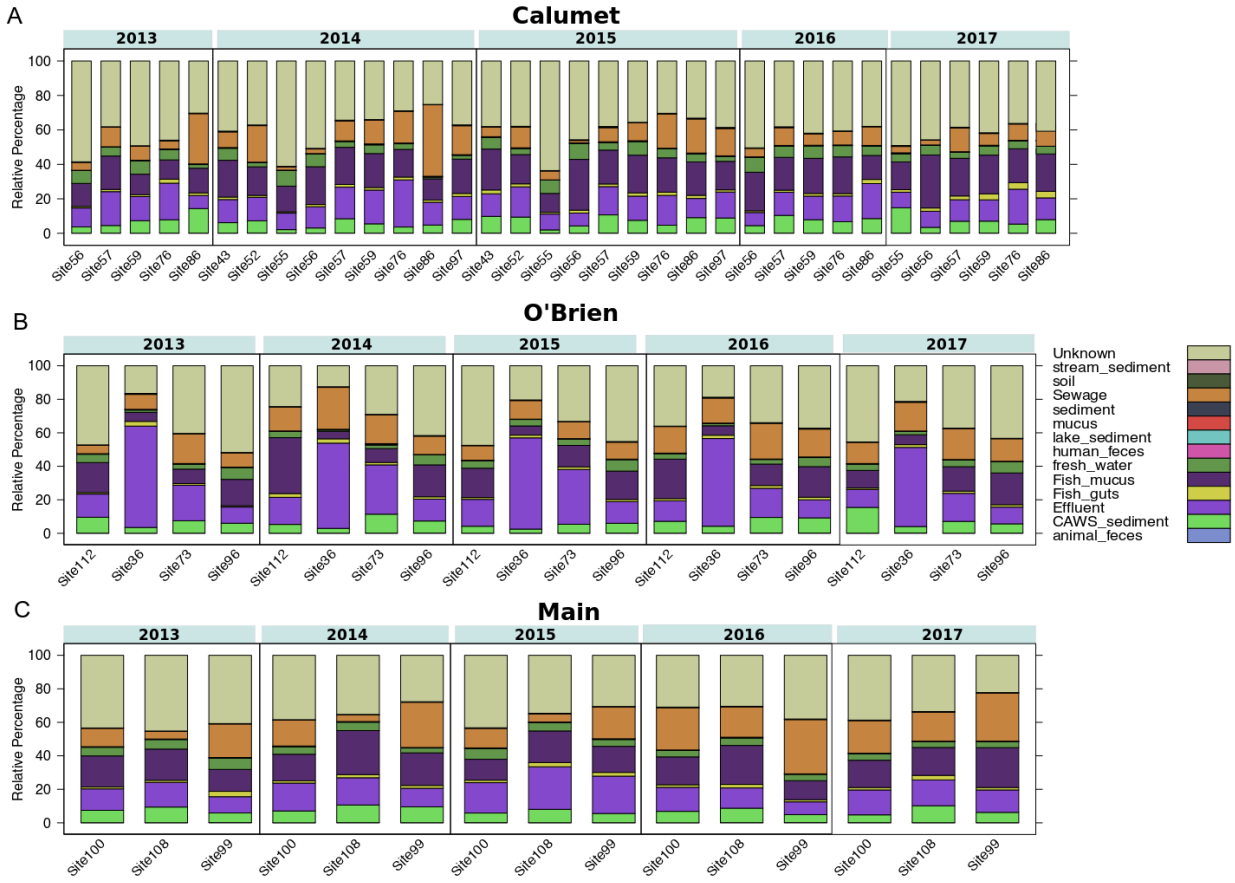


FIGURE 5 CAWS Microbial Community Sources Using Earth Microbiome Project Database. SourceTracker 2.0 Analysis of Water Column Samples by Sampling Site for Years 2013-2017 Using a Curated Database for (A) Calumet, (B) O'Brien, and (C) Main. A curated database was built using CAWS samples (i.e. effluent, sewage, sediment, fish gut and mucus) and additional 100,000 samples from the Earth Microbiome Project (EMP) 2017 release version. The sources from the EMP database included- animal feces, human feces, fresh water, soil, and stream sediment.

from the North region have a dominant effluent signature. Strikingly, the contribution made by human fecal matter across all water column samples was extremely low. However, there remains a large proportion of bacterial taxonomic diversity in the CAWS that cannot be reliably attributed to a 'source'. This is reflective of the sampling bias in the available gene libraries wherein urban river microbiomes are underrepresented. Additionally, it is likely that these are endemic but extremely rare taxa that are only found in the Chicago River.

CAWS-FIB MODEL

Introduction and Methods

The second goal of the study is to develop the CAWS-FIB model, a data-driven model for predicting fecal indicator bacteria (FIB) in the Chicago Area Waterway System. The

CAWS-FIB model predicts the FIB concentrations at any point along the CAWS using Gradient Boosting Machines (GBM), a machine learning approach. The research focus to date has been to use a GBM algorithm to predict fecal coliform concentration or density in the water column given a set of predetermined relevant environmental variables. We developed initial modeling results using data from the three Calumet WRP river sites (e.g., 56 (Upstream), 57 (Tributary), and 76 (Downstream)) with the largest number of observations available among the 12 sampling sites during the pre-disinfection period (2013-2015). Tributaries stations 52, 96 and 97 are not on the CAWS. Data for the CAWS-FIB model included three major categories such as 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, sediment, and heavy metals, etc.) data.

In addition to predicting FIB density given a set of relevant features or environmental variables, the model is capable of estimating the probability that a predicted FIB density will exceed a threshold number (probability of exceedance (POE), similar to the VirtualBeach model (Cyterski et al., 2013). The threshold number is a function of the regulatory limit (RL) and a decision value (DV). A DV is 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. In this preliminary model tests, the CAWS limit for fecal coliform of 200 CFU/100 mL was used as the RL and DV. 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. Thus, the regulatory limit (RL) is on the scale of actual observations, while the DV is on the scale of the model predictions. When we raise or lower the DV, we are inherently adjusting for the differences in scale between model predictions and the actual observations.

Results. The model results indicate that the most important explanatory variables for fecal coliform include pH and dissolved oxygen, as well as nitrate (NO_3) and water temperature. Net radiation, air temperature, and dissolved oxygen are also important explanatory variables followed by, discharge, water temperature, specific conductance, rainfall, and stage.

Model training and testing indicated that the CAWS-FIB model performed well when classifying whether a predicted FIB value is a false positives, true positives, false negatives, and true negatives, but produced lower accuracy values during testing. However, the 10-variable model consistently performed well with classification accuracy over 0.7 across sites. The limited amount of data for model training and testing as well as the large dynamic range of the data seem to limit the assignment of a POE to a predicted fecal coliform value to 0 and 100% only. For example, a predicted fecal coliform value less than RL and DV is assigned a 0% POE regardless whether it is very close to 200 CFU/100 mL (e.g. 198 CFU/100 mL) or well below 200 CFU/100 mL (e.g., 16 CFU/100 mL).

CAWS-FIB predicted fecal coliform concentrations ranged from 16 to 2,523 CFU/100 mL for site 56, 182 to 1,626 CFU/100 mL for site 57, and 452 to 89,657 CFU/100 mL for site 76. These values are within the range of the observed fecal coliform densities during 2013-2015.

Overall, the model training and testing results should be considered preliminary, and may improve after the nine additional sites are incorporated into the model training and testing. Additional approaches such as artificial neural networks (ANNs) and XGBoosting (XGB) will also be explored and the best performance will be used as the main algorithm for the CAWS-FIB model. Following the development of the predictive model for FIB, we will attempt to apply the model for predicting other fecal indicator bacteria (bacteriodes). This will only be possible for the post-disinfection years (2016-2019) due to the limited amount of pre-disinfection data available for the aforementioned FIBs in training the model.

This page intentionally left blank.

EXECUTIVE SUMMARY

This report summarizes the research conducted by Argonne National Laboratory (Argonne) for the Metropolitan Water Reclamation District of Greater Chicago (MWRD) for the first five years (2013-2017) of a seven-year study with emphasis on the impact of disinfection (2016-2017) on the microbial community of the Chicago Area Waterway System (CAWS). The study examines microbial communities following disinfection of secondary treated effluents at the T.J. O'Brien (UV) and Calumet (chlorine and dechlorination) Water Reclamation Plants (WRPs) and the phased implementation of the Tunnel and Reservoir Plan (TARP). The Thornton Composite Reservoir (TCR) was completed in 2015. It 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 overflow into CAWS in rainy weather. In addition to the effects of TARP and disinfection, we present an analysis of potential sources of the microbes found in the CAWS and the results of 16S rRNA amplicon gene sequencing which was used to characterize microbial community variability as a function of location, season, and environmental conditions.

A total of 2,077 samples were collected from 2013 to 2017 from WRPs and CAWS and analyzed using high-throughput sequencing. The whole microbial community composition using all the genetic material in the environmental sample was characterized using 16S rRNA gene-based analysis, while functional subsystem attributes of these microbial communities were characterized using shotgun metagenomics. We used a Bayesian statistical tool, SourceTracker 2.0 to determine the likely sources of microbes found in CAWS by sampling location and sampling period.

The report also describes the development and preliminary testing of the CAWS-Fecal Indicator Bacteria (CAWS-FIB) model. The CAWS-FIB model uses Gradient Boosting Machines (GBM), a machine learning algorithm, to relate FIB concentration to weather, water chemistry, and hydrology related variables. The model can be used to predict FIB concentrations at any point along the CAWS for a given set of predetermined relevant environmental variables.

Based on the metagenomic analyses of CAWS samples from the year 2013-2015 (included in phase I report), microbial communities showed distinct differences across different sampling locations (biogeography), in which the main differentiator was the sample type (river water, sediment, effluent, etc.). These communities appear to be stable (in their diversity and composition) across these sampling years and sampling seasons. Our analysis also showed that microorganisms associated with final WRP effluent from disinfection can be tracked downstream and typically showed increased abundance in proximity to the secondary treated final effluent location. These included human fecal indicators including *Bifidobacterium* and sewage contamination indicators including *Acinetobacter*.

In 2016 MWRD implemented disinfection of secondary treated effluent at both O'Brien and Calumet WRPs. In Phase II, we therefore, aim to understand the impact of disinfection on microbial communities of both river water and sediment samples, using 16S rRNA gene sequencing data from 2013-2017 with the years 2013-2015 serving as baseline

(i.e. pre-disinfection). Approximately, 40 million reads for the 16S rRNA gene were generated for a total 2,077 samples collected from 2013-2017. To date, this is the first study to investigate the longitudinal and spatial impact of disinfection on the microbial ecology of an urban river. For this, we extensively sampled CAWS river water and sediment as well as treated effluent discharged from two WRPs over the course of three years (2013-2015) prior to and two years (2016-2017) following the implementation of new disinfection processes. Based on sample type, CAWS sediment samples were the most diverse followed by sewage, effluent, and river water samples ($p < 0.05$). Similarly, the beta diversity (number of taxa [microbial subgroups] across sampling media [i.e. sediment, river water, effluent]), analyses based on weighted and unweighted UniFrac distance matrices, showed significant clustering of river water, sediment, and effluent samples when ordinated using PCoA plots ($p = 0.001$).

Whilst maintaining a relatively stable microbial composition during the sampling period from 2013 to 2015, post-disinfection years (2016-17) were characterized by significant differences in microbial composition for river water, sediment, effluent, and sewage samples when compared to the pre-disinfection time period. 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 followed by a significant increase ($p < 0.05$) in 2017 when compared to 2016. However, for the effluent and sewage, we observed a different pattern, whereby the alpha diversity increased in 2016 followed by a significant decrease in 2017 ($p < 0.05$). It is uncertain whether these trends in diversity for river water, sediment, effluent, and sewage samples will continue or whether they are simply examples of natural inter-annual variation.

Current USEPA limits for *E. coli* detection are 126 CFU / 100 mL as geometric mean, and 410 CFU / 100mL as a statistical threshold value. *E. coli* is the EPA-recommended indicator of fecal pollution, however fecal coliforms (FC) continues to be widely used for water quality monitoring. In Illinois, FC is the National Pollutant Discharge Elimination System (NPDES) permit program required indicator bacteria for disinfected effluent monitoring, with a monthly geomean threshold not to exceed 200 CFU/ 100 ml and 10 % of sample to be less than 400 CFU/100 ml. The traditional plate count monitoring for FC bacteria levels downstream of the O'Brien WRP and the Calumet WRP showed higher concentrations in the pre-disinfection period (2013 to 2015) compared to the post-disinfection period (2016-2017). There was significant reduction in the final effluent, suggesting that disinfection was effective at reducing FC bacteria levels meeting the two WRPs NPDES permit compliance target (200 CFU/100 ml monthly geometric mean). However, similar decreases in FC concentrations were not seen in river waters upstream of the WRPs and tributaries.

Using the compositional analyses of the 16S data, we identified a significant post-disinfection reduction of known sewage and human fecal indicators such as *Acinetobacter*, *Cloacibacterium*, *Bifidobacterium*, and Clostridiales in both river water and sediment samples. There were significant differences between the two years i.e. 2016 and 2017 post disinfection. We observed a further reduction in sewage indicators such as bacterial taxa Lachnospiracae, Paraprevotellacae, Bacteroides, and Clostridiales in 2017 when compared to 2016. The Bioball® experiment demonstrated that as few as 100 *E. coli* cells can be reliably detected across all four spiked, sample types at a sequencing depth as low as 1000 sequences per sample. These sample

types included: (i) an O'Brien WRP secondary treated effluent sample, (ii) a CAWS water column sample, (iii) a 0.22 μm filtered CAWS water column sample, and (iv) a Phosphate buffered saline (PBS) control sample.

We used a Bayesian statistical tool, SourceTracker 2.0 to determine the likely sources of microbes found in CAWS water column samples. For this analysis, a curated database was built using CAWS samples (effluent, sewage, sediment, fish gut and mucus) and additional 100,000 samples from the Earth Microbiome Project (EMP) 2017 release version. This database was used to determine potential sources of microbes that occur in CAWS water column samples at each sampling site by sampling year. SourceTracker 2.0 indicated that the sources of microbial diversity across all river water samples can be largely attributed to effluent, sewage, CAWS sediment, freshwater, and fish. The three CAWS regions i.e. North, Main, and Calumet regions, 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 WRP 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. Strikingly, the contribution made by human fecal matter across all water column samples was extremely low. As already reported for the sampling years from 2013 to 2015, there remains a large proportion of bacterial taxonomic diversity in the CAWS that cannot be reliably attributed to a 'source'.

Twenty-four CAWS water column samples from 2014 and 2015 were selected for deep metagenome sequencing to determine the functional attributes of the CAWS microbial community. The samples included 10 upstream and 14 downstream sites of Calumet WRP. In addition to genera like Flavobacterium (fresh-water bacterial biomarker) which were associated with downstream samples in our 16S data analyses, we identified many other genera which were significantly enriched in abundance in downstream sites, such as Burkholderia, Rhodococcus, Methylobacterium, Methylibium, and Alicycliphilus. Interestingly, multiple species from these genera have been shown to degrade organic contaminants. In contrast to microbial diversity, both upstream and downstream sites were functionally more conserved with no significant differences identified at the subsystem level. Overall, the most abundant functional categories included amino acid associated pathways (biosynthesis and metabolism), carbohydrate metabolism, protein metabolism, and RNA metabolism. Interestingly, pathways related to phages, and other transposable elements were the most variable between sampling sites, however, there was no specific enrichment pattern between the upstream and downstream samples. Pathways including sulfur metabolism, nitrogen metabolism phosphorous metabolism, iron transport, secondary metabolism were the most consistent across all samples and were relatively rare. However, these results are at sub-system level and we will next perform functional annotations at higher resolution i.e. at pathways and enzyme level. Additionally, these samples were collected from Calumet WRP for the years 2014-2015. We will next sequence samples from 2016-2017 in order to investigate the effect of disinfection in details. This will also be extended to O'Brien WRP.

The second goal of the study is to develop the CAWS-FIB model, a data-driven model for predicting fecal indicator bacteria (FIB) in the Chicago Area Waterway System (CAWS) The CAWS-FIB model predicts 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 The research focus to date has been to use a GBM algorithm to predict

fecal coliform indicator bacteria (FIB) concentration or density in the water column given a set of predetermined relevant environmental variables. We developed initial modeling results using the three Calumet WRP sites (e.g., 56, 57, and 76) with the largest number of observations available among the 12 sampling sites during the pre-disinfection period (2013-2015).

Data for the CAWS-FIB model 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, sediment, and heavy metals, etc.) data. The model can take environmental variables that come from frequent (hourly to daily) and one-time manual measurements.

GBM uses decision or regression trees rather than linear equations. Each decision or regression tree is composed of virtual branches and nodes and controlled by a set of decision rules. The preprocessed data were subdivided into training and testing with 85%-15% split. For each sampling site, the training set was used for determining or learning the model parameters and assessing the initial model performance, while the testing set 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. A number of measures were taken to avoid 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 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. In addition to predicting FIB density given a set of relevant features or environmental variables, the model is capable of estimating the probability of exceedance (POE) which is the probability (%) that a predicted FIB density will exceed a threshold number. The threshold number is a function of the regulatory limit (RL) and a decision value (DV). In our current model the RL is fixed, the DV can be variable. Setting the DV to some value not equal to RL may be confusing. However, when we adopt a modeling approach, we have decided to base our advisory decisions (to close a waterbody/beach from human contact or not) on a statistical model derived from historical data. For instance, a model prediction of 150 CFUs/100 mL may be approximately equivalent to an actual or "real" observation of 175 CFUs/100 mL, or a "real" value of 125 CFU/100 mL, depending on the specific model. Therefore, we should not think of model predictions as actual FIB concentrations, but only some quantity that is related to actual FIB concentrations.

The model results indicated that top 15 most important FIB explanatory variables are comprised of both the one-time, manually collected during the sampling period and more frequently measured or time-lagged environmental factors. The manually measured relevant explanatory variables include pH and dissolved oxygen (DOMan), which rank as top variables for two of the sites, as well as nitrate (NO_3) and water temperature (TwMan). Net radiation (RnSDhrs), air temperature (TaSDhrs or TaMeanhrs), and dissolved oxygen (e.g., DOSDhrs or DOMeanhrs) are the most common time-lagged explanatory variables followed by, discharge (LQ10Maxhrs or LQ10Minhrs), water temperature (TwSDhrs), specific conductance (SpCondSDhrs or SpCondMeanhrs), rainfall (RsqrSumhrs), and stage (HDiffhrs).

Model training and testing indicated that the CAWS-FIB model performed well when classifying whether a predicted FIB value is a false positive (FP), true positive (TP), false negative (FN), and true negative (TN), but produced lower accuracy values during testing. However, the 10-variable model consistently performed well with classification accuracy over 0.7 across sites. Predicted FIB concentrations using the hypothetically generated dataset were within the range of the observed FIB densities during 2013-2015. The limited amount of data for model training and testing as well as the large dynamic range of the data seem to limit the assignment of a POE to a predicted FIB value to 0 and 100% only. For example, a predicted FIB value less than RL and DV is assigned a 0% POE regardless whether it is very close to 200 CFU/100 mL (e.g. 198 CFU/100 mL) or well below 200 CFU/100 mL (e.g., 16 CFU/100 mL).

Overall, the model training and testing results should be considered preliminary, and may improve after the nine additional sites are incorporated into the model training and testing. Additional approaches such as artificial neural networks (ANNs) and XGBoosting (XGB) will also be explored and the best performance will be used as the main algorithm for the CAWS-FIB model. Following the development of the predictive model for FIB, we will attempt to apply the model for predicting other FIB (Bacteriodes, etc.). This will only be possible for the post-disinfection years (2016-2019) due to the limited amount of pre-disinfection data available for the aforementioned FIBs in training the model.

This page intentionally left blank.

1 INTRODUCTION

Traditional laboratory-culture methods such as fecal bacteria counts and select pathogen Polymerase Chain Reaction (PCR)-based methods have 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. In addition, these methods do not completely describe the diversity of microbial communities present in the CAWS. Together with 16S rRNA gene sequencing, metagenome sequencing can supersede, for qualitative analyses, typical culture-based methods that currently only detect approximately 8% of known microbes. 16S rRNA gene sequencing has been used for the accurate and reliable qualitative identification and classification of microorganisms. Metagenomics-based sequencing can capture all genes present in a microbial community giving insight into the functional potential of microbes present in that sample. Overall, these molecular methods reveal substantially more information about the diversity of the microbes present in the CAWS, their potential function, and their activity (e.g., nutrient cycling, ability to cause disease etc.), and they can be used to predict with greater accuracy the common sources of microbes in these waters. Thus, these methods can help us discover which microbes are present in the CAWS, and what these microbes are capable of doing in the CAWS.

This report summarizes the research conducted by Argonne National Laboratory (Argonne) for the Metropolitan Water Reclamation District of Greater Chicago (MWRD) for the first five years (2013-2017) of a seven-year study with emphasis on the impact of disinfection (2016-2017) on the microbial community of the Chicago Area Waterway System (CAWS). The study examines microbial communities following disinfection of secondary treated effluents at the O'Brien (UV) and Calumet (chlorine and dechlorination) Water Reclamation Plants (WRPs) and the phased completion of the Tunnel and Reservoir Plan (TARP). In addition to the effects of disinfection, we present an analysis of potential sources of the microbes found in the CAWS and the results of 16S rRNA amplicon gene sequencing which was used to characterize microbial community variability as a function of location, season, and environmental conditions.

The report also describes the development and preliminary testing of the CAWS-Fecal Indicator Bacteria (CAWS-FIB) model. The CAWS-FIB model use a machine learning algorithm to predict FIB concentrations at any point along the CAWS for a given set of predetermined relevant environmental variables related to weather, water chemistry, and hydrology. The model also specifies the most important explanatory variables in predicting fecal coliform concentration and provides the probability that FIB concentrations will exceed U.S. Environmental Protection Agency (USEPA) standards.

This page intentionally left blank.

2 16S RRNA-BASED ANALYSIS OF RIVER WATER AND SEDIMENT SAMPLES COLLECTED POST DISINFECTION AT CALUMET AND O'BRIEN WATER RECLAMATION PLANTS

2.1 INTRODUCTION

Urban river ecosystems are greatly perturbed by untreated wastewater. Wastewater contains organic pollutants that stimulate eutrophication of freshwater ecosystems thus, reducing overall water quality (Carey and Migliaccio, 2009). Water reclamation plants are essential for the treatment of municipal waste as they remove a large fraction of C-, N-, and/or P-based nutrients, before discharging the treated effluent into receiving water bodies (Gücker, Brauns, and Pusch 2006). Water reclamation plants (WRPs) provide a valuable service; however, it remains crucial that additional measures are taken to examine and monitor the microbial ecology of ecosystems receiving treated effluent. Particularly, addressing questions such as: How will disinfection impact the structure and function of receiving water and sediment environments? At what taxonomic level (i.e. at family or genus or strain level) do we observe changes driven by disinfection? How does a varying community structure ultimately affect the functional capabilities of the microbial community?

There are many studies that examine the structure and composition of microbial communities associated with urban water systems, however, this is the first study to investigate the impact of disinfection on the microbial ecology of river water and sediments (Van Rossum et al. 2015; McLellan, Fisher, and Newton 2015; Payne et al. 2017). Previous studies on urban river systems have observed decreased microbial diversity and disrupted nitrogen metabolism in water bodies receiving secondary treated effluent (Drury, Rosi-Marshall, and Kelly 2013; Lu and Lu 2014; Wakelin, Colloff, and Kookana 2008). This observation may vary at different sites, as the impact of a WRP on the downstream water and sediment microbiota depends on different factors such as type of WRP, the chemical composition of the effluent, the population size of the urban area, climate and geography of the region, the size and flow rate of the river, and the buffering and self-purification capacity of the stream sediments to stresses.

This study, which started in 2013, aims to better understand the composition and sources of the microbial community associated with the CAWS using state-of-the-art 16S rRNA gene amplicon- and metagenome-based sequencing. This study particularly aims to provide an understanding of the sources of the CAWS' microbial communities--are they from specific sources and are they widespread or constrained to particular sections of the CAWS? Potential sources include effluent from water reclamation plants (WRPs), direct stormwater runoff, and combined sewer overflows (CSOs). Using 16S rRNA gene sequence data for the years 2013-2015, we observed no significant differences in overall microbial diversity, suggesting a stable riverine ecosystem. Results also showed that microbial diversity did not change significantly between samples collected during dry weather (dry events) and samples collected after precipitation (wet). However, the sediment samples when compared to river water samples showed significantly higher bacterial diversity (alpha diversity). Further, we didn't observe significant differences in microbial richness (alpha diversity) between different seasons or months. Microbial community profile similarities between samples (beta diversity) showed that there were significant differences in microbial community composition across sampling media,

including beach water, fish gut, fish mucous, mixed liquor, secondary treated final effluent, sediment, sewage, and river water, but no significant differences in beta-diversities were observed by sampling month or year. Both alpha and beta diversity analyses demonstrated significant variation between sampling medium suggesting higher impact of sampling medium on microbial community composition when compared to sampling month, year or seasons. The samples were further analyzed for human fecal and sewage contamination indicators. We identified presence of human fecal indicators such as *Bifidobacterium, Bacteroides* and sewage indicators such as *Acinetobacter and Arcobacter* in both river water and sediment samples. While, presence of human fecal indicators such as *Bifidobacterium* and *Bacteroides* was observed, the overall relative proportions were very low i.e. 0.2% and 2%, respectively, in water. Among, the sewage indicators, *Arcobacter* is known to be abundant in sewage water with relative proportions ranging between 5 to 11% as per previous metagenomic studies (Fisher et al. 2014). In our data, we identified *Arcobacter* to be ~10% abundant in our sequence data which is in agreement with previous reports. Similarly, *Acinetobacter* was about 12% abundant in the water samples, which is an ubiquitous genus and has been found to be associated with sewage, sludge in the past (Al Atrouni et al. 2016; Doughari et al. 2011). The previous studies on the genus *Acinetobacter* are precisely species-specific and hence we will further investigate the relevance of different *Acinetobacter* species using the whole genome metagenome sequencing effort.

In 2016 MWRD started disinfecting the secondary treated effluents at the O'Brien and Calumet WRPs. The WRPs at Calumet and O'Brien use different disinfections processes i.e. chlorination/dechlorination and UV treatment, respectively. Therefore, for Phase II, we focused on documenting potential changes in the microbial communities post-disinfection (i.e. 2016-2017) and by disinfection type. The Phase I data generated from years 2013-2015 will serve as baseline. This report describes changes in microbial signatures over time in WRP influent (raw sewage), WRP final disinfected effluent, and in sediment and river water samples collected from the North Shore Channel, North Branch Chicago River, Tributaries, Main Stem Chicago River, South Branch Chicago River, Grand Calumet River, Little Calumet River, and Cal-Sag Channel.

2.2 MATERIALS AND METHODS

2.2.1 Assessing Microbial Community Structure in CAWS Samples Over Five Years Using 16S rRNA Amplicon Gene Sequencing

We utilized 16S rRNA gene amplicon sequencing to characterize the microbial communities in CAWS samples during years, 2013-2017. We collected 261 effluent samples, 472 sediment samples, 537 water column samples, 205 wet weather river samples, 61 influent sewage samples, 339 (bottle, filter, equipment) blanks, 129 fish samples, and 73 other sources (Table 1). Sewage and effluent samples were collected from at O'Brien and Calumet WRPs (Fig.1, Table1). Disinfection processes were implemented in 2016 at the O'Brien (UV) and Calumet (chlorination/dechlorination) WRPs. In 2017, we observed the phased implementation of the Tunnel and Reservoir Plan (TARP) at the Calumet WRP. Thereafter, we continued sampling treated effluent, water, and sediment samples both upstream and downstream of the two WRP sites (Table 2).

TABLE 1 Total Number of Samples Collected per Sample Type from 2013 to 2017

Year	Final Effluent	Sediment	River water	Wet/Dry	Raw sewage	Bottle blank	Filter blank	Equip blank	Trip blank	Fish samples	Spiked	Plate	Lake Bypass	Beach	Total
2013	55	78	82	0	0	15									230
2014	72	84	133	54	9	1	22	16	17	0					408
2015	76	99	109	54	17	1	33	36	33	48	9			7	522
2016	41	104	106	44	19	8	32	24	18	47					443
2017	17	107	107	53	16	7	31	23	22	34		55	2		474
Total	261	472	537	205	61	339				129	9	55	2	7	2077

TABLE 2 Details of Location for Each Site for the Two Water Reclamation Plants

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 ~ 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
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 ~ 4.4 Miles from Calumet WRP
55[b,c]	Calumet River	130th St.	Upstream Tributary ~ 5.7 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 River ~ 1.7 Miles from Calumet WRP
52[b,c]	Little Calumet River	Wentworth Ave.	Tributary River ~ 1.7 Miles from Calumet WRP
97[b,c]	Thorn Creek	170th St.	Tributary River ~ 1.7 Miles 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.

Additional information on the number of river water and sediment samples that were processed by sampling site is included in Table 3. DNA was extracted from these samples using an extraction protocol described in Appendix A (Protocol#2 and #3) of the work plan.

2.2.2 Amplicon Based Microbial Community Sequencing Analysis

The 16S rRNA gene was amplified using a protocol described in Appendix A (Protocol#4) of the work plan. 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

TABLE 3 Summary of Sediment and Water Column Samples by Sites on the CAWS from 2013-2017

		Water column					Sediment				
Site	Address	2013	2014	2015	2016	2017	2013	2014	2015	2016	2017
36	North Shore Channel @ Touhy Ave.	7	8	9	15	10	6	10	7	9	8
43	Cal-Sag Channel @ Route # 83	0	13	7	0	5	0	0	0	0	0
52	Little Calumet River @ Wentworth Ave.	0	13	7	0	5	0	0	0	0	0
55	Calumet River @ 130th St.	0	12	7	0	5	0	0	0	0	0
56	Little Calumet River @ Indiana Ave.	6	15	16	9	14	8	8	8	8	9
57	Little Calumet River @ Ashland Ave.	6	14	16	9	14	8	7	8	9	9
59	Cal-Sag Channel @ Cicero Ave.	7	14	16	9	14	7	8	8	8	9
73	North Branch Chicago River @ Diversey Ave.	7	8	9	15	10	6	8	8	9	9
76	Little Calumet River @ Halsted St.	7	14	16		14	7	8	8	8	9
86	Grand Calumet River @ Burnham Ave.	6	13	16	9	13	7	8	8	9	9
96	North Branch Chicago River @ Albany Ave.	0	0	0	0	0	0	0	0	0	0
97	Thorn Creek @ 170th St.	6	9	8	15	10	6	9	8	9	9
99	South Fork, South Branch Chicago River @ Archer Ave.	0	13	7		5	0	0	0		0
100	Chicago River Main Stem @ Wells St.	7	9	9	13	10	6	9	9	9	9
108	South Branch Chicago River @ Loomis St.	7	9	9	14	10	7	8	9	9	9
112	North Shore Channel @ Dempster Street	7	8	9	13	10	7	9	9	9	9
	Total	80	180	169	233	160	80	101	98	105	107

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).

2.2.3 16S rRNA Gene Sequence Analyses

For 16S rRNA gene analysis, the 16 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 ESVs (Exact 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 ESVs with less than 10 reads were removed.

2.2.4 Statistical Analyses

Analysis of the resulting biome 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). Alpha diversity is defined as species richness (number of taxa) within a single sample. Beta-diversity was determined using weighted and unweighted UniFrac distance matrices (Lozupone et al. 2011). Beta diversity is defined as diversity in the microbial community between different environmental samples. Differences in microbial alpha diversity (based on Shannon and Inverse Simpson indices) and beta diversity were assessed for significance using permutational multivariate analysis of variance (PERMANOVA) (Anderson Marti J. 2014). Analysis of composition of microbiome (ANCOM) was used to identify differentially abundant bacterial ESVs in different sample types across different sampling periods (2013-2017) (p-value cut-off of 0.05 following Benjamini-Hochberg FDR correction) (Mandal et al. 2015).

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

Shotgun metagenomic data was generated using the DNA extracts following the Illumina Paired-End Prep kit protocol. The sequencing was performed using a 2 × 100 bp sequencing run on the Illumina GAIIx. Paired-end metagenomic reads for 24 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). Taxonomic and functional information was assigned to each read using MetaPhlAn (Overbeek et al. 2014) and MGRAST

(Meyer et al. 2008), respectively. The MGRAST annotations will be done using SEED database. The SEED database is a gene annotation database containing gene functions, either within a single organism or a set of gene/protein families across a set of organisms. The SEED is a constantly updated integration of genomic data which is then annotated using public genomes annotated by RAST, expert user annotations, metabolic modeling data, expression data, literature references verifying annotations and links to data from other popular resources including Swiss-Prot (UniProt Consortium 2014), GenBank (Benson et al. 2013), IMG (Markowitz et al. 2012), KEGG (Kanehisa et al. 2012), and CDD (Marchler-Bauer et al. 2013).

2.3 RESULTS

2.3.1 Alpha- and Beta- Diversity Comparison of the CAWS Samples from 2013-2017

We performed quality control by analyzing a total of 339 blank samples comprising equipment blanks, filter blanks, and trip blanks for the five sampling years. Blanks serve as indicators of microbial contamination associated with any equipment or reagent used for sampling and analysis. The blank samples (n=339) showed DNA concentrations below 1 ng/µL. Samples containing <1 ng/µL are typically considered 'sterile' as they contain DNA quantities that cannot be reliably amplified by a standard PCR reaction (Brandt and Albertsen 2018; Castelino et al. 2017). These samples failed during sequencing with very low sequence data generated (< 1000 reads) and were not included in the analyses. This also supported that there was no contamination due to the sample processing and sequencing process, which was also the ultimate goal of collecting and using black samples as controls. Out of 1738 experimental samples, 13 samples were also excluded from the analyses due to low reads count (<1000 reads). For the remaining 1725 samples, Illumina MiSeq generated approximately 40 million 16S rRNA nucleotide reads for all the samples collected from 2013-2017 with an average of 10,824 reads per sample representing 45,892 unique ESVs in the total dataset.

Microbial community species diversity (alpha diversity) was estimated using the Shannon and Inverse Simpson metrics. Shannon and Inverse Simpson indices are both positively correlated with species richness and evenness (the distributed abundance of those species), with Shannon weighted to rare species and Simpson weighted towards abundant species. Based on the Shannon index, sediment samples were the most diverse followed by mixed liquor and effluent samples ($p_{PERMANOVA} < 0.05$, Figure 1A). No significant differences were observed in microbial diversity for river water, sewage, and beach water samples ($p_{PERMANOVA} > 0.05$, Figure 1A). Fish associated gut and mucus samples were the least diverse of all sample types ($p_{PERMANOVA} < 0.05$, Figure 1A).

We observed significant differences in beta-diversity between sample types based on weighted and unweighted (data not shown) UniFrac distance measures ($p_{PERMANOVA} < 0.05$, Figure 1B). CAWS water column samples formed a distinct cluster with beach water samples that separated from the other sample types ($p_{PERMANOVA} < 0.05$). Sediment and effluent samples ordinated into two separate clusters in the principal coordination analyses (PCoA) plot (Figure 1B). While no significant differences were seen between fish gut and mucus samples

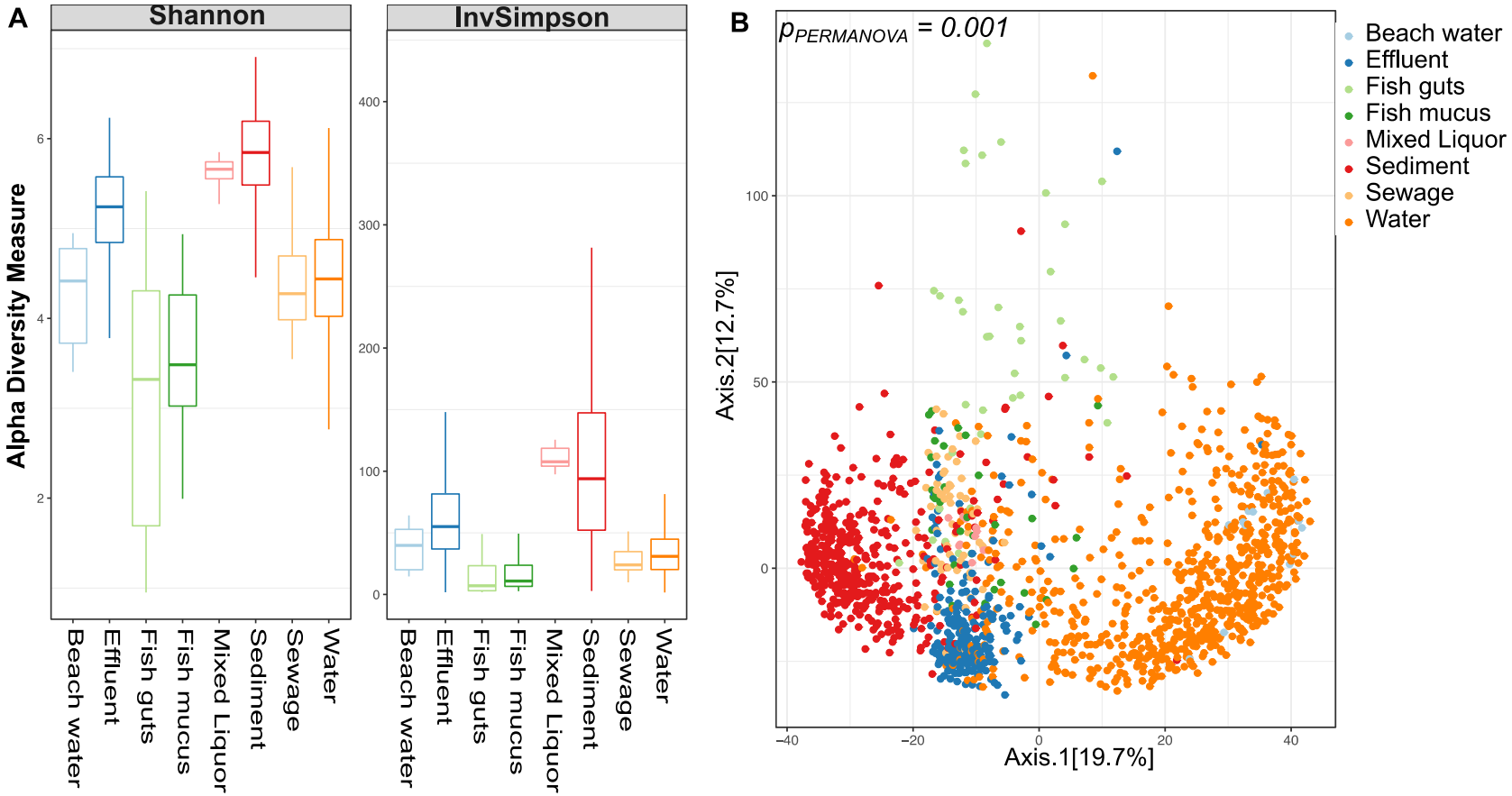


FIGURE 1 The Alpha and Beta Diversity Analyses of CAWS Samples Collected from 2013-2017. (A) The distribution of alpha diversity indices- Shannon and Inverse Simpson for each sample type consolidated for all five sampling years (2013-2017); the sediment samples are the most diverse followed by effluent and water samples, with fish associated samples least diverse. (B) Principal Coordinate Analyses (PCoA) plot based on the weighted UniFrac distance matrix showing clustering patterns of different sample types, i.e. beach water, effluent, fish guts and mucus, mixed liquor, 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 ($p < 0.05$).

based on the alpha diversity, beta diversity analyses demonstrated a clear separation between the sample types ($p_{PERMANOVA} < 0.05$, Figure 1B).

As reported in the Phase I Report, no significant differences in alpha diversity were observed between sampling periods i.e. 2013-2015 for any sample type http://pepportal.mwrd.local:50100/irj/portal/anonymous/Microbiome. However, the disinfection years 2016-2017 were characterized by significant changes in microbial diversity in sewage, effluent, river water, and sediment samples. Sediment samples collected from 12 sites across the CAWS demonstrated a significant reduction ($p_{PERMANOVA} < 0.05$) in microbial diversity in 2016 (post disinfection) as compared to 2015 (pre-disinfection) (Figure 2). Interestingly, a significant increase in microbial diversity of sediment samples was observed in 2017 as compared to 2016 ($p_{PERMANOVA} < 0.05$, Figure 2). This pattern was consistent across all the sediment samples from the 12 sampling sites (Figure 2). Likewise, river water samples collected from 17 sampling sites in 2016 were characterized by lower microbial diversity when compared to 2015, then followed by a significant increase in 2017 ($p_{PERMANOVA} < 0.05$, Figure 2).

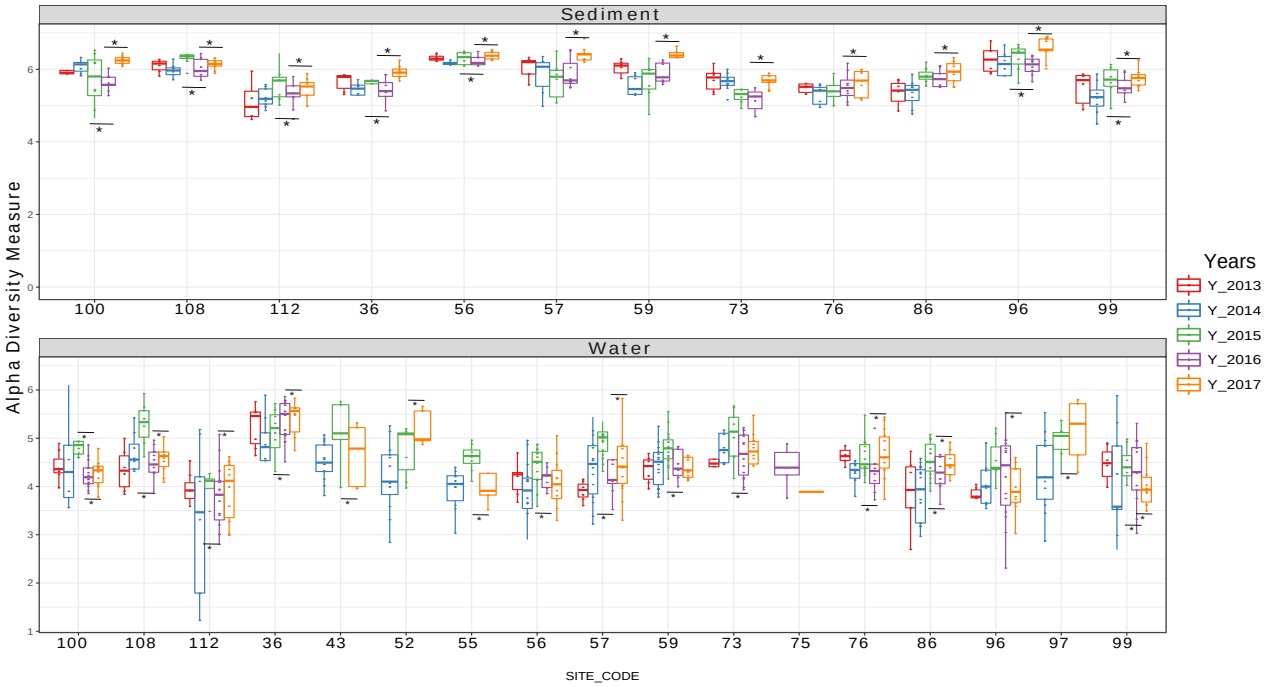


FIGURE 2 Alpha Diversity Analyses for Sediment and Water Samples Collected at Different Sites Over a Period of Five Years. The distribution of Shannon diversity index was shown for sediment and river water samples collected from 12 and 17 sites, respectively. The boxplots are grouped by sampling year. The asterisks sign represents statistically significant differences (PERMANOVA, $p < 0.05$) between years 2015-2016 and 2016-2017. Across the sediment samples, as of general trend, there is a significant decrease in diversity from 2015-2016 (i.e. first year post disinfection) and then a significant increase in diversity from 2016-2017 (i.e. second year post-disinfection). Similar pattern was also seen for river water samples.

Next, effluent and sewage samples were investigated for significant differences in microbial diversity at the Calumet and O'Brien WRPs. Based on Shannon and Inverse Simpson indices, Calumet WRP effluent samples demonstrated a significant increase in microbial diversity in 2016 as compared to 2015, followed by a significant reduction in 2017 ($p_{PERMANOVA} < 0.05$, Figure 3). In contrast, effluent samples from O'Brien WRP didn't demonstrate any significant change in microbial diversity post-disinfection. These variations between the two sites can be attributed to the local water chemistry, different sources of water from different tributaries, and the dynamic micro-environment. For sewage samples collected at Calumet WRP, we observed no significant differences in alpha diversity between 2015 and 2016. However, a significant reduction in alpha diversity was noted in 2017 when compared to 2016 ($p_{PERMANOVA} < 0.05$, Figure 3). While O'Brien WRP (UV disinfection) effluent samples did not show significant differences in microbial diversity post-disinfection, a significant reduction was observed in 2017 when compared to 2016 in agreement with the Calumet WRP (chlorine disinfection) sewage samples ($p_{PERMANOVA} < 0.05$, Figure 3). These results are interesting since sewage samples represent the influent coming into the plant, and would not be affected by changes in the disinfection process. Therefore, the significant reduction in microbial diversity in 2017 suggests a compositional variation in the incoming sewage samples between the sampling years 2016 and 2017. However, the precise cause of this variation upstream is difficult to attribute to a specific causal factor (for instance water chemistry).

We then investigated the post-disinfection alpha diversity patterns across all river water and sediment samples from sites immediately upstream and downstream of the two WRPs i.e. Calumet (site 56- upstream, site 76- downstream) and O'Brien (site 112-upstream and site 36-downstream). Overall, the microbial diversity of river water samples at the two WRP sites increased downstream as compared to samples collected upstream ($p_{PERMANOVA} < 0.05$, Figure 3). Sediment samples, however, demonstrated a different pattern. For Calumet WRP CAWS area sediment samples, we observed a significant reduction in microbial diversity in samples collected downstream as compared to those collected upstream. At the O'Brien WRP downstream North Shore channel sediment samples showed no significant difference by sampling location (upstream *vs.* downstream) (Figure 3). Calumet WRP area river water samples showed a significant reduction in microbial diversity ($p_{PERMANOVA} < 0.05$) post-disinfection (i.e. 2016). However, river water samples collected from the same site in 2017 showed a significant increase in microbial diversity ($p_{PERMANOVA} < 0.05$, Figure 3) compared to 2016. At O'Brien WRP, the disinfection did not significantly alter overall microbial diversity in river water samples collected downstream as compared to those collected upstream (site 36). At this site, sediment samples demonstrated a significant reduction in microbial diversity post-disinfection in 2016 followed by a significant increase in 2017 ($p_{PERMANOVA} < 0.05$, Figure 3). While, these differences are significant, we have further looked into compositionality of samples using differential abundance of taxa in order to capture unique microbial signatures that can be attributed to the disinfection process at two WRPs and/or the TCR completion.

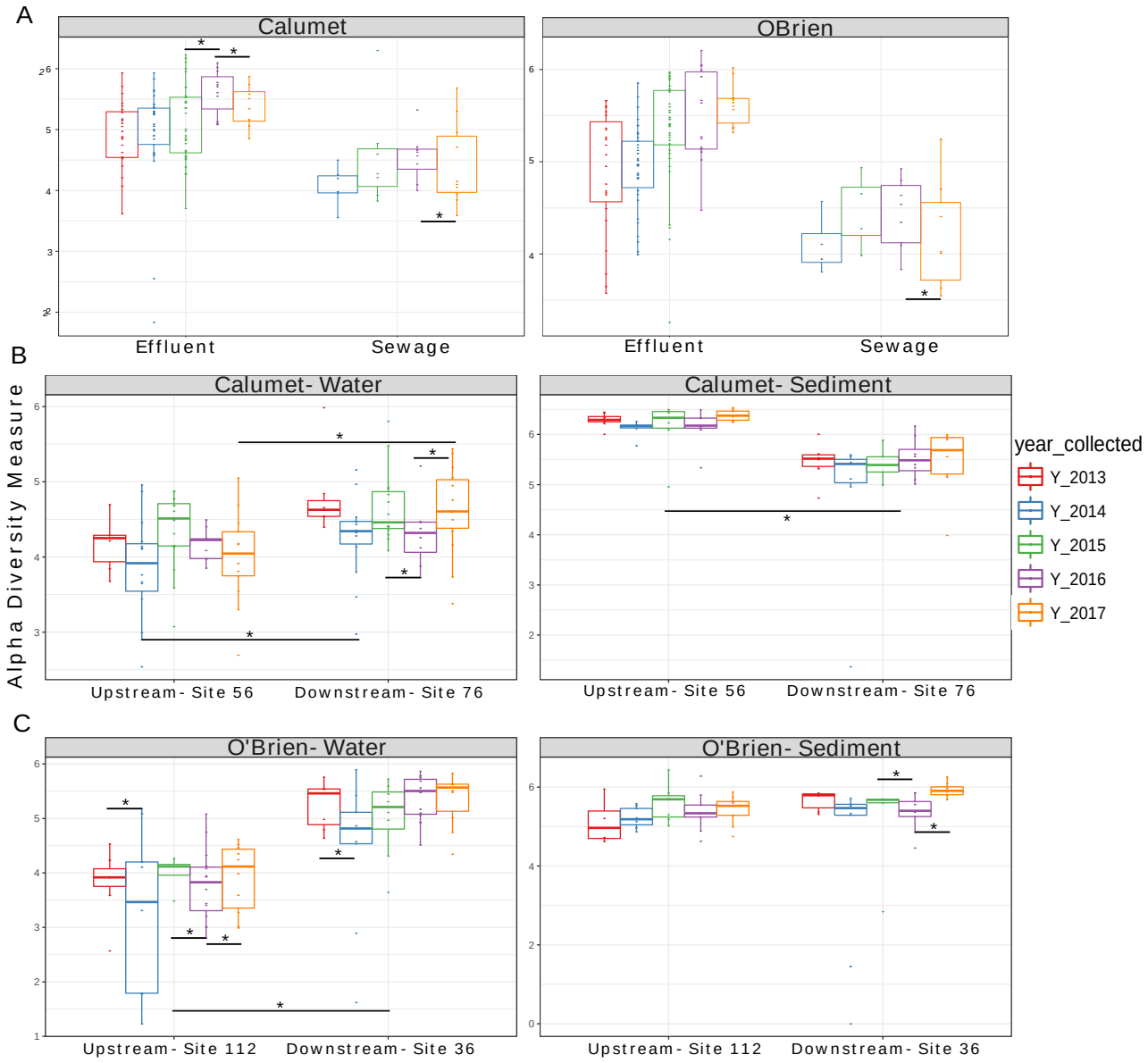


FIGURE 3 Shannon Alpha Diversity Indices of (A) Sewage, Effluent, (B-C) River Water and Sediment Samples from Calumet and O'Brien WRPs Over a Period of Five Years. The black bars with asterisks stand for pairwise comparisons with significant differences i.e. PERMANOVA *$p < 0.05$*. (A) At Calumet WRP, there was a significant increase in diversity of effluent in 2016 (when compared to 2015) and then significant decrease in diversity in 2017 (when compared to 2016). At O'Brien WRP, a significant decrease was observed from 2016 to 2017. (B) When comparing the river water and sediment samples from immediate upstream site (56) and downstream site (76) of Calumet WRP, overall the downstream water samples were characterized by an increase in diversity when compared to the upstream site, whereas the downstream sediment samples had lower diversity when compared to the upstream samples. Precisely, post-disinfection, the river water samples showed a significant decrease in diversity in 2016 (when compared to 2015) and then a significant increase in diversity in 2017 (when compared to 2016). However, there was no significant pattern for the sediment samples. (C) At O'Brien WRP, the diversity reduced ($p < 0.05$) in 2016 of the sediment samples and then increased in 2017. No significant trends were observed for the river water samples based on alpha diversity.

2.3.2 Compositional Variation Among River Water and Sediment Samples Collected Pre- and Post-Disinfection

Genomic Data and Culture Based Fecal Indicator Bacteria (FIB) Monitoring Method

Current USEPA limits for *E. coli* detection are 126 CFU / 100 mL as geometric mean, and 410 CFU / 100mL as a statistical threshold value. *E. coli* is the EPA-recommended indicator of fecal pollution, however fecal coliform (FC) continues to be widely used for bacteria quality monitoring. In Illinois, FC is tested per the National Pollutant Discharge Elimination System (NPDES) permit requirement for disinfected effluent to meet the monthly geometric mean of 200 CFU/ 100 ml and 10 % of sample to be less than 400 CFU/100 ml. The traditional plate count monitoring for FC bacteria levels in effluent at sites downstream of the O'Brien WRP (Fig. 4) and the Calumet WRP (Fig. 5) showed higher concentrations in the pre-disinfection period (2013 to 2015) compared to the post-disinfection period (2016-2017). There was significant FC bacteria reduction in the final effluent and immediately downstream of both WRPs, suggesting that disinfection was effective at reducing FC bacteria levels meeting the WRPs permit compliance required levels (Fig. 4 and 5). However, similar decreases in FC concentrations were not seen in river waters upstream of the WRPs and in tributaries. The FC concentration was found to be above the water quality standard in select tributaries and further downstream river locations.

As described above, plate counts indicated there was a decrease in FC counts immediately downstream (Site 36) of the O'Brien WRP from the pre-disinfection to the post-disinfection period. The microbial community analysis also indicated significant decreases in sewage associated bacteria at Site 36 following disinfection in 2016. These results emphasize the impact of the disinfection in greatly reducing sewage and human fecal indicators in the CAWS. We used non-parametric, two-group statistical tests to compare river water and sediment samples downstream of Calumet WRP (site 56- upstream, site 76- downstream) and O'Brien WRP (site 112-upstream and site 36-downstream), pre- (2013-2015) and post-disinfection (2016-2017). At Calumet WRP, sewage indicators such as *Arcobacter, Acinetobacter,* and *Rubrivivax* were significantly lower post-disinfection across all river water samples (Table 4 and Figure 5). The genus *Flavobacterium*, which is associated with freshwater, increased significantly post-disinfection (Eiler and Bertilsson 2007). Similarly, Calumet WRP area CAWS sediment showed a significant reduction of genera such as *Methylotenera* and *Rhodobacter* (Table 5 and Figure 5). The *Methylotenera* species are methanotrophs which are capable of methanol-dependent denitrification- a well described phenomenon occurring in sewage treatment plants (Mustakhimov et al. 2013).

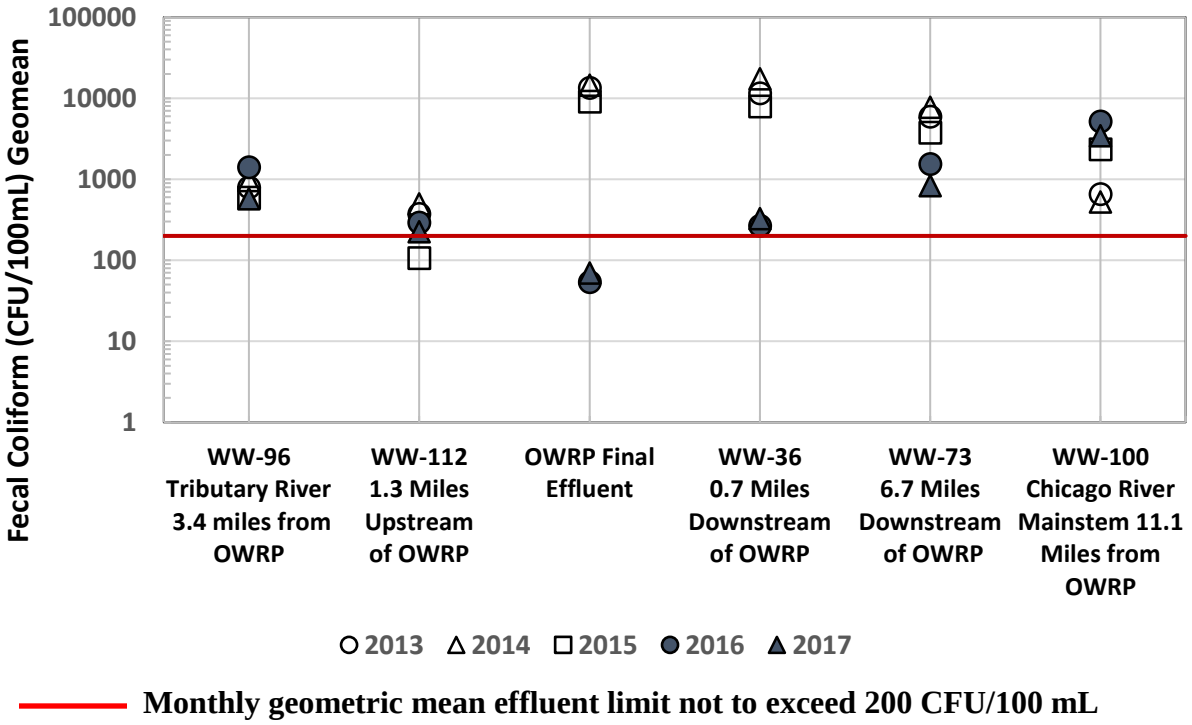


FIGURE 4 Geometric Mean Fecal Coliform Concentrations Observed March-November (2013-2017) in Final Effluent (UV treated), Upstream and Downstream of O'Brien WRP (OWRP), North Shore Channel and North Branch Chicago River Locations

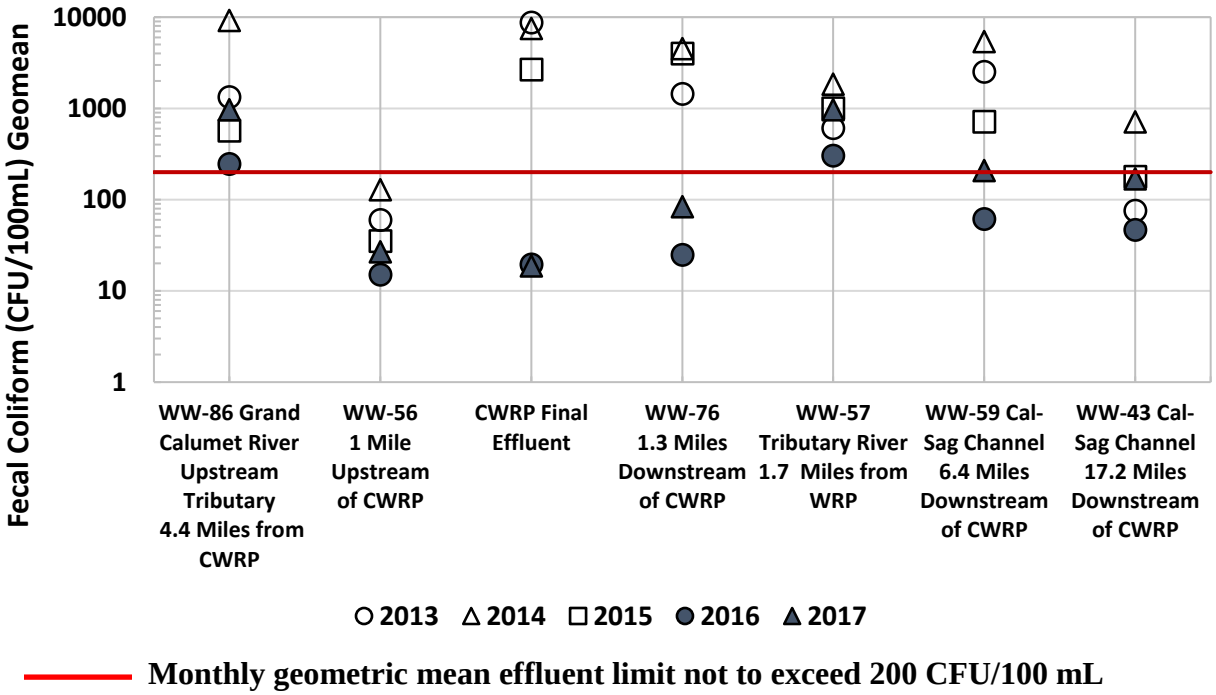


FIGURE 5 Geometric Mean Fecal Coliform Concentrations Observed March-November (2013-2017) in Final Effluent (chlorinated/dechlorinated treated), Upstream and Downstream of Calumet WRP (CWRP), Grand Calumet, Little Calumet, and Cal-Sag Channel River Locations

TABLE 4 List of Significantly (FDR[a] Corrected P-value < 0.05) Differential Genera between the Pre- (2013-2015) and Post-Disinfection Period (2016-2017) across CAWS Water Samples Collected at Downstream of the Calumet WRP (Site 76)

Genus	Pre: mean rel. freq. (%)	Post: mean rel. freq. (%)	p-values (corrected)
Acinetobacter	12.90644487	3.995119769	0.001121504
Arcobacter	10.52677164	5.28688027	0.029532862
Candidatus Xiphinematobacter	2.909938948	0.919121044	0.043291693
Cellvibrio	0.047804076	0.354479635	0.005516746
Flavobacterium	11.56586688	26.6662068	0.005526302
Gemmatimonas	0.42285776	0.196085057	0.044447796
Geobacter	0.080158157	0.182318373	0.042762295
Rubrivivax	0.685177455	0.137438901	0.016836849

[a] FDR-False Discovery Rate

TABLE 5 List of Significantly (FDR[a] Corrected P-value < 0.05) Differential Genera between the Pre- (2013-2015) and Post-Disinfection Period (2016-2017) across CAWS Sediment Samples Collected Downstream of the Calumet WRP (Site 76)

Genus	Pre: mean rel. freq. (%)	Post: mean rel. freq. (%)	p-values (corrected)
C1_B004	0.259176929	0.73244762	0.010147061
Candidatus Methanoregula	0.530009951	0.82932609	0.041394347
Crenothrix	1.006379281	0.419443938	0.016354514
Desulfobulbus	1.695571252	0.635342425	0.009156576
Methanobacterium	0.53907654	0.936060453	0.029247068
Methanosaeta	0.400049071	0.752905423	0.037467046
Methylotenera	2.366876287	0.119944423	0.023359578
Rhodobacter	0.625503018	0.202517475	0.035982652
SHD-231	0.572380273	1.024747306	0.039348063
WCHB1-05	1.894230548	4.487016887	0.000653432

[a] FDR-False Discovery Rate

At O'Brien WRP, the post-disinfection, downstream river water samples demonstrated a significant decrease in the abundance of genera such as *Acinetobacter* and *Claocibacterium*, which are known sewage indicators (Nouha, Kumar, and Tyagi 2016; Wiedmann-al-Ahmad, Tichy, and Schön 1994; Table 6 and Figure 6). There was an increase in the abundance of the genus, *Hydrogenophaga* which is known to be associated with waste-water treatment plants (Magic-Knezev, Wullings, and Kooij 2009). At O'Brien WRP CAWS area sediment samples also showed differential microbial signatures pre- and post-disinfection (Table 7 and Figure 6). Genera like *Anaerolinea*, *Bifidobacterium*, *Devosia*, and *Paracoccus* significantly reduced

post-disinfection whereas we observed a significant increase in the abundance of genera such as *Dechloromonas* and *Rhodoplanes*. The genera *Devosia* and *Paracoccus* are also known to be enriched in sludge across wastewater disinfection plants (Fan et al. 2017). *Dechloromonas* increased post-disinfection. Members of this genus are known to be a part of the bacterial community in wastewater treatment plants and significantly correlate with improved performance of wastewater treatment (Yang et al. 2011). Hence, *Dechloromonas* might have been introduced downstream due to disinfection process taking place at the WRPs.

TABLE 6 List of Significantly (FDR[a] Corrected P-value < 0.05) Differential Genera between the Pre- (2013-2015) and Post-Disinfection Period (2016-2017) across CAWS Water Samples Collected at Downstream of the O'Brien WRP (Site 36)

Genus	Pre: mean rel. freq. (%)	Post: mean rel. freq. (%)	p-values (corrected)
Acinetobacter	20.09028832	11.61669693	0.010649209
Blvii28	0.060352358	0.328137987	0.012667184
Cloacibacterium	3.082879083	1.101313115	0.047592908
Hydrogenophaga	2.231574851	3.648483378	0.005319124
Paludibacter	0.388134143	0.777058703	0.003909511
Thiothrix	0.347289946	0.933636726	0.028569958
Tolumonas	0.976228671	2.297157768	0.002381874

[a] FDR-False Discovery Rate

TABLE 7 List of Significantly (FDR[a] Corrected P-value < 0.05) Differential Genera between the Pre- (2013-2015) and Post-Disinfection Period (2016-2017) across CAWS Sediment Samples Collected at Downstream of the O'Brien WRP (Site 36)

Genus	Pre: mean rel. freq. (%)	Post: mean rel. freq. (%)	p-values (corrected)
Anaerolinea	1.533156685	2.698181816	0.029316676
Bifidobacterium	0.042293256	0.051825212	0.001512844
Dechloromonas	4.101596691	14.84149949	0.015113633
Desulfobacca	1.376984555	2.189353192	0.032972873
Devosia	0.851189813	0.423209334	0.026738279
Paracoccus	5.345063936	2.800263437	0.004915891
Prosthecobacter	0.577765459	1.718962053	0.022324131
Rhodoplanes	3.020667309	1.213667512	0.006684938
Rickettsiella	0.52932096	1.012592663	0.04083851
SHD-231	0.58933563	2.40563896	0.000929976
Spirochaeta	0.08695668	0.487537042	0.001591618
WCHB1-05	1.626627827	4.645426229	0.003557609

[a] FDR-False Discovery Rate

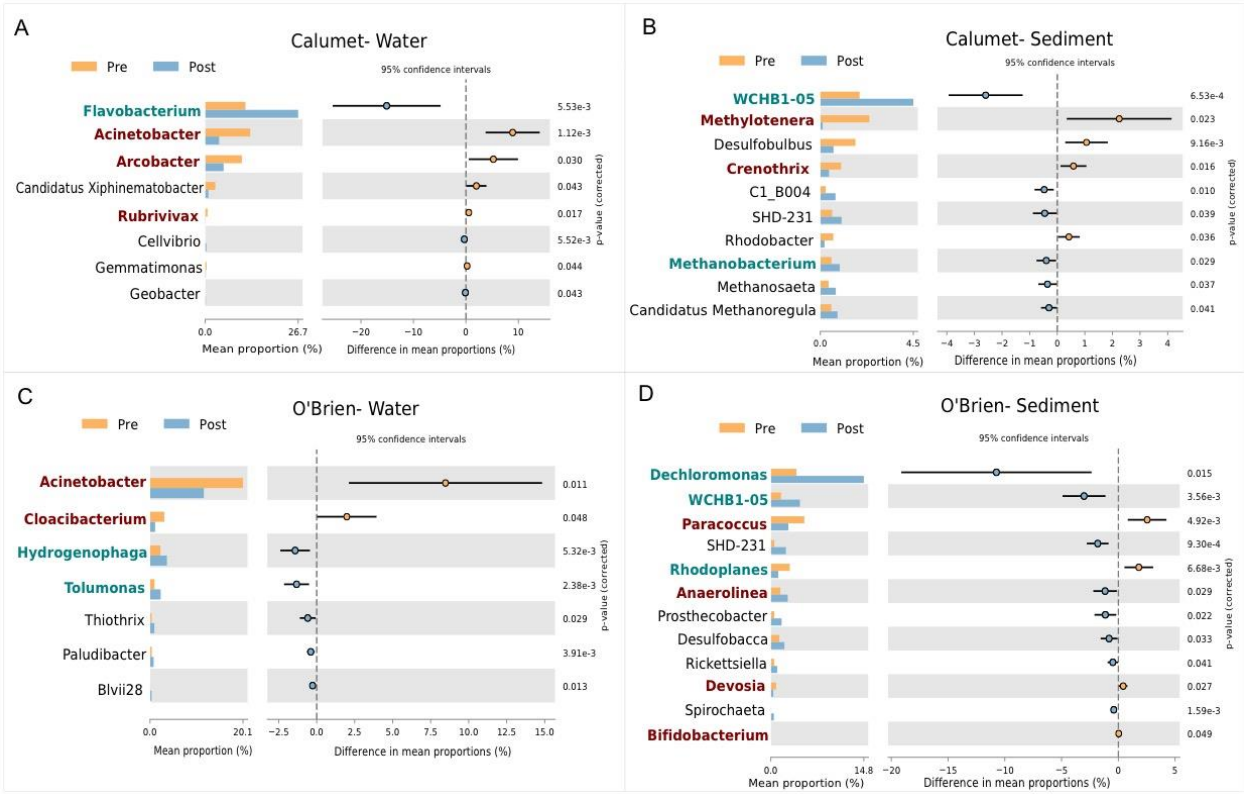


FIGURE 6 Non-Parametric Two Group Tests (ANCOM) Done Between Pre- and Post-disinfection River Water and Sediment Samples Downstream of the Two Water Reclamation Plants i.e. (A-B) Calumet WRP and (C-D) O'Brien WRP. This figure shows list of statistically differential bacterial genera with Benjamini-Hochberg FDR corrected p-values (< 0.05) labelled for each taxon. For all four different sample types, in each figure there are two sub-panels : i) Mean proportion (%), stands for relative abundance/proportion of the taxa in the data, ii) Difference in mean proportions (%) stands for the percentage increase or decrease of the specific taxa in one of the groups over the other group compared (which in this case are: Pre and Post disinfection. The indicators for sewage (e.g. *Acinetobacter, Cloacibacterium*) and human fecal material (e.g. *Arcobacter, Bifidobacterium*) have been highlighted using red color. Green colored labels represent the bacterial genera associated with fresh water (e.g. *Flavobacterium*) and also the ones which are known to be introduced by wastewater treatment plants (e.g. *Hydrogenophaga, WCHB1, Dechloromonas*). It was also interesting to note here that we identified were able to identify the fecal indicators such as *Bifidobacterium* prior to disinfection, however overall abundance of *Enterococcus* was very low (less than 0.005%) in the water samples.

Our results from the alpha diversity analyses demonstrated a significant decrease in 2016 i.e. immediately post-disinfection followed an increase in 2017 in both river water and sediment samples. These findings indicate at a potentially different community structure between both the years post-disinfection. Hence, we investigated differential microbial community between 2016 and 2017. Significant differences were also seen among the sediment and water samples between the two years of disinfection i.e. 2016 and 2017. At Calumet WRP, the year 2017 was characterized by a reduction of the genus *Bacteroides* and the phylum *TM7*, order *Streptophyta*, and MLE1-12, when compared to 2016 ($p_{FDR} < 0.05$) (Figure 7). *Bacteroides* is also a known marker of sewage pollution and has been used as a tracer of ecosystem health (Ahmed, Hughes, and Harwood 2016). Likewise, *TM7* is also known to be associated with wastewater sludge which significantly reduced in year 2017 (Ju and Zhang 2015). The order MLE1-12 belongs to phylum *Cyanobacterium* which has been recently found in water treatment systems in the Ohio river region (Stanish et al. 2016).

Similarly, the sediment samples downstream of the Calumet WRP showed significant differences between the two disinfection years i.e. 2016 and 2017. We observed a significant reduction of ESVs belonging to genera *Sediminibacterium, Epulopiscium*, family *Lachnospiraceae*, and order *Clostridales* in the year 2017 when compared to 2016 ($p_{FDR} < 0.05$) (Figure 7). *Lachnospiraceae* and *Clostrdiales* are used as human fecal indicators for the influent sewage samples (McLellan et al. 2013). However, there was a significant increase in genus *Rhodococcus* in 2017. Another, ESV belonging to the genus *Dehalococcoides* also showed a significant increase in the year 2017. *Dehalococcoides* are strictly anaerobic bacteria which are capable of metabolizing water pollutants (such as chlorinated ethenes, polychlorinated biphenyls) produced during water disinfection processes such as reductive dechlorination (Islam, Edwards, and Mahadevan 2010). This is interesting because Calumet WRP's disinfection process is based on chlorination and dechlorination.

Both Calumet and O'Brien WRPs had significantly different microbial signatures post-disinfection which can be attributed to different methods of disinfection. At O'Brien WRP, ESVs belonging to order *Clostridiales*, families *Bacteroidacaea, Paraprevotellacaea,* and *Sphingobacteraceae* were significantly reduced in 2017 across all river water samples, which are well known sewage indicators (Figure 7) (McLellan et al. 2013). Similarly, among the sediment samples, there was a significant decrease in abundance of ESVs belonging to phylum OD1, family *Lachospiraceae*, and *Paraprevotellace* downstream in 2017 when compared to 2016 (Figure 7). Phyla OD1 is a phosphorous metabolizing group which significantly correlates with the total phosphorous content of the waste-water sludge (Niu et al. 2015).

Additionally, we also investigated the taxonomic structure of the two WRPs downstream CAWS water samples, since each uses different methodologies for disinfection. While Calumet WRP used chlorination/de-chlorination technique, at O'Brien WRP, UV based disinfection was used. We identified genera with significantly different relative proportions between downstream water samples collected from Calumet and O'Brien WRPs. We observed differences between downstream water samples of O'Brien and Calumet WRPs in both pre-disinfection as well as post-disinfection. The differences seen before disinfection are not surprising since both these WRPs represent very different water channel systems. Pre-disinfection, we identified genera such as *Dechloromonas, Flavobacterium, Synechoccus,* and *Thalassiosira* to be significantly more abundant ($p<0.05$) at Calumet WRP (Table 8). However, bacterial genera namely *Delftia, Hydrogenophaga,* and *Methylibium* were more abundant in O'Brien WRP area water samples (Table 8).

As expected, we also identified genera which were significantly different between two WRPs (Calumet and O'Brien) post-disinfection period. Genera like *Acinetobacter, Arcobacter, Delftia,* and *Tolumonas* were significantly ($p<0.05$) higher at O'Brien WRP area (Table 9). However, *Flavobacterium, Polynucleobacter, Rhodobacter, Sediminbacter, Synechococcus,* and *Thalassiosira* were more abundant at Calumet WRP area (Table 9). These results suggest that after disinfection, between Calumet and O'Brien WRPs, O'Brien WRP has higher abundance of sewage indicators such as *Acinetobacter* and *Arcobacter*. On the other hand, at Calumet WRP, higher abundance of fresh water indicator such as *Flavobacterium and Polynucleobacter* were observed. This may be attributed to the completion of TCR in 2015, which captured the combined stormwater and sewage from Calumet WRP in rainy weather condition.

TABLE 8 Differentially Abundant Genera Between O'Brien WRP and Calumet WRP Downstream Water Samples Before Disinfection

Genus	O'Brien WRP_Post: mean rel. freq. (%)	Calumet WRP _Post: mean rel. freq. (%)	p-values (corrected)
Candidatus Xiphinematobacter	0.27	2.91	0.02
Crenothrix	0.01	2.06	0.02
Dechloromonas	0.65	2.37	0.02
Delftia	7.70	1.30	0.02
Flavobacterium	4.51	11.57	0.02
Haliscomenobacter	1.40	0.09	0.01
Hydrogenophaga	2.23	1.07	0.05
Methylibium	1.90	0.59	0.01
Synechococcus	0.01	3.75	0.02
Thalassiosira	0.02	2.20	0.02

TABLE 9 Differentially Abundant Genera Post-Disinfection

Genus	O'Brien WRP_Post: mean rel. freq. (%)	Calumet WRP _Post: mean rel. freq. (%)	p-values (corrected)
Acinetobacter	11.62	4.00	0.02
Arcobacter	14.44	5.29	0.02
Delftia	7.84	1.08	0.03
Flavobacterium	6.90	26.67	0.04
Hydrogenophaga	3.65	0.91	0.02
Polynucleobacter	1.04	3.88	0.02
Rhodobacter	0.90	2.19	0.01
Sediminibacterium	4.22	8.46	0.03
Synechococcus	0.02	6.46	0.02
Thalassiosira	0.10	2.85	0.03
Tolumonas	2.30	0.52	0.02

2.3.3 Determining the Sources of Microbial Community in the CAWS

We used a Bayesian statistical tool, SourceTracker 2.0 to determine the potential sources of microbial ESVs (exact genomic sequence variant) associated with each sample by sampling location and sampling period (Mustakhimov et al. 2013). SourceTracker determines the microbial signature of each environment that is unique and exclusive to that environment. The aim of this analysis is to determine the likely sources of microbes found in CAWS water column samples. For this analysis, a curated database was built using CAWS samples (effluent, sewage, sediment, fish gut and mucus) and additional 100,000 samples from the Earth Microbiome Project (EMP) 2017 release version. The sources from the EMP database included- animal feces, fresh water, soil, and stream sediment. This database was used to determine potential sources of microbes that occur in CAWS water column samples at each sampling site by sampling year (Figure 8). The sources of microbial diversity across all river water samples can be largely attributed to effluent, sewage, CAWS sediment, freshwater, and fish associated samples. The three CAWS regions i.e. North, Main, and Calumet regions, 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 (Figure 8).

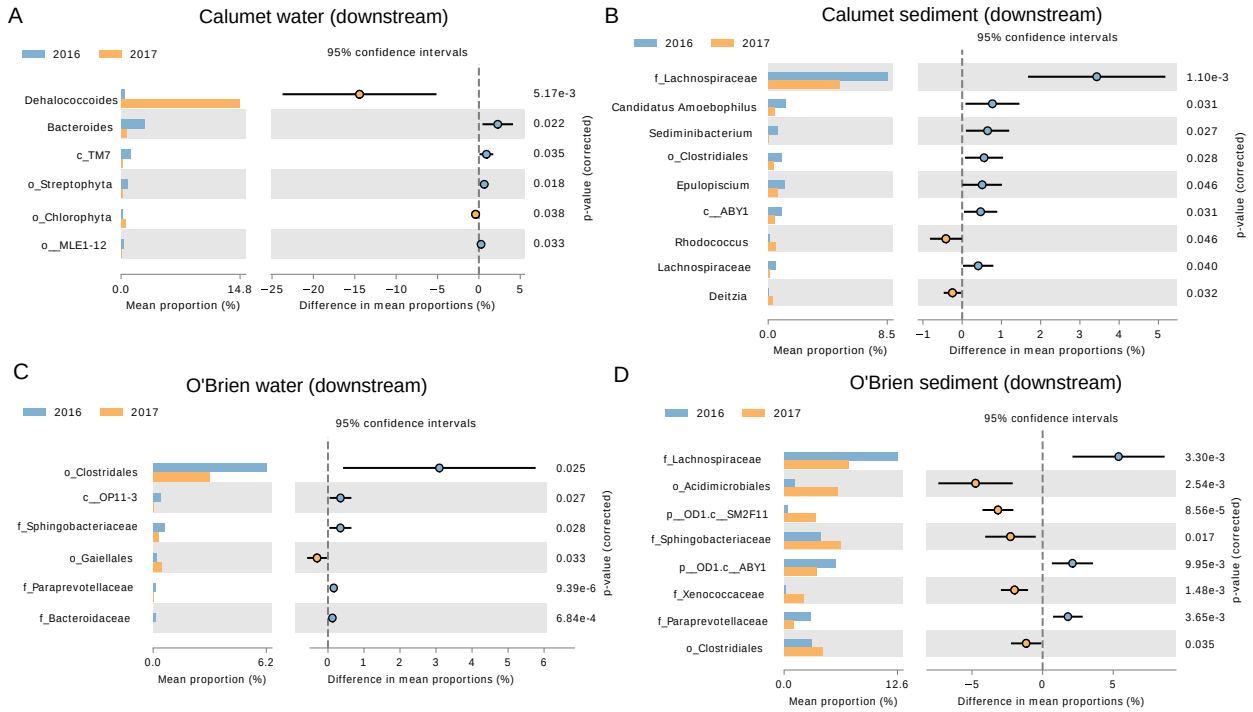


FIGURE 7 Non-Parametric Two Group Tests (ANCOM) Done between the Two Disinfection Years i.e. 2016 and 2017 for River Water and Sediment Samples Downstream of the Calumet WRP (A-B) and O'Brien WRP (C-D). This figure shows list of statistically differential bacterial genera with Benjamini-Hochberg FDR corrected p-values (< 0.05) labelled for each taxon. For all four different sample types, in each figure there are two sub-panels : i) Mean proportion (%), stands for relative abundance/proportion of the taxa in the data, ii) Difference in mean proportions (%) stands for the percentage increase or decrease of the specific taxa in one of the groups over the other group compared which in this case are: Pre and Post disinfection. The differentially abundant ESVs were assigned taxonomy status at different levels such as genus, class, family and order. At both Calumet and O'Brien WRPs, for river water samples, the ESVs belong to sewage indicators such as *Clostridiales*, families *Bacteroidacaea*, *Paraprevotellacaea,* and *Sphingobacteraceae* were significantly reduced in 2017. Similarly, the sediment samples demonstrated a significant decrease in ESVs belonging to families such as *Lachospiraceae* and *Paraprevotellace* which are also human fecal indicators.

Strikingly, the contribution made by human fecal matter across all water column samples was extremely low. Overall, there remains a large proportion of bacterial taxonomic diversity in the CAWS that cannot be reliably attributed to a 'source', as already reported for the sampling years from 2013 to 2015. This is reflective of the sampling bias observed in the EMP project wherein urban river microbiomes are underrepresented. Additionally, it is likely that these are endemic but extremely rare taxa that are only found in the Chicago River.

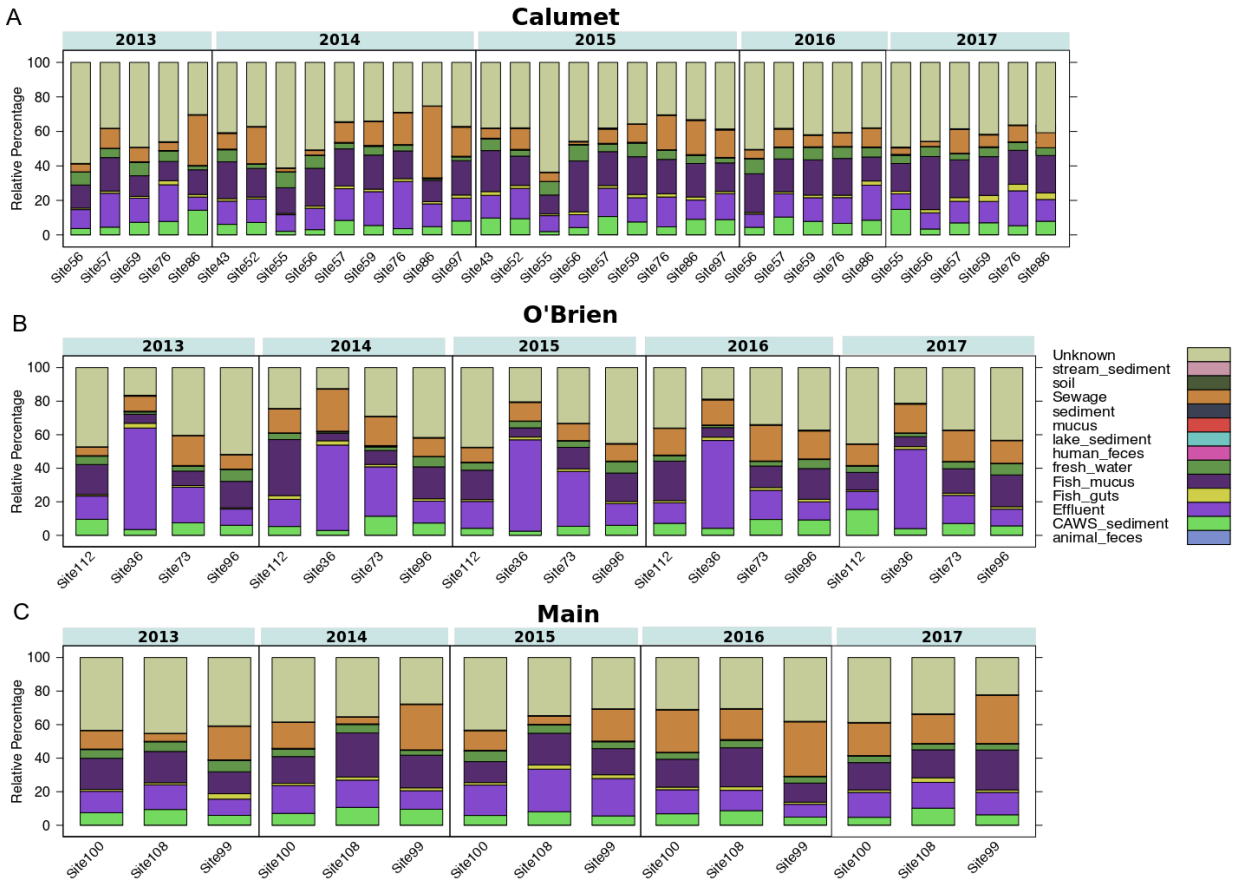


FIGURE 8 SourceTracker 2.0 Analysis of Water Column Samples by Sampling Site for Years 2013-2017 Using a Curated Database for (A) Calumet, (B) O'Brien, and (C) Main. A curated database was built using CAWS samples (i.e. effluent, sewage, sediment, fish gut and mucus) and additional 100,000 samples from the Earth Microbiome Project (EMP) 2017 release version. The sources from the EMP database included- animal feces, fresh water, soil, and stream sediment.

2.3.4 Bioball Experiment

The purpose of the Bioball experiment was to determine the depth of sequencing required to detect the *Escherichia coli* 16S rRNA gene in our 16S rRNA amplicon libraries. A known concentration of *E. coli* cells (10^2, 10^3, 10^4, 10^5, 10^6, and 10^7 cells) were added to 100 ml of each sample type. These values were picked because the current USEPA limits for *E. coli* detection are 126 CFU/100 mL as geometric mean, and 410 CFU/100 mL as a statistical threshold value. Therefore, we wanted to determine if we could detect *E. coli* abundances of ~100 cells per 100mL of water using our 16S rRNA amplicon detection techniques. We included four sample types: (i) one O'Brien WRP secondary treated effluent (STE) sample, (ii) a CAWS waterway (WW) column sample, (iii) a 0.22 µm filtered CAWS (filtered) water column sample, and (iv) a Phosphate buffered saline (PBS) control sample.

All samples were deeply sequenced on the Illumina MiSeq platform. Sequencing depth information is as reported in Table 10. To further improve our ability to recover *Escherichia* sequences from this dataset, we performed 16S rRNA amplicon analysis using a newer, more sensitive technique - DADA2 (Callahan et. al, 2016) with the RDP database (Cole et. al, 2014). To normalize for variable sequencing depth, all samples were rarefied to depth of 10,000 and 1000 sequences per sample. We observed that for both rarefaction depths, as few as 100 *E. coli* cells were reliably detected across all four sample types (Figures 9 and 10). Thus, the apparent lack of *E. coli* sequences in CAWS 16S rRNA amplicon datasets is a result of a true low abundance of *E. coli* and not inadequate sequencing or detection potential. These results are being confirmed with quantitative PCR targeting *E. coli* strains.

TABLE 10 Summary of Sequence Depth per Sample. Samples types include phosphate buffered saline control samples (PB), 0.22 µm filtered CAWS water column samples (Filtered WW), CAWS water column samples (WW), and O'Brien WRP secondary treated effluent samples (Obrien STE).

Sample Id	Sample type	*E. coli* cell conc.	Total reads
2475	PB	10^2	15535
2478	PB	10^3	87098
2481	PB	10^4	212003
2486	PB	10^5	290623
2489	PB	10^6	272934
2491	PB	10^7	279227
2495	Filtered WW	10^2	41585
2498	Filtered WW	10^3	55506
2501	Filtered WW	10^4	180643
2502	Filtered WW	10^5	265540
2507	Filtered WW	10^6	249031
2508	Filtered WW	10^7	283116
2511	WW	10^2	94197
2514	WW	10^3	106048
2517	WW	10^4	173137
2521	WW	10^5	251241
2524	WW	10^6	287219
2526	WW	10^7	300048
2529	Obrien STE	10^2	78489
2534	Obrien STE	10^3	51340
2536	Obrien STE	10^4	168308
2539	Obrien STE	10^5	255279
2541	Obrien STE	10^6	212176
2545	Obrien STE	10^7	250512

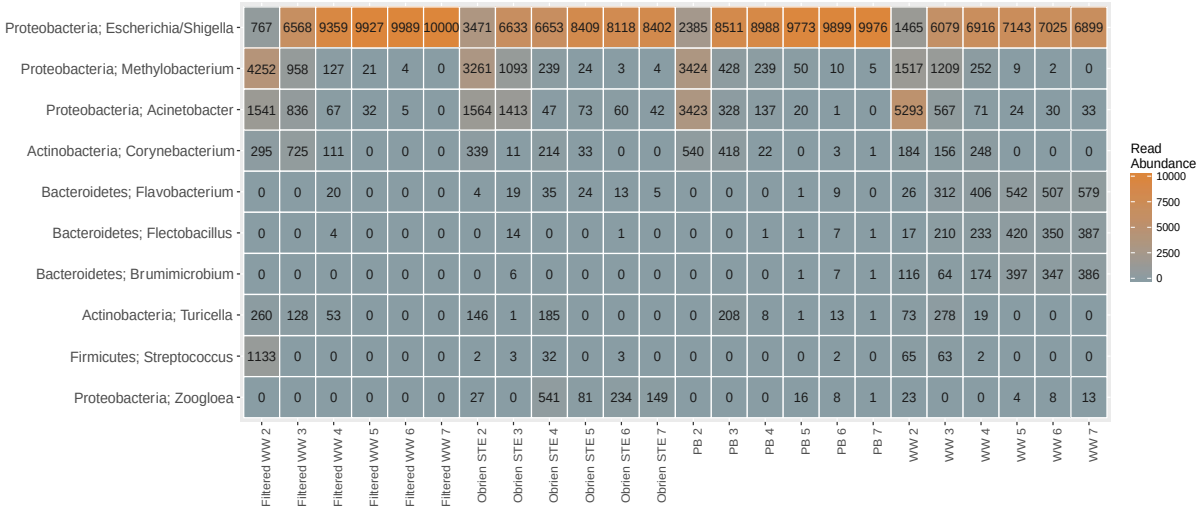


FIGURE 9 Heatmap Showing Distribution of 10,000 Sequence Reads Assigned to the 10 Most Abundant Bacterial Genera by Sample. Filtered WW = 0.22 um filtered CAWS water column samples; O' Brien STE = O'Brien WRP secondary treated effluent samples; PB = PBS control samples; and WW = CAWS water column samples. The numbers 2, 3, 4, 5, 6, and 7 refer to *E. coli* cell concentrations 10^2, 10^3, 10^4, 10^5, 10^6, and 10^7cells, respectively.

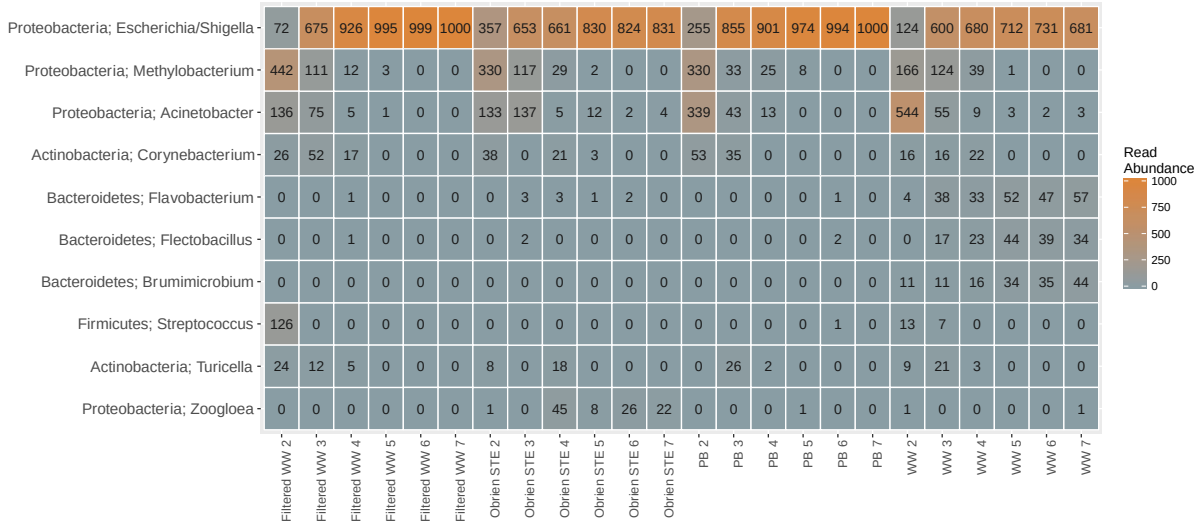


FIGURE 10 Heatmap Showing Distribution of 1000 Sequence Reads Assigned to the 10 Most Abundant Bacterial Genera by Sample. Filtered WW = 0.22 um filtered CAWS water column samples; Obrien STE = O'Brien WRP secondary treated effluent samples; PB = PBS control samples; and WW = CAWS water column samples. The numbers 2, 3, 4, 5, 6, and 7 refer to *E. coli* cell concentrations 10^2, 10^3, 10^4, 10^5, 10^6, and 10^7cells, respectively.

2.3.5 Taxonomic and Functional Annotations of Metagenomic Samples

Twenty-four CAWS water column samples from 2014 and 2015 were selected from Calumet WRP region for deep metagenome sequencing (Table 11). These were selected from a candidate list of 54 to represent water column samples upstream and downstream of the Calumet WRP. Selection criteria were based on the (expected) strain level variation in sewage, human and non-human fecal indicators. We used these shotgun data for a complete taxonomic and functional characterization of all samples. Based on taxonomic signatures, we will reconstruct bacterial genomes (species/strain level accuracy) from these samples and check for strain-level variation which might be modulated by sampling season and/or wastewater disinfection regimes. Also, it is important to note that these were the samples collected from 2014-2015. We next aim to select a subset of samples from 2016-2017 to include samples post-disinfection for comparison.

TABLE 11 Summary of Samples Chosen with Reference to Calumet WRP for Deep Metagenome Sequencing

Sample ID	Site Code	Miles from WRP	Reference to WRP	Sampling Period	Rainfall
521	76	1.24	Downstream	14-Aug	Wet
380	76	1.24	Downstream	14-Jul	Wet
850	76	1.24	Downstream	15-Jun	Wet
650	76	1.24	Downstream	15-Mar	Wet
325	76	1.24	Downstream	14-Apr	Wet
891	76	1.24	Downstream	15-Aug	Dry
736	76	1.24	Downstream	15-May	Wet
294	76	1.24	Downstream	14-May	Wet
696	76	1.24	Downstream	15-May	Dry
351	76	1.24	Downstream	14-Apr	Wet
710	57	1.74	Downstream	15-Apr	Dry
847	59	6.39	Downstream	15-Jun	Wet
739	59	6.39	Downstream	15-May	Dry
528	59	6.39	Downstream	14-Aug	Wet
732	56	1.05	Upstream	15-May	Dry
484	55	5.74	Upstream	14-Aug	Wet
371	55	5.74	Upstream	14-Jul	Wet
842	55	5.74	Upstream	15-Jun	Wet
481	86	4.43	Upstream	14-Aug	Wet
386	86	4.43	Upstream	14-Jul	Wet
7161239	86	4.43	Upstream	14-Jun	Wet
297	86	4.43	Upstream	14-May	Wet
2079.750774	86	4.43	Upstream	15-Oct	-
316	86	4.43	Upstream	14-Apr	Wet

We taxonomically characterized the 24 samples from Calumet WRP using MG-RAST. Overall, these samples contained an average of 99.03% bacteria, 0.6% eukaryotes, 0.2% archaea, 0.1% viruses, and the remainder 0.006% of sequences were unclassified. Among bacteria, at phylum level, *Proteobacteria* dominated all samples with an average of 56% abundance, followed by *Bacteroidete*s (22%), *Actinobacteria* (13%), and *Firmicutes* (2%). At the family level, these samples showed a consistent pattern with *Comamonadaceae* being the most dominant taxon (20%). Recently, a study focused on identifying the core community across several activated sludge systems demonstrated high activity or growth rate of one particular ESV belonging to the *Comamonadaceae* (Saunders et al. 2016). This result supports the dominance of this taxon in our dataset. We observed a similar compositional pattern across all samples at higher taxonomic levels. At the genus level, however, these samples showed significant sample-to-sample variation and differential abundance patterns between upstream *vs.* downstream groups (Figure 11). Of the 2000 bacterial genera annotated across all samples, we plotted the 30 most variable genera between the sample sites (Figure 11). Members of the genus, *Polynucleobacter* were the most variable between sample sites (Figure 11). *Polynucleobacter* was more abundant in upstream samples, particularly at site 55 that is 5 miles upstream as compared to those collected downstream from Calumet WRP. (Figure 11). *Acinetobacter* was the most abundant a mile upstream from Calumet WRP, there on we observe a decrease in its abundance downstream (Figure 11). Member of the genus, *Thalassiosira* belonging to the family of diatoms showed an increase in abundance downstream of Calumet WRP. This genus is frequently observed in effluents discharged from wastewater treatment plants. Member of *Flavobacterium*, a freshwater organism was also increased in samples immediately downstream of Calumet WRP. Genera like *Dechloromonas, Aquibacter,* and *Aeromonas* showed an increase in abundance in downstream sites. Their abundance can be attributed to the Calumet WRP effluent since these genera are known to be enriched in wastewater treatment sludge (Figure 11). These preliminary results indicate variable abundance patterns of different genera at upstream and downstream sites.

We next grouped the upstream and downstream river water samples to identify statistically significant candidates which differentiate these groups. In addition to *Flavobacterium,* we identified other genera that also demonstrated a significant increase in their abundances at downstream sites, including *Burkholderia, Rhodococcus, Methylobacterium, Methylibium,* and *Alicycliphilus* (Figure 12A). Interestingly, multiple species from these genera have been previously used for the degradation and remediation of organic contaminants. *Burkholderia* and *Rhodococcus* species are known to degrade a diverse set of organic contaminants such as chlorinated solvents, herbicides, and pesticides (Wang et al. 2014). *Alicycliphilus* species, specifically *A. denitrificans*, are cyclohexanol-degrading nitrate-reducing betaproteobacterium known for their ability to remediate chlorine contaminated water (Weelink et al. 2008). *Methylobacterium* species have been isolated from wastewater plants and are utilized to degrade organo-halogenated pollutant dichloromethane (Mustakhimov et al. 2013). Additionally, a study on the Zenne river in Belgium demonstrated an enrichment in these taxa in post-wastewater disinfection water column and sediment samples, which were associated with a depletion in organic carbon and nitrogen sources and an increase in oxygen concentration; these conditions also drove a significant decrease in the abundance of sulfate reducers and methanogens (Atashgahi et al. 2015). We also separately analyzed and compared sites immediately upstream and downstream from Calumet WRP (i.e. 1 mile) and observed different

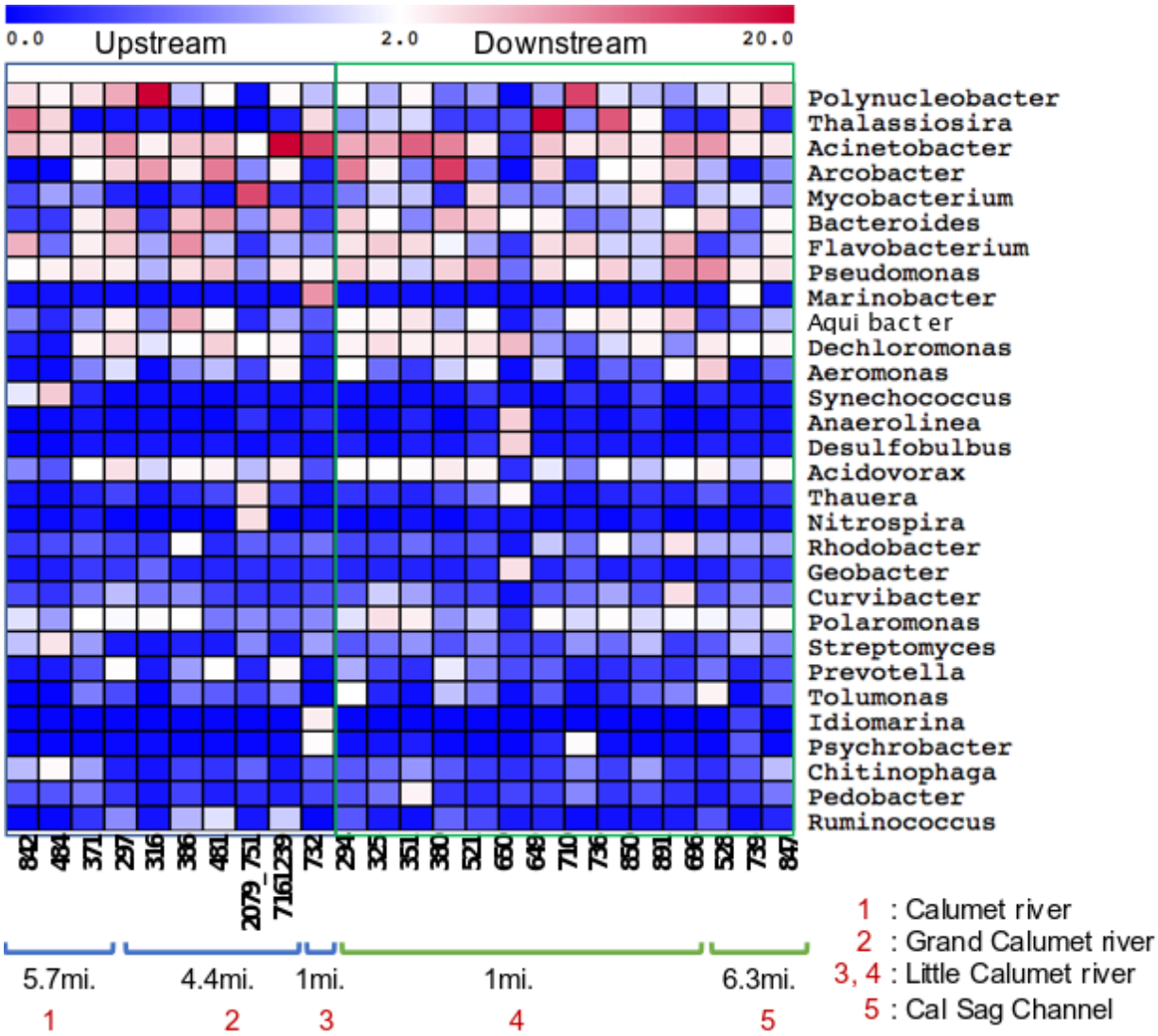


FIGURE 11 Heatmap Showing the Relative Abundance of the 30 Most Variable Bacterial Genera Distributed Across the Three Upstream and Three Downstream Sites from the Calumet WRP. Samples were collected from three different rivers with sample sites classified as upstream or downstream according to their location in relation to the Calumet WRP (Table 2 and Table 3). Site 55 (1 Calumet River, 5.74 mile upstream), Site 86 (2- Grand Calumet River, 4.4 miles upstream), 56 (3- Little Calumet River upstream, 1 mile), 76 (4- Little Calumet River Downstream, 1.24 miles), 57 (4- Little Calumet River, 1.7 miles downstream), and 59 (5- Cal Sag Chanel, Downstream).

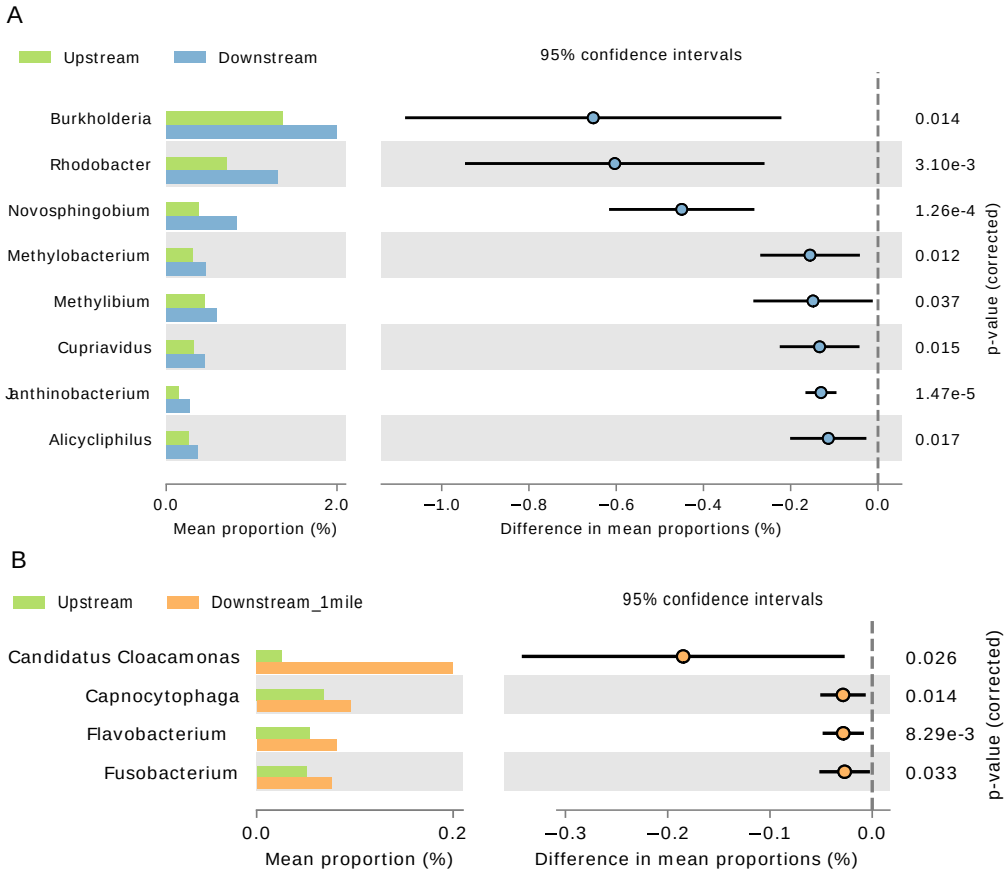


FIGURE 12 Two-Group Tests (Welch's t-test) Performed to Identify Statistically Differential Bacterial Genera between Sites, Upstream and Downstream of the Calumet WRP. This figure shows a list of statistically differential bacterial genera with Benjamini-Hochberg FDR corrected p-values (< 0.05) labelled for each taxon. A significant increase was observed for genera like *Burkholderia, Methylobacterium, Methylibium,* which have been used for degradation and remediation of organic contaminants in the water systems. These results also highlight the increased bioremediation potential of the downstream samples in comparison to the upstream samples after processing in the Calumet WRP.

microbial signatures. For example, the genus Candidatus *Cloacamonas*, often observed in wastewater disinfection plants (Siezen and Galardini 2008), showed a significant increase in its abundance downstream as compared to its abundance upstream. Overall, these results demonstrate the significant impact of wastewater disinfection on downstream CAWS microbial communities.

We further annotated the 24 shotgun samples at a subsystem level using the SEED database employed in MG-RAST. In contrast to their taxonomy, the samples were functionally conserved and there was no significant variation between the upstream and downstream samples. Overall, the abundant functional categories included amino acid associated pathways

(biosynthesis and metabolism), carbohydrate metabolism, protein metabolism, and RNA metabolism (Figure 13). Interestingly, pathways related to phages, and other transposable elements were the most variable between sites, however, there was no specific enrichment pattern between the upstream and downstream samples. Pathways including sulfur metabolism, nitrogen metabolism phosphorous metabolism, iron transport, and secondary metabolism were the most consistent across all the samples (Figure 13). The functional category for virulence was low in abundance (<2%) across all the samples.

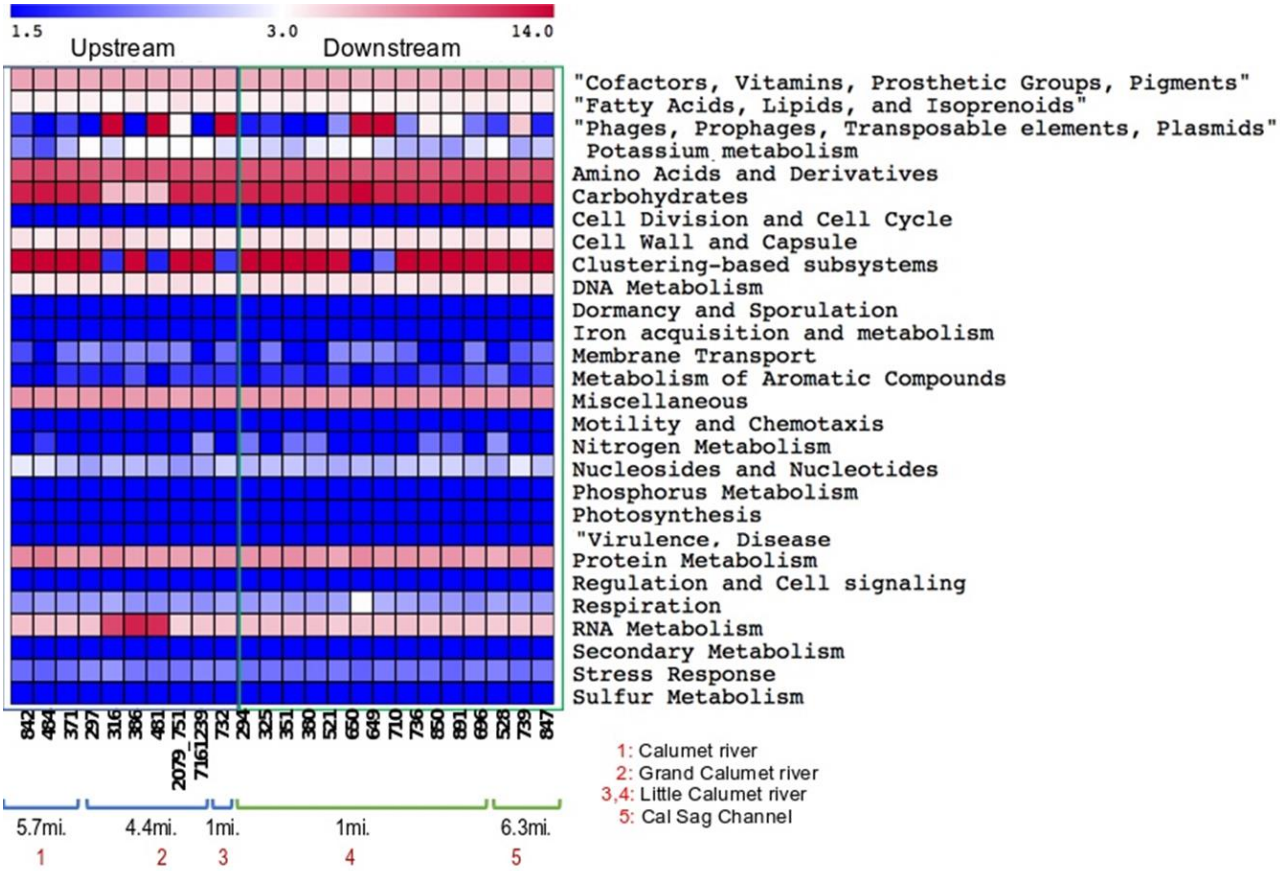


FIGURE 13 Heatmap Showing Relative Abundance of the SEED Subsystems Annotated from the 24 Shotgun Samples. Samples were collected from three different rivers with sample sites classified as upstream or downstream according to their location in relation to the Calumet WRP (Table 2 and Table 3). Site 55 (1 Calumet River, 5.74 mile upstream), Site 86 (2- Grand Calumet River, 4.4 miles upstream), 56 (3- Little Calumet River upstream, 1 mile), 76 (4- Little Calumet River Downstream, 1.24 miles), 57 (4- Little Calumet River, 1.7 miles downstream), and 59 (5- Cal Sag Chanel, Downstream).

We next aim to select a subset of samples from the years 2016-2017 in order to study the affect of disinfection and the phased TARP completion in modulating the taxonomic structure (at species/strain level resolution) as well functional genes. This will be done for both O'Brien and Calumet WRPs.

2.4 CONCLUSIONS

We performed a comparative analysis to investigate the effects of disinfection as well as phased TARP completion during 2016 and 2017 using sampling years, 2013-2015 as the baseline (pre-disinfection). Based on sample type, sediment samples were the most diverse followed by influent sewage, final disinfected effluent, and the CAWS river water samples ($p < 0.05$). Similarly, beta diversity analyses based on weighted and unweighted UniFrac distance matrices, demonstrated a clear, distinct separation of river water, sediment, and effluent samples when ordinated using PCoA plots ($p = 0.001$). We did not observe significant temporal changes in microbial community composition and structure within each sample type from 2013 to 2015. However, the post-disinfection years of 2016 and 2017 were characterized by significant differences for river water, sediment, effluent, and sewage samples when compared to pre-disinfection samples (2013-2015). For river water and sediment samples collected across the CAWS from 2016-2017, a significant increase ($p < 0.05$) in alpha diversity was observed in 2016 when compared to 2015 followed by a significant increase ($p < 0.05$) in 2017 when compared to 2016. However, for effluent and sewage, we observed a different pattern, whereby the alpha diversity increased in 2016 followed by a significant decrease in 2017 ($p < 0.05$). It is uncertain why there was a difference in diversity trends between the different sample types.

Using the compositional analyses of the 16S data, we identified a significant reduction of well-established sewage and human fecal indicators such as *Acinetobacter*, *Cloacibacterium*, *Bifidobacterium*, and *Clostridiales* post disinfection. We also observed significant differences between the two years 2016 and 2017. We observed a further reduction in sewage indicators such as families *Lachnospiracae*, *Paraprevotellacae*, *Bacteroides*, and *Clostridiales* in 2017 when compared to 2016. These results suggest that the disinfection process is significantly impacting the downstream river water systems by reducing the sewage and fecal indicators. The sequencing depth used across all CAWS-associated 16S amplicon analyses were appropriate given our ability to detect rarer organisms. The Bioball® experiment demonstrated that as few as 100 *E. coli* cells can be reliably detected across all four spiked, sample types at a sequencing depth as low as 1000 sequences per sample.

We then annotated 24 shotgun samples from Calumet WRP CAWS locations (10 upstream and 14 downstream) in 2013-2015. In addition to the genus, *Flavobacterium* (fresh marker) which was associated with downstream samples in our 16S data analyses, we identified many other genera which demonstrated significant increase in downstream sites, such as *Burkholderia*, *Rhodococcus*, *Methylobacterium*, *Methylibium*, *Alicycliphilus*. Interestingly, multiple species from these genera have been used for the degradation and remediation of organic contaminants. Overall, the results demonstrated a significant impact of Calumet WRP and TARP on downstream CAWS in detoxification and improving the health of the CAWS ecosystem. In contrast to the variable microbial diversity, both upstream and downstream sites

were functionally more conserved with no significant differences identified at the subsystem level. The most abundant functional categories included amino acid associated pathways (biosynthesis and metabolism), carbohydrate metabolism, protein metabolism, and RNA metabolism. Interestingly, pathways related to phages, and other transposable elements were the most variable between sampling sites, however there was no specific enrichment pattern between the upstream and downstream samples. Pathways including sulfur metabolism, nitrogen metabolism phosphorous metabolism, iron transport, secondary metabolism were the most consistent across all the samples and were relatively less abundant. Overall, these results emphasize the impact of disinfected effluent as well as the phased TARP completion in modifying the CAWS microbial diversity downstream to Calumet WRP.

Future Goals for the Year 2019

1. We will utilize the shotgun data to reconstruct microbial genomes in order to analyze the species/strain specific metabolic trade-offs between significant community members. The whole genome sequences (draft or complete) from the shotgun data are important for functional genomic analyses which can help us in assessing specific impact of environmental changes (dry/wet) and the MWRD improvement efforts (disinfection and TARP completion) on the CAWS bacterial species/strains.

2. We will increase the functional resolution by analyzing the metagenome sequences at gene and pathways using de novo metagenome assembly. This will be performed using the shotgun data already generated for 24 samples and additional samples that will be sequenced from 2016 and 2017, which will include samples from the O'Brien WRP. The sub-system level annotations are very broad and therefore functional pathways and gene-level annotations will provide a better resolution of the functional potential of the residing bacterial population.

3. We will analyze the shotgun metagenome assemblies for genes associated with pathogenicity. The assemblies will be screened for virulent genes using Virulence Factor Database (VFDB) which is an integrated and comprehensive online resource for curating information about virulence factors of bacterial pathogens (Chen et al. 2016). VFDB includes genes belonging to 6 broad classes- toxins, adherence and invasion associated, Type III, IV, V, VI, VII secretion systems, defensive virulence factors, and genes associated with regulation of virulence genes.

4. We will target fecal indicator bacteria for qPCR quantification. We have selected following targets:

    a. *E. coli*

b. *Enterococci* (we will use already established primers to amplify *Enterococcus faecium*, *Enterococcus faecalis*, and *Enterococcus casseliflavus*. These species are well known fecal indicators)
c. *Bacteroides* (HF183 primer will be used to target human specific fecal *Bacteroides* species mainly *Bacteroides fragilis* and *Bacteroides thetaiotamicron*).
d. *Bifidobacterium adoloscentis* will be targeted as potential indicator of human fecal pollution in environmental waters.

5. We will further select more samples for shotgun sequencing, preferably the same upstream and downstream sites already sequenced but from the years 2016 and 2017. We will also include samples from 2018 and 2019, this will provide us with a baseline comparison between pre,post-disinfection, and the phased TARP completion. The current sample set is from Calumet WRP and therefore we will extend this sample set to include upstream and downstream samples from O'Brien WRP as well. In addition, we plan to include comparison between dry and wet weather samples from upstream and downstream locations from two WRPs.

6. We will continue to collect samples for 2018 and 2019 to determine if diversity and functional trends seen post-disinfection and with TARP completion are maintained. We will specifically focus on effluent, sewage and water samples since sediment samples stay stable over time.

2.5 REFERENCES

Ahmed, Warish, Bridie Hughes, and Valerie J. Harwood. 2016. "Current Status of Marker Genes of Bacteroides and Related Taxa for Identifying Sewage Pollution in Environmental Waters." *Water* 8 (6): 231. https://doi.org/10.3390/w8060231.

Al Atrouni, Ahmad, Marie-Laure Joly-Guillou, Monzer Hamze, and Marie Kempf. 2016. "Reservoirs of Non-Baumannii Acinetobacter Species." *Frontiers in Microbiology* 7 (February). https://doi.org/10.3389/fmicb.2016.00049.

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.

Atashgahi, Siavash, Rozelin Aydin, Mauricio R. Dimitrov, Detmer Sipkema, Kelly Hamonts, Leo Lahti, Farai Maphosa, et al. 2015. “Impact of a Wastewater Treatment Plant on Microbial Community Composition and Function in a Hyporheic Zone of a Eutrophic River.” *Scientific Reports* 5 (November): 17284. https://doi.org/10.1038/srep17284.

Benson, Dennis A., Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2013. “GenBank.” *Nucleic Acids Research* 41 (Database issue): D36-42. https://doi.org/10.1093/nar/gks1195.

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.

Brandt, Jakob, and Mads Albertsen. 2018. “Investigation of Detection Limits and the Influence of DNA Extraction and Primer Choice on the Observed Microbial Communities in Drinking Water Samples Using 16S RRNA Gene Amplicon Sequencing.” *Frontiers in Microbiology* 9 (September). https://doi.org/10.3389/fmicb.2018.02140.

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.

Carey, Richard O., and Kati W. Migliaccio. 2009. “Contribution of Wastewater Treatment Plant Effluents to Nutrient Dynamics in Aquatic Systems: A Review.” *Environmental Management* 44 (2): 205–17. https://doi.org/10.1007/s00267-009-9309-5.

Castelino, Madhura, Stephen Eyre, John Moat, Graeme Fox, Paul Martin, Pauline Ho, Mathew Upton, and Anne Barton. 2017. “Optimisation of Methods for Bacterial Skin Microbiome Investigation: Primer Selection and Comparison of the 454 versus MiSeq Platform.” *BMC Microbiology* 17 (January). https://doi.org/10.1186/s12866-017-0927-4.

Chen, Lihong, Dandan Zheng, Bo Liu, Jian Yang, and Qi Jin. 2016. “VFDB 2016: Hierarchical and Refined Dataset for Big Data Analysis--10 Years On.” *Nucleic Acids Research* 44 (D1): D694-697. https://doi.org/10.1093/nar/gkv1239.

Doughari, Hamuel James, Patrick Alois Ndakidemi, Izanne Susan Human, and Spinney Benade. 2011. “The Ecology, Biology and Pathogenesis of Acinetobacter Spp.: An Overview.” *Microbes and Environments* 26 (2): 101–12.

Drury, Bradley, Emma Rosi-Marshall, and John J. Kelly. 2013. “Wastewater Treatment Effluent Reduces the Abundance and Diversity of Benthic Bacterial Communities in Urban and Suburban Rivers.” *Applied and Environmental Microbiology* 79 (6): 1897–1905. https://doi.org/10.1128/AEM.03527-12.

Eiler, Alexander, and Stefan Bertilsson. 2007. “Flavobacteria Blooms in Four Eutrophic Lakes: Linking Population Dynamics of Freshwater Bacterioplankton to Resource Availability.” *Applied and Environmental Microbiology* 73 (11): 3511–18. https://doi.org/10.1128/AEM.02534-06.

Fan, Xiangyu, Ying Zhu, Pengfei Gu, Yumei Li, Guiqing Xiao, Dongxue Song, Yiwei Wang, Rong He, Huajun Zheng, and Qiang Li. 2017. “Bacterial Community Compositions of Propylene Oxide Saponification Wastewater Treatment Plants.” *RSC Advances* 7 (36): 22347–52. https://doi.org/10.1039/C6RA27808F.

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.

Gücker, Björn, Mario Brauns, and Martin T. Pusch. 2006. “Effects of Wastewater Treatment Plant Discharge on Ecosystem Structure and Function of Lowland Streams.” *Journal of the North American Benthological Society* 25 (2): 313–29. https://doi.org/10.1899/0887-3593(2006)25[313:EOWTPD]2.0.CO;2.

Islam, M. Ahsanul, Elizabeth A. Edwards, and Radhakrishnan Mahadevan. 2010. “Characterizing the Metabolism of Dehalococcoides with a Constraint-Based Model.” *PLOS Computational Biology* 6 (8): e1000887. https://doi.org/10.1371/journal.pcbi.1000887.

Ju, Feng, and Tong Zhang. 2015. “Bacterial Assembly and Temporal Dynamics in Activated Sludge of a Full-Scale Municipal Wastewater Treatment Plant.” *The ISME Journal* 9 (3): 683–95. https://doi.org/10.1038/ismej.2014.162.

Kanehisa, Minoru, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. 2012. “KEGG for Integration and Interpretation of Large-Scale Molecular Data Sets.” *Nucleic Acids Research* 40 (Database issue): D109-114. https://doi.org/10.1093/nar/gkr988.

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.

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.

Lu, Xiao-Ming, and Peng-Zhen Lu. 2014. “Characterization of Bacterial Communities in Sediments Receiving Various Wastewater Effluents with High-Throughput Sequencing Analysis.” *Microbial Ecology* 67 (3): 612–23. https://doi.org/10.1007/s00248-014-0370-0.

Magic-Knezev, A., B. Wullings, and D. Van der Kooij. 2009. "Polaromonas and Hydrogenophaga Species Are the Predominant Bacteria Cultured from Granular Activated Carbon Filters in Water Treatment." *Journal of Applied Microbiology* 107 (5): 1457–67. https://doi.org/10.1111/j.1365-2672.2009.04337.x.

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.

Marchler-Bauer, Aron, Chanjuan Zheng, Farideh Chitsaz, Myra K. Derbyshire, Lewis Y. Geer, Renata C. Geer, Noreen R. Gonzales, et al. 2013. "CDD: Conserved Domains and Protein Three-Dimensional Structure." *Nucleic Acids Research* 41 (Database issue): D348-352. https://doi.org/10.1093/nar/gks1243.

Markowitz, Victor M., I.-Min A. Chen, Krishna Palaniappan, Ken Chu, Ernest Szeto, Yuri Grechkin, Anna Ratner, et al. 2012. "IMG: The Integrated Microbial Genomes Database and Comparative Analysis System." *Nucleic Acids Research* 40 (Database issue): D115-122. https://doi.org/10.1093/nar/gkr1044.

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, Sandra L., Ryan J. Newton, Jessica L. Vandewalle, Orin C. Shanks, Susan M. Huse, A. Murat Eren, and Mitchell L. Sogin. 2013. "Sewage Reflects the Distribution of Human Faecal Lachnospiraceae." *Environmental Microbiology* 15 (8): 2213–27. https://doi.org/10.1111/1462-2920.12092.

Meyer, F., D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, et al. 2008. "The Metagenomics RAST Server – a Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes." *BMC Bioinformatics* 9 (1): 386. https://doi.org/10.1186/1471-2105-9-386.

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.

Niu, Lihua, Yi Li, Peifang Wang, Wenlong Zhang, Chao Wang, and Qing Wang. 2015. "Understanding the Linkage between Elevation and the Activated-Sludge Bacterial Community along a 3,600-Meter Elevation Gradient in China." *Applied and Environmental Microbiology* 81 (19): 6567–76. https://doi.org/10.1128/AEM.01842-15.

Nouha, Klai, Ram Saurabh Kumar, and R. D. Tyagi. 2016. “Heavy Metals Removal from Wastewater Using Extracellular Polymeric Substances Produced by Cloacibacterium Normanense in Wastewater Sludge Supplemented with Crude Glycerol and Study of Extracellular Polymeric Substances Extraction by Different Methods.” *Bioresource Technology* 212 (July): 120–29. https://doi.org/10.1016/j.biortech.2016.04.021.

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.

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.

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.

Siezen, Roland J., and Marco Galardini. 2008. “Genomics of Biological Wastewater Treatment.” *Microbial Biotechnology* 1 (5): 333–40. https://doi.org/10.1111/j.1751-7915.2008.00059.x.

Stanish, Lee F., Natalie M. Hull, Charles E. Robertson, J. Kirk Harris, Mark J. Stevens, John R. Spear, and Norman R. Pace. 2016. “Factors Influencing Bacterial Diversity and Community Composition in Municipal Drinking Waters in the Ohio River Basin, USA.” *PloS One* 11 (6): e0157966. https://doi.org/10.1371/journal.pone.0157966.

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.

UniProt Consortium. 2014. “Activities at the Universal Protein Resource (UniProt).” *Nucleic Acids Research* 42 (Database issue): D191-198. https://doi.org/10.1093/nar/gkt1140.

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.

Wakelin, Steven A., Matt J. Colloff, and Rai S. Kookana. 2008. “Effect of Wastewater Treatment Plant Effluent on Microbial Function and Community Structure in the Sediment of a Freshwater Stream with Variable Seasonal Flow.” *Applied and Environmental Microbiology* 74 (9): 2659–68. https://doi.org/10.1128/AEM.02348-07.

Wang, Xiaohui, Yu Xia, Xianghua Wen, Yunfeng Yang, and Jizhong Zhou. 2014. “Microbial Community Functional Structures in Wastewater Treatment Plants as Characterized by GeoChip.” *PLOS ONE* 9 (3): e93422. https://doi.org/10.1371/journal.pone.0093422.

Weelink, Sander A. B., Nico C. G. Tan, Harm ten Broeke, Corné van den Kieboom, Wim van Doesburg, Alette A. M. Langenhoff, Jan Gerritse, Howard Junca, and Alfons J. M. Stams. 2008. "Isolation and Characterization of Alicycliphilus Denitrificans Strain BC, Which Grows on Benzene with Chlorate as the Electron Acceptor." *Applied and Environmental Microbiology* 74 (21): 6672–81. https://doi.org/10.1128/AEM.00835-08.

Wiedmann-al-Ahmad, M, H V Tichy, and G Schön. 1994. "Characterization of Acinetobacter Type Strains and Isolates Obtained from Wastewater Treatment Plants by PCR Fingerprinting." *Applied and Environmental Microbiology* 60 (11): 4066–71.

Yang, Chao, Wei Zhang, Ruihua Liu, Qiang Li, Baobin Li, Shufang Wang, Cunjiang Song, Chuanling Qiao, and Ashok Mulchandani. 2011. "Phylogenetic Diversity and Metabolic Potential of Activated Sludge Microbial Communities in Full-Scale Wastewater Treatment Plants." *Environmental Science & Technology* 45 (17): 7408–15. https://doi.org/10.1021/es2010545.

This page intentionally left blank.

3 CHICAGO AREA WATERWAY SYSTEM FECAL INDICATOR BACTERIA (CAWS-FIB) MODEL DEVELOPMENT

3.1 INTRODUCTION

The main objective of this task is to develop the CAWS-FIB model, a data-driven model for predicting fecal indicator bacteria (FIB) in the CAWS. The CAWS-FIB model model predicts 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. A number of ML algorithms including gradient boosting machine (GBM), artificial neural networks (ANNs), and XGBoosting (XGB) were explored and compared. Eventually, the algorithm that produces the best predictive ability based on the results of model performance evaluation metrics will be chosen as the main algorithm for the CAWS-FIB. The research focus to date has been to use a GBM algorithm to predict fecal coliform (Fecal) concentration or density in the water column given a set of predetermined relevant environmental variables. We developed initial modeling results using the Calumet WRP three sites (e.g., 56, 57, and 76) with the largest number of observations available among the 12 sampling sites during the pre-disinfection period (2013-2015). In addition, the limitations of the current modeling activity and future directions of this task will be discussed.

This task had the following objectives and activities:

1. Explore the applicability of a ML-based model to predict Fecal density at a location in the CAWS given a set of pre-determined relevant variables. This objective was met with two activities:

    - Data management and streamlining from multiple sources including the Metropolitan Water Reclamation District (MWRD) of Greater Chicago, Illinois State Water Survey (ISWS), U.S. Geological Survey (USGS), and NOAA - National Centers for Environmental Information (NCEI).

    - Development of a ML-based model using the GBM algorithm in Python.

2. Determine relevant site specific Fecal explanatory variables by relative importance, analyze/evaluate model training and testing performance, and demonstrate model functionality.

3.2 MATERIALS AND METHODS

3.2.1 Selected Sampling Sites Used for Model Development

The number of sampling sites where fecal coliform (Fecal) data were analyzed during the pre-disinfection period (2013-2015) was inconsistent (Table 12). In 2013, there were 12 sampling sites ranging from sites 36 to 112 (Figure 14). In 2014, a total of four sampling sites including sites 43, 52, 55, and 97 were added to the 12 sites in 2013. In 2015, site 39 were added relative to the 2014 list, but sites 43, 52, and 55 were excluded. To ensure having the largest sample size for each site available for model training and testing, sites with consistent record available across the three years (2013-2015) were selected for model development. Therefore, only the 12 sampling sites in 2013 were used in model development. In this preliminary report, only results from sites 56, 57, and 76 were shown and discussed.

3.2.2 The CAWS-FIB Conceptual Modeling Framework

The main components of the CAWS-FIB model and their connections are shown in Figure 15. 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 ML-based model (with GBM as the main algorithm in this case) that best describes the underlying complex mathematical relationship between FIB densities and environmental variables at specific points along the CAWS. The third component summarizes predicted FIB densities (Fecal 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.

TABLE 12 Summary of Sampling Sites with Fecal Data During the Pre-Disinfection (2013-2015) Period

Year	Sampling Sites
2013	36, 56, 57, 59, 73, 76, 86, 96, 99, 100, 108, 112
2014	36, 43, 52, 55, 56, 57, 59, 73, 76, 86, 96, 97, 99, 100, 108, 112
2015	36, 43, 56, 57, 59, 73, 76, 86, 96, 97, 99, 100, 108, 112

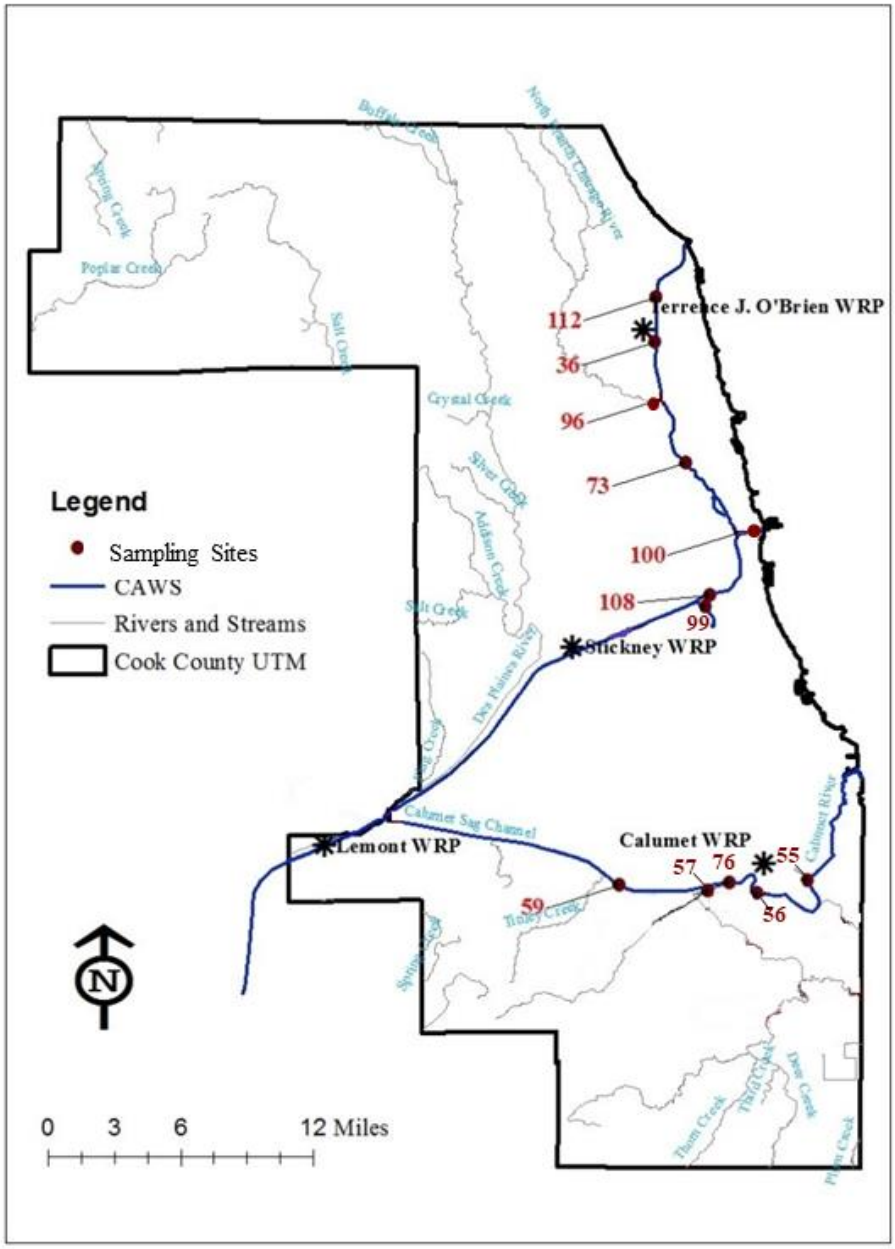


FIGURE 14 The 12 Sampling Sites to Be Used for Model Development

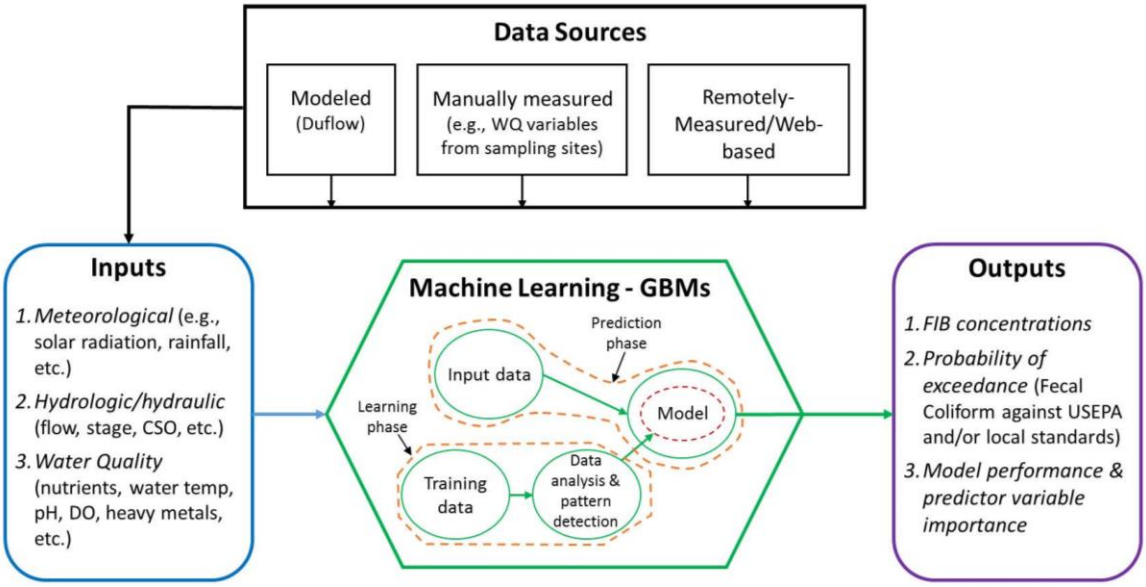


FIGURE 15 Schematic of the CAWS FIB Modeling Process

3.2.2.1 Data and Data Sources

Data for the CAWS-FIB model include daily FIB concentrations and multiple environmental variables (Figure 15) from various sources including the MWRD of Greater Chicago, ISWS, USGS, NOAA-NCEI. Environmental variables included three major categories such as 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, sediment, and heavy metals, etc.) 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., turbidity, air and 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 A.1 shows the list of the environmental variables and the corresponding summary statistics used at each of the sampling points (Sites 56, 57, and 76) 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 to 2015.

3.2.3 GBM Model Development

3.2.3.1 Overview of the GBM Algorithm

The 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). The CAWS-FIB GBM-based model was developed and implemented in Python 3.6.1 using the scikit-learn package (Pedregosa *et al.*, 2011).

3.2.3.2 Model Training and Testing

The first step in the model estimation was to subdivide the preprocessed (cleansed and enriched) 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 (Fecal) variable of the training dataset only, instead of the underlying general

structure representing the entire environmental and Fecal 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 avoid 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 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 A.1, there are a multitude of features or explanatory variables used in model development. There was a total of 183 features derived from manually and more frequently (time-lagged) measured environmental variables (Table A.1), while the number of site observations ranged from 25-38. Thus, after the 183 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) x 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. 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.

3.2.3.3 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. Due to lack of observed data outside of the training and testing dataset, a hypothetical dataset was generated consisting of the mean of the observed values for the 183 explanatory variables (MEV) for a given site from 2013 to 2015, one standard deviation of the MEV (SD1), 1.1 x MEV, 0.90 x MEV, MEV + SD1, MEV - SD1 for predictions I-VI (Table A.2).

3.2.3.4 Model Capability and Performance Metrics

In addition to predicting Fecal 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 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 this preliminary model tests, the CAWS limit for Fecal of 200 CFU/100 mL was used as the RL and DV.

In our current model the RL is fixed, while the DV can be variable. Setting the DV to some value not equal to RL may be confusing. However, when we adopt a modeling approach, we have decided to base our advisory decisions (to close a waterbody/beach from human contact or not) on a statistical model derived from historical data. For instance, a model prediction of 150 CFUs/100 mL may be approximately equivalent to an actual or "real" observation of 175 CFUs/100 mL, or a "real" value of 125 CFUs/100 mL, depending on the specific model. The regulatory limit (RL) is on the scale of actual observations, while the DV is on the scale of the model predictions. Therefore, we should not think of model predictions as actual FIB concentrations, but only some quantity that is related to actual FIB concentrations.

When we raise or lower the DV, we are inherently adjusting for the differences in scale between model predictions and the actual observations. We understand that some difference exists between the scale of mode predictions and actual concentrations, and we are factoring that difference in the decision making process. We use the plot of model fits vs. actual observations to help us choose the DV value that "optimizes" model performance (striking a balance between sensitivity, specificity, and/or overall model accuracy.

A user is free to decrease or increase the DV to see what DV optimizes the model's sensitivity and specificity. High specificity means you don't often close the water body for human contact/swimming unnecessarily, while high sensitivity means you rarely expose swimmers to high FIB concentrations. Most beach managers or advisory decision-makers emphasize safety and therefore value sensitivity over specificity. By lowering the DV, we may incur on additional false positive, but if we lose one false negative, that may be worthwhile. However, if we incur multiple false positives to delete one false negative by lowering the DV, that may not be palatable. Some people pick the DV that maximizes the overall accuracy of the model. In the end, the DV is a cutoff that says "if the model prediction is above this number, we will issue an advisory to close the water body/beach from human contact; if the model prediction is below this number we won't." It is completely up to the advisory decision maker, while the RL is set by law or proclamation and should not be adjusted.

Model performance metrics include 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 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 16). 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 concentrations, while high specificity corresponds to minimizing the chances of putting up a water quality advisory when the actual Fecal density is below the threshold. 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.

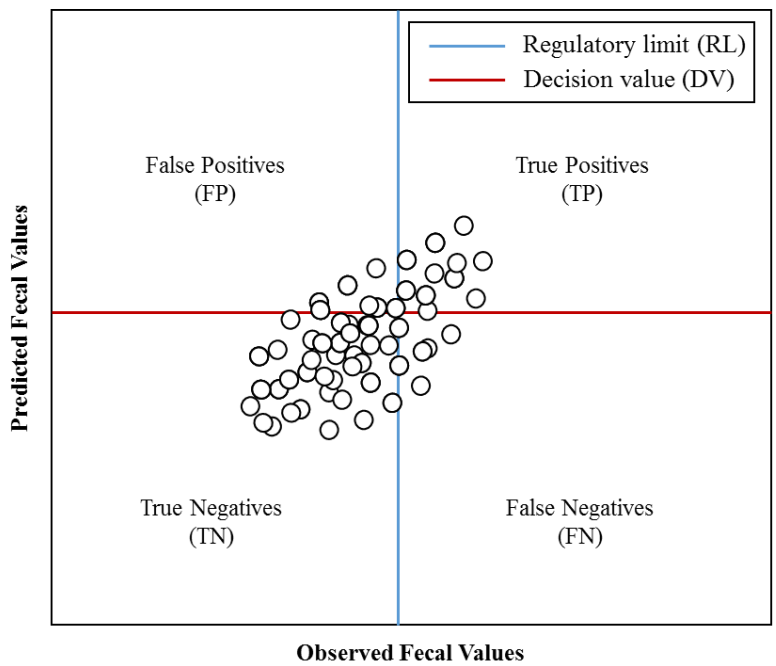


FIGURE 16 A Schematic of a Scatter Plot of Observed and Predicted Fecal Values During Model Training and Testing Showing False Positives (FP), True Positives (TP), True Negatives (TN), and False Negatives (FN)

3.3 RESULTS

3.3.1 Model Training and Testing

A plot of a long list of explanatory variables for Fecal ranked from the most to the least important is shown in FIGURE A.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 the GBM-based model to find variable importance was necessary. Initially, the GBM-based model was run 30 times over the 30 differing random configurations of the test and training data use the median of the variable importance was used to determine the most relevant explanatory features (Figure A.1). The final model development was then conducted by running the GBM-based model on the 15, 10, and 5 most relevant explanatory variables for each site. Figure 17 shows the top 15 most relevant explanatory variables for sites 56, 57, and 76. The top 15 most explanatory variables across the three sites are comprised of both the one-time, manually collected during the sampling period and more frequently measured or time-lagged environmental factors. The manually measured relevant explanatory variables include pH and dissolved oxygen (DOMan), which rank as top variables for two of the sites, as well as nitrate (NO_3) and water temperature (TwMan). Net radiation (RnSDhrs), air temperature (TaSDhrs or TaMeanhrs), and dissolved oxygen (e.g., DOSDhrs or DOMeanhrs) are the most common time-lagged explanatory variables followed by, discharge (LQ10Maxhrs or LQ10Minhrs), water temperature (TwSDhrs), specific conductance (SpCondSDhrs or SpCondMeanhrs), rainfall (RsqrSumhrs), and stage (HDiffhrs).

Figure 18 shows the deviance plots for sites 56, 57, and 76 for models that have 15 features, 10 features, and 5 features, respectively. The training error or deviation between observation and prediction decreases as the number of boosting iterations or regression trees increases. The optimal number of boosting iterations or number of regression trees where convergence is met ranges from 1000 to 10000. This is true across sites and regardless of whether the model has 15, 10, or 5 features or variables. Conversely, the testing error is generally unchanged by the number of boosting iterations across the three sites, and in some cases, slightly increased as the number of boosting iterations is increased. The aforementioned is a manifestation of overfitting, which has been described in section 2. This observation is also supported by the values of numerical model performance metrics including R^2, MSE, and RMSE, which consistently show excellent and very poor values during training and testing, respectively, across sites and models (Tables 13 to 15). Similarly, the models perform well during training in terms of classifying whether a predicted Fecal value is a TP, TN, FP, and FN, consistently showing an accuracy of 1.0 across the three sites regardless of the number of relevant variables. Model accuracy during testing varied across sites and number of variables (Tables 13 to 15). The 10-variable model consistently performed well during testing with classification accuracy of 0.85, 0.75, and 0.875 for sites 56, 57, and 76, respectively. In contrast, a 5-variable model consistently showed the lowest accuracy values of 0.714, 0.375, and 0.375 for sites 56, 57, and 76.

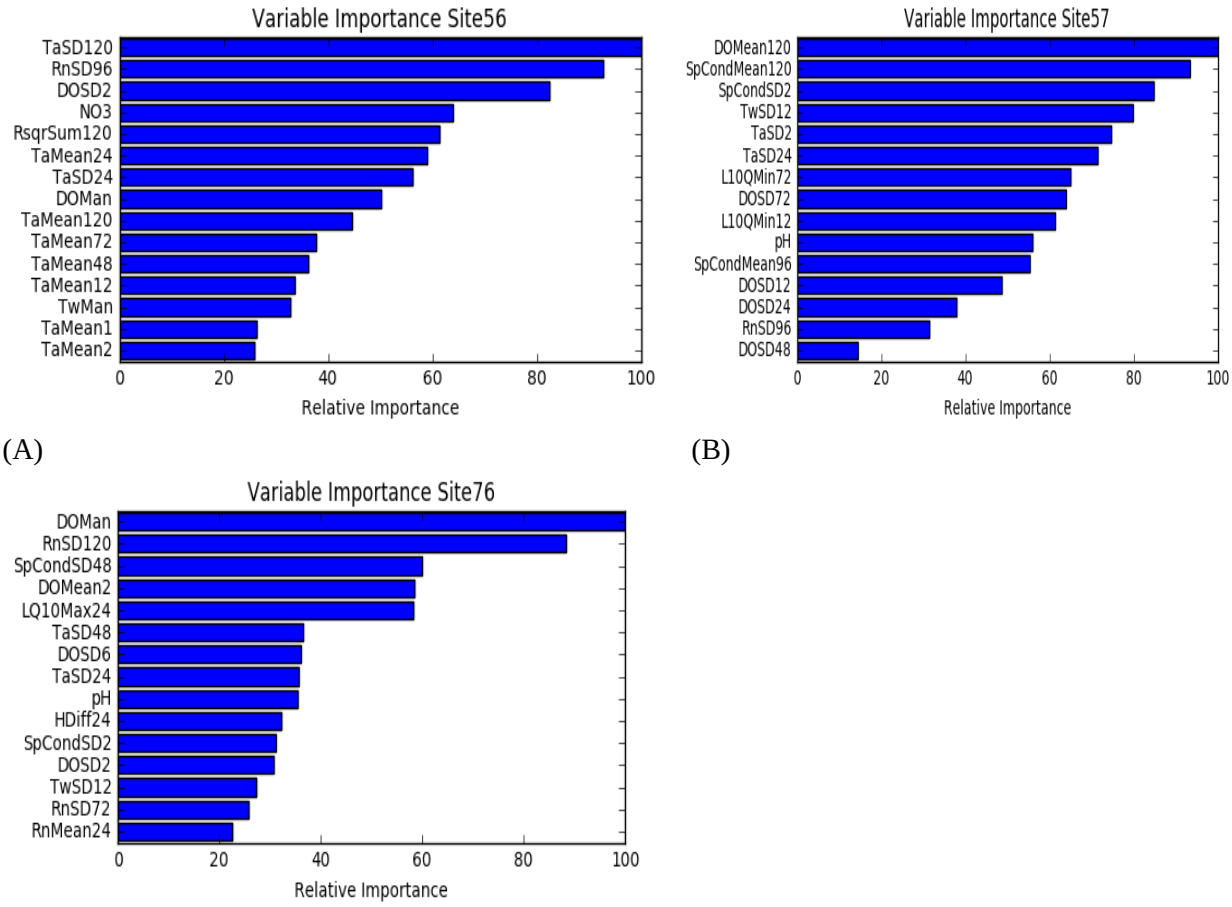
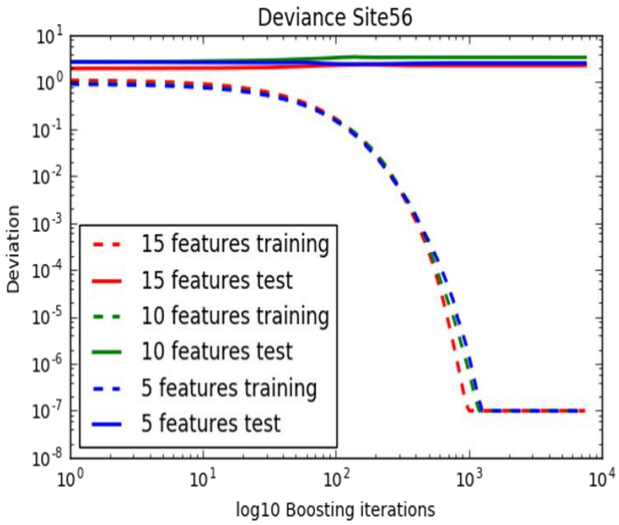
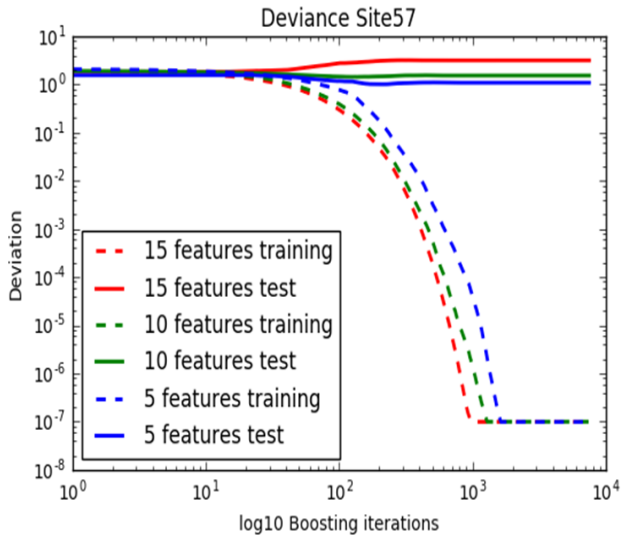


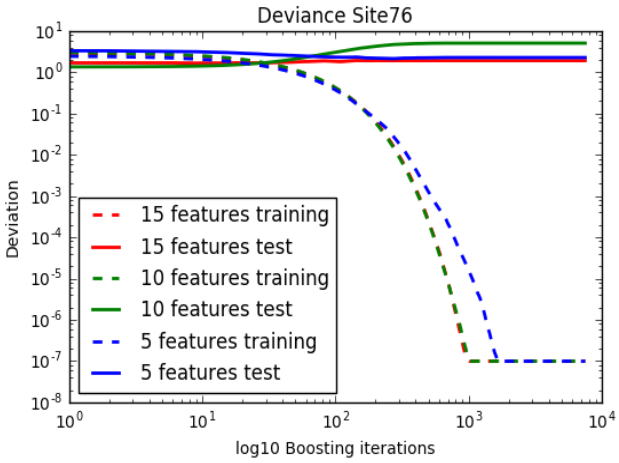
FIGURE 17 Plots of the 15 Most Relevant Explanatory Variables for (A) Site 56, (B) Site 57, and (C) Site 76. Explanatory variables, in general, are named by an abbreviation for the environmental variable, followed by the transformation used (if any, i.e., mean and standard deviation (SD)), and the time-lag (1-120 hours). For instance, standard deviation of net 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 $LQ_{10}Max_{24}$.



(A)



(B)



(C)

FIGURE 18 Plots of Deviance for (A) Site 56, (B) Site 57, and (C) Site 76 between Training (broken line) and Testing (solid line) of Models that Include 15 Features (red line), 10 Features (green line), and 5 Features (blue line)

TABLE 13 Values of the Model Training and Testing Performance Metrics for Site 56

Phase	Model	R^2	MSE	RMSE	TP	TN	FP	FN	Accuracy	Specificity	Sensitivity
Training	15-Var.	1.0	0.00	0.00	6	22	0	0	1.0	0.78	1.0
	10-Var.	1.0	0.00	0.00	6	22	0	0	1.0	0.82	1.0
	5-Var.	1.0	0.00	0.00	5	23	0	0	1.0	0.82	1.0
Testing	15-Var.	-0.96	223.8	14.8	0	6	0	1	0.85	1.0	0.10
	10-Var.	-0.08	6.3	2.51	0	6	0	1	0.85	1.0	0.0
	5-Var.	0.08	3.23	1.80	1	4	1	1	0.714	0.8	0.5

Var. = variables; MSE = mean squared error; RMSE = root mean squared error; TP = true positives; TN = true negatives; FP = false positives; FN = false negatives.

TABLE 14 Values of the Model Training and Testing Performance Metrics for Site 57

Phase	Model	R2	MSE	RMSE	TP	TN	FP	FN	Accuracy	Specificity	Sensitivity
Training	15-Var.	1.0	0.00	0.00	6	22	0	0	1.0	0.79	1.0
	10-Var.	1.0	0.00	0.00	6	22	0	0	1.0	1.0	1.0
	5-Var.	1.0	0.00	0.00	5	23	0	0	1.0	0.82	1.0
Testing	15-Var.	-0.38	645.55	25.4	3	2	1	0	0.625	0.4	0.6
	10-Var.	-0.19	19952	141.25	5	1	2	0	0.75	0.17	1.0
	5-Var.	-0.68	104712	323.59	1	2	1	4	0.375	0.67	0.2

Var. = variables; MSE = mean squared error; RMSE = root mean squared error; TP = true positives; TN = true negatives; FP = false positives; FN = false negatives.

TABLE 15 Values of the Model Training and Testing Performance Metrics for Site 76

Phase	Model	R^2	MSE	RMSE	TP	TN	FP	FN	Accuracy	Specificity	Sensitivity
Training	15-Var.	1.0	0.00	0.00	22	7	0	0	1.0	0.24	1.0
	10-Var.	1.0	0.00	0.00	20	9	0	0	1.0	1.0	1.0
	5-Var.	1.0	0.00	0.00	21	8	0	0	1.0	0.28	1.0
Testing	15-Var.	0.27	116380290	10787	5	0	3	0	0.625	0.00	1.0
	10-Var.	-0.24	910690737	30177	7	0	1	0	0.875	0.00	1.0
	5-Var.	-2.53	205810776	14346	6	1	1	0	0.375	0.14	1.0

Var. = variables; MSE = mean squared error; RMSE = root mean squared error; TP = true positives; TN = true negatives; FP = false positives; FN = false negatives.

3.3.2 Model Prediction

Predicted Fecal concentrations using the hypothetically generated dataset (Table A.2) and probability of exceedance (POE) values based on the RL and DV of 200 CFUs/100 mL are shown in Table 16, Table 17, and Table 18 for sites 56, 57, and 76, respectively. Predicted Fecal concentrations ranged from 16 to 2523 CFUs/100 mL for site 56, 182 to 1626 CFUs/100 mL for site 57, and 452 to 89657 CFUs/100 mL for site 76. These values are within the range of the observed Fecal densities during 2013-2015 with the lowest value of <1 CFU/100 mL for all three sites and highest values of 2000, 20,000, and 100,000 CFUs/100 mL for sites 56, 57, and 76, respectively. The limited amount of data for model training and testing as well as the large dynamic range of the data seem to limit the assignment of a POE to a predicted Fecal value to 0 and 100% only (Tables 16 to 18). For example, a predicted Fecal value less than RL and DV is assigned a 0% POE regardless whether it is very close to 200 CFUs/100 mL (e.g. 198 CFUs/100 mL) or well below 200 CFUs/100 mL (e.g., 16 CFUs/100 mL).

TABLE 16 Model Predicted Fecal Coliform (Fecal) 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

Prediction	Input[a]	SITE CODE	RL [CFUs/ 100 mL]	DV [CFUs/ 100 mL]	Predicted Fecal [CFUs/100 mL]	POE [%]
I	MEV	Site56	200	200	29	0
II	SD1	Site56	200	200	2,523	100
III	1.1 x MEV	Site56	200	200	16	0
IV	0.9 x MEV	Site56	200	200	41	0
V	MEV + SD1	Site56	200	200	707	100
VI	MEV - SD1	Site56	200	200	37	0

[a] MEV=mean of the observed values for the 183 explanatory variables from 2013 to 2015; SDI = one standard deviation of the MEV.

TABLE 17 Model Predicted Fecal Coliform (Fecal) Density (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

Prediction	Input[a]	SITE CODE	RL [CFUs/ 100 mL]	DV [CFUs/ 100 mL]	Predicted Fecal [CFUs/100 mL]	POE [%]
I	MEV	Site57	200	200	182	0
II	SD1	Site57	200	200	304	100
III	1.1 x MEV	Site57	200	200	185	0
IV	0.9 x MEV	Site57	200	200	198	0
V	MEV + SD1	Site57	200	200	1,554	100
VI	MEV - SD1	Site57	200	200	1,626	100

[a] MEV=mean of the observed values for the 183 explanatory variables from 2013 to 2015; SDI = one standard deviation of the MEV.

TABLE 18 Model Predicted Fecal Coliform (Fecal) Density (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 76

Prediction	Input[a]	SITE CODE	RL [CFUs/ 100 mL]	DV [CFUs/ 100 mL]	Predicted Fecal [CFUs/100 mL]	POE [%]
I	MEV	Site76	200	200	1,547	100
II	SD1	Site76	200	200	89,657	100
III	1.1 x MEV	Site76	200	200	2,173	100
IV	0.9 x MEV	Site76	200	200	452	100
V	MEV + SD1	Site76	200	200	3,204	100
VI	MEV - SD1	Site76	200	200	5,819	100

[a] MEV=mean of the observed values for the 183 explanatory variables from 2013 to 2015; SDI = one standard deviation of the MEV.

3.4 LIMITATIONS OF THE CURRENT MODELING WORK AND FUTURE WORK

The CAWS hydraulic model construction and validation for 2014 and 2015 using the DuFlow model is still underway. Therefore, discharge (Q) and stage (H) as input data in the current GBM-based model were taken from the nearest USGS Gage stations instead of using DuFlow-simulated Q and H at or near each sampling site since USGS Gage no. 05536357 at Grand Calumet River was used for sites 56 and 76, while USGS Gage no. 05536340 at Midlothian Creek was used for site 57.

Model training and testing indicated that overfitting was a problem for the CAWS-FIB model, as the model performed well when classifying whether a predicted Fecal value is a TP, TN, FP, and FN, but produced lower accuracy values during testing. However, the 10-variable model consistently performed well with classification accuracy over 0.7 across sites. The Fecal concentrations predicted using the hypothetically generated dataset were within the range of the Fecal densities observed during 2013-2015, suggesting the model produced reasonable Fecal estimates for the CAWS. However, the model produced a binary (0 or 100%) POE assignment to the predicted Fecal values, likely due to the limitations of the dataset.

Overall, the model training and testing results should be considered preliminary, and may improve after the nine additional sites are incorporated into the model training and testing. The current GBM-based model will be applied to the remaining sampling sites (sites 36, 59, 73, 86, 96, 99, 100, 108, 112) to evaluate how the model performs and whether the list of explanatory variables change across the 12 sampling sites. Additional approaches such as artificial neural networks (ANNs) and XGBoosting (XGB) will also be explored. Eventually, the algorithm that consistently shows the best performance will be used as the main algorithm for the CAWS-FIB model. Additionally, input regularization or standardization (mean of 0) will be applied across the 183 explanatory variables, as an alternative to mixed-scale (original and log-transformed) inputs suggested by previous studies. This proposed action is warranted due to the wide range of values (<0 to >1000000) of the explanatory variables. Following the development of the

predictive model for fecal coliform, we will attempt to apply the model for predicting other fecal indicator bacteria (bacteriodes, Alpha diversity, etc.). This will only be possible for the post-disinfection years (2016-2019) due to the limited amount of pre-disinfection data available for the aforementioned FIBs in training the model.

3.5 REFERENCES

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

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.

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.

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.

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

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

APPENDIX A

APPENDIX A

TABLE A.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 in mm									
Sum		✓	✓	✓	✓	✓	✓	✓	✓	✓
Q	Discharge, m^3 s^{-1}									
Difference				✓						
$\log_{10}Q$	Logarithm of discharge measured in m^3 s^{-1}									
Mean		✓	✓	✓	✓	✓	✓	✓	✓	✓
Min		✓	✓	✓	✓	✓	✓	✓	✓	✓
Max						✓				
H	Stage, m									
Mean		✓	✓	✓	✓	✓	✓	✓	✓	✓
Difference				✓		✓				
CSO_int	Intensity of the combined sewer overflows (CSOs), gph									
Mean		✓	✓	✓	✓	✓	✓	✓	✓	✓
Difference			✓	✓		✓				
log10 CSO_int	Logarithm of the intensity of CSOs measured in gph									
Mean		✓	✓	✓	✓	✓	✓	✓	✓	✓
Min		✓	✓	✓	✓	✓	✓	✓	✓	✓
Max						✓				
CSO_mag	Magnitude of CSO, gal									
Sum		✓	✓	✓	✓	✓	✓	✓	✓	✓
Mean		✓	✓	✓	✓	✓	✓	✓	✓	✓
Std. Dev.			✓	✓	✓	✓	✓	✓	✓	✓
log10 CSO_mag	Logarithm of the magnitude of CSO measured in gal									
Mean		✓	✓	✓	✓	✓	✓	✓	✓	✓
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								

TABLE A.1 (Cont.)

Variable and Statistics	Description and Units	Time Windows/Lags (hrs) 1	2	6	12	24	48	72	96	120
TOC	Total organic carbon, mg L^{-1}									
DO	Dissolved oxygen, mg L^{-1}									
TDS	Total dissolved solids, mg L^{-1}	Manual								
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 A.2 General Description of the Hypothetical Dataset Used in Model Prediction to Showcase Model Predictive Functionality and Other Computations

Prediction	Input
I	MEV
II	SD1
III	1.1 x MEV
IV	0.9 x MEV
V	MEV + SD1
VI	MEV - SD1

MEV = mean of the observed values for the 183 explanatory variables for a given site from 2013-2015; one standard deviation of the mean (SD1).

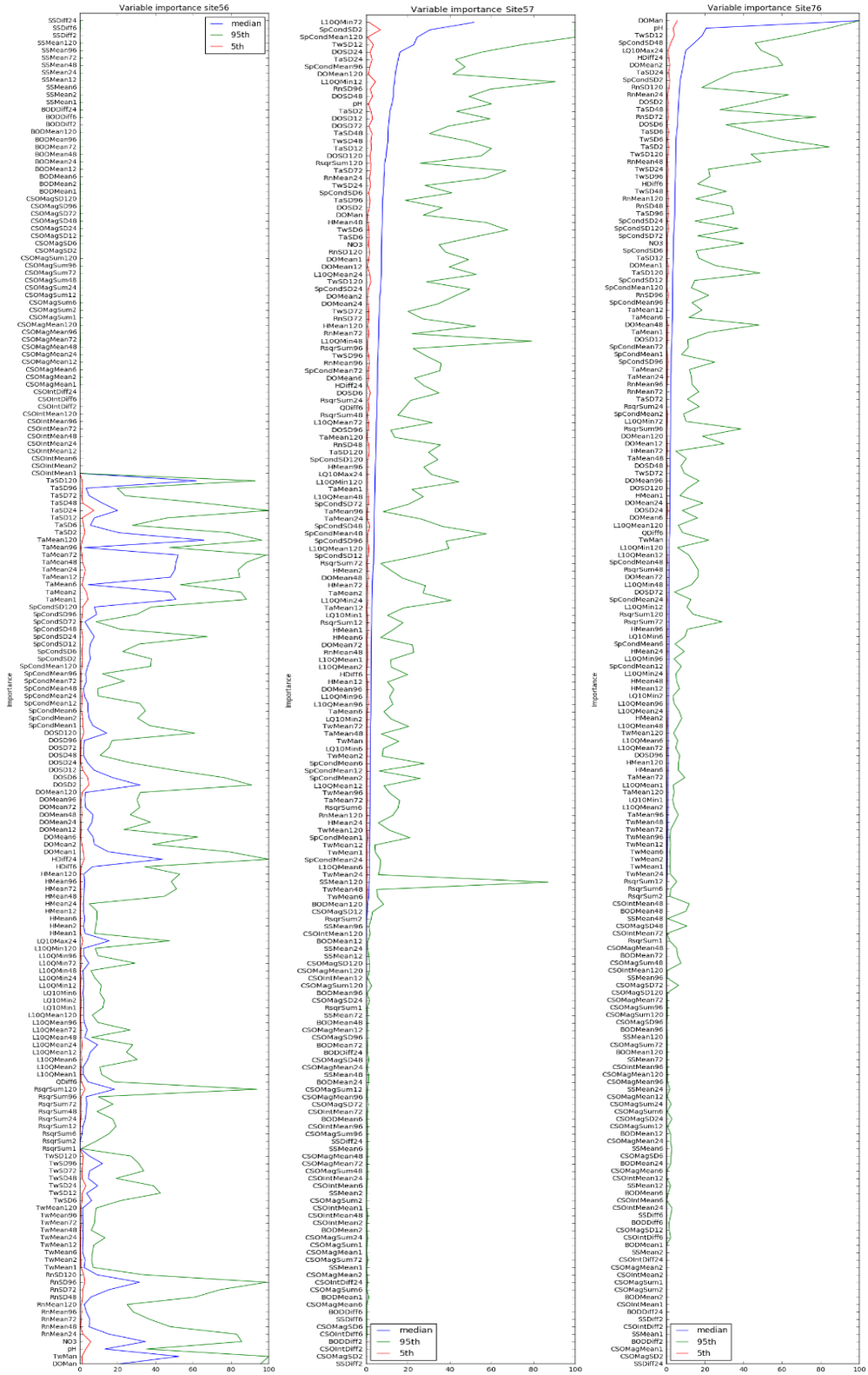


FIGURE A.1 A Sample Plot of the Long List of Explanatory Variables for Fecal from the Most to the Least Relevant Variable Used for the Dimensionality Reduction Step for Each Site

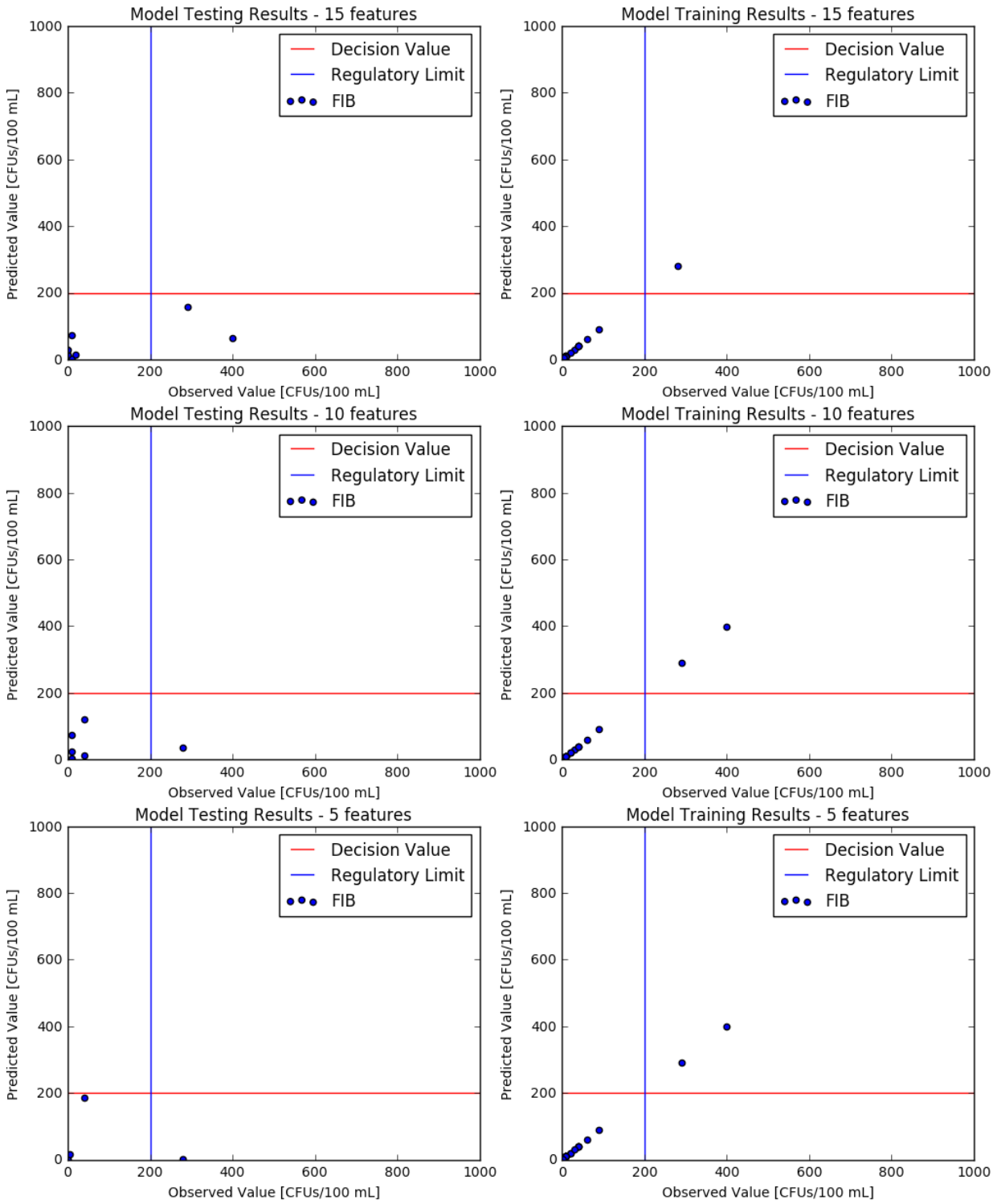


FIGURE A.2 Predicted vs. Observed Plot of Fecal Values for Site 56 with Regulatory Limit (RL, vertical blue line) of 200 CFU/100 mL and Decision Value (DV, horizontal red line) of 200 CFU/100 mL in Both Model Training and Testing and Using 15-, 10- and 5-Most Relevant Explanatory Variables

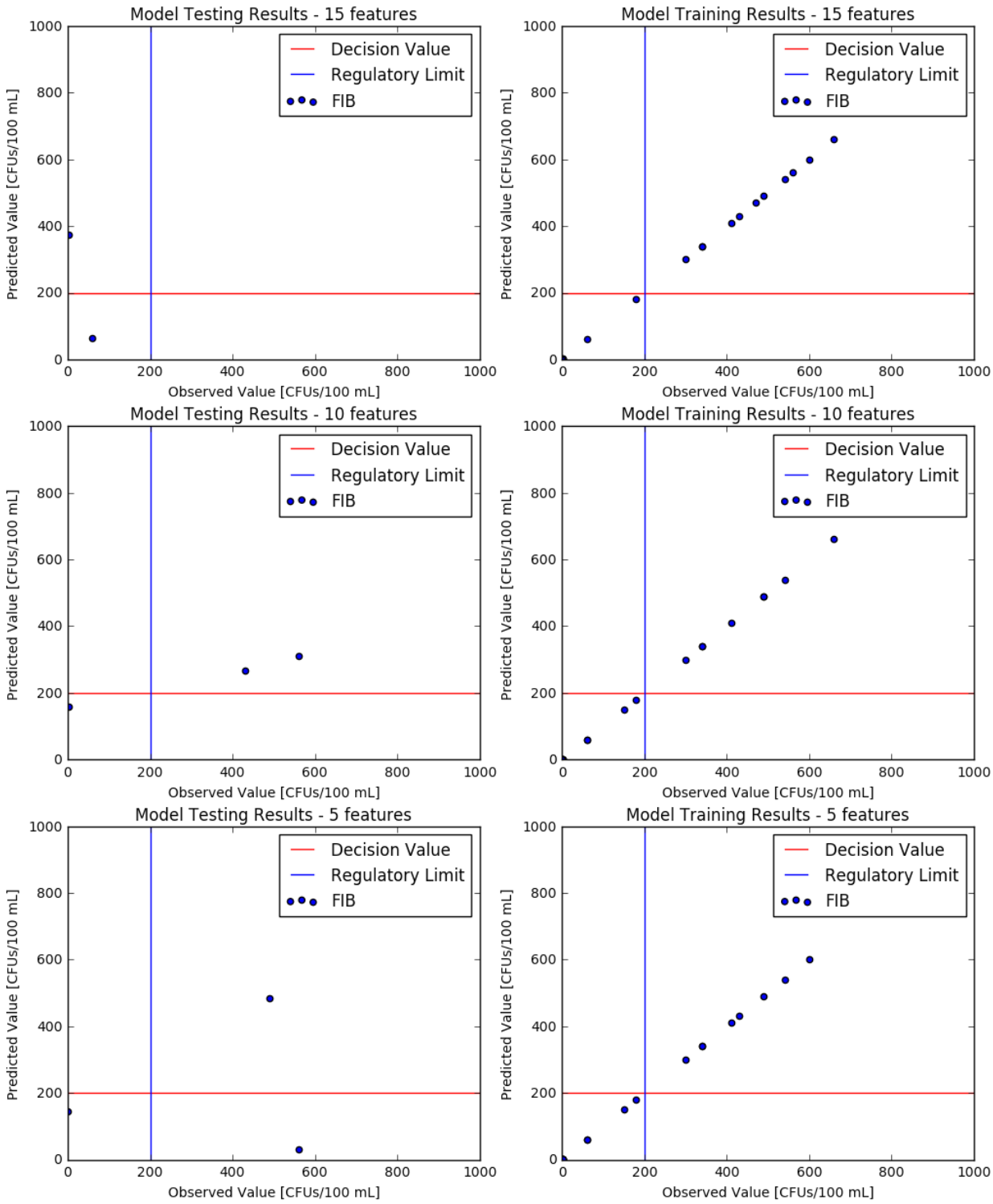


FIGURE A.3 Predicted vs. Observed Plot of Fecal Values for Site 57 with Regulatory Limit (RL, vertical blue line) of 200 CFU/100 mL and Decision Value (DV, horizontal red line) of 200 CFU/100 mL in Both Model Training and Testing and Using 15-, 10- and 5-Most Relevant Explanatory Variables

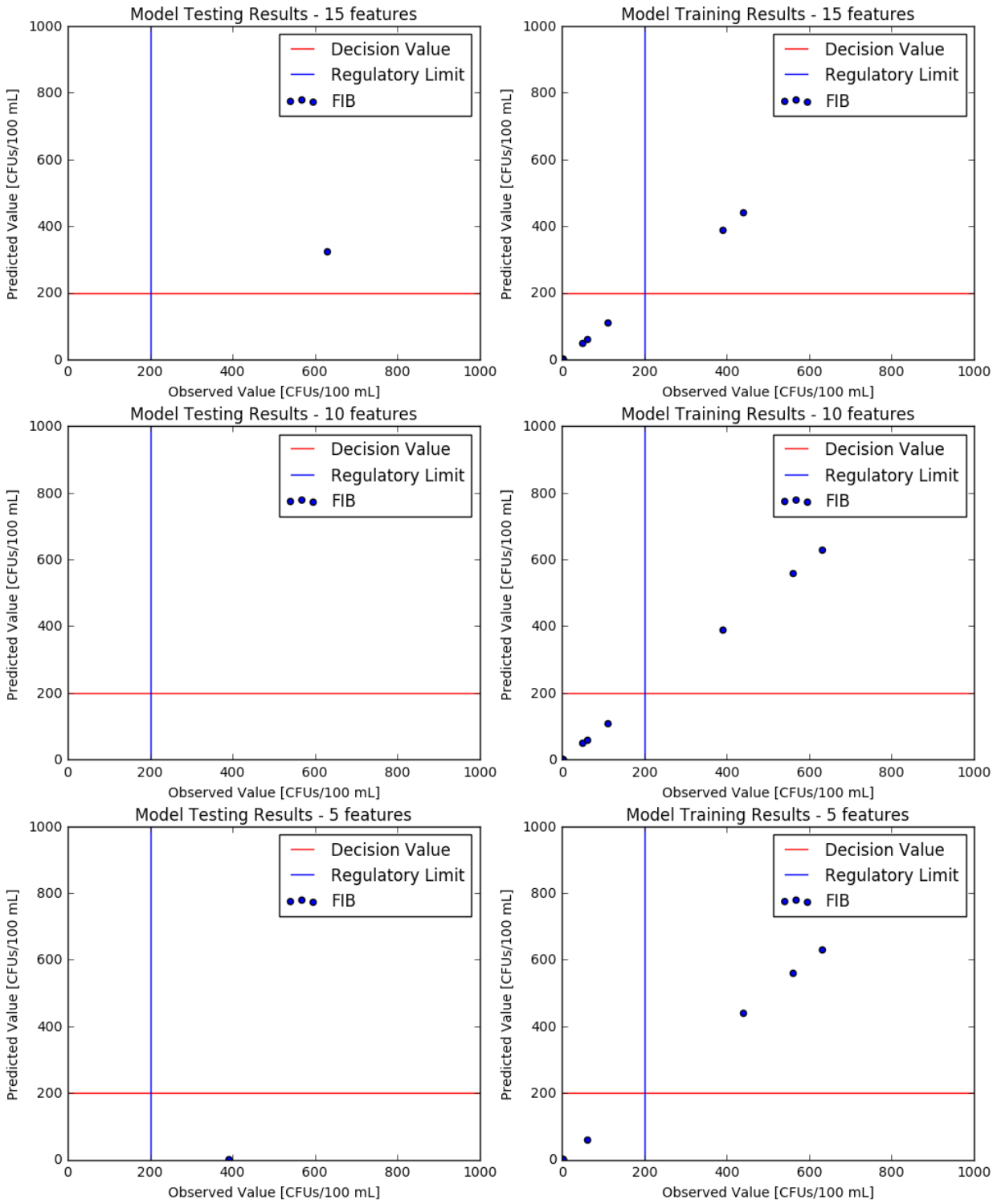


FIGURE A.4 Predicted vs. Observed Plot of Fecal Values for Site 76 with Regulatory Limit (RL, vertical blue line) of 200 CFU/100 mL and Decision Value (DV, horizontal red line) of 200 CFU/100 mL in Both Model Training and Testing and Using 15-, 10- and 5-Most Relevant Explanatory Variables

Environmental Science Division
Argonne National Laboratory
9700 South Cass Avenue, Bldg. 240
Argonne, IL 60439-4854

www.anl.gov

Argonne National Laboratory is a U.S. Department of Energy
laboratory managed by UChicago Argonne, LLC