Argonne NATIONAL LABORATORY

DRAFT REPORT
FOR SPONSOR REVIEW ONLY

Phase I Interim Report for Chicago Area Waterways System Microbiome Research December 2013–December 2015

Biosciences Division
Energy Systems Division

DRAFT REPORT
FOR SPONSOR REVIEW ONLY

Phase I Interim Report for Chicago Area Waterways System Microbiome Research December 2013–December 2015

Principal Investigators
M. Cristina Negri and Jack A. Gilbert

Research Team
Melissa Dsouza, Herbert Ssegane, Jarrad Hampton-Marcell,
Naseer Sangwan, and Patty Campbell

Prepared for: Metropolitan Water Reclamation District of Greater Chicago

July 18, 2016

CONTENTS

CONTENTS (CONT.)

FIGURES

FIGURES (CONT.)

TABLES

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 3 years of a 7-year study to investigate the typical sources and distribution of microbial communities in the Chicago Area Waterway System (CAWS) and the response of the CAWS microbial ecology to the disinfection mangement process to be employeed during this term.

Microbial communities are key players in maintaining the CAWS health. Traditional laboratory culture methods such as fecal bacteria counts have been extensively used to characterize the CAWS microbial quality for regulatory purposes; however, these methods cannot resolve the source of the contamination. This study, which started in 2013, aims to understand the composition and sources of the CAWS microbial population using state-of-the-art amplicon and metagenomic science. Metagenomics is the study of whole microbial communities in a given sample by analyzing all the genetic material in that sample instead of analyzing select marker genes. This method enables us to identify common sources of microbial organisms (bacteria, archaea, and viruses) and assemblages, as well as their function in the Chicago River system.

There are many potential sources of the microorganisms in the CAWS, including treated effluent from wastewater treatment plants, land-based stormwater runoff, combined sewer overflows, sediment resuspension, storm drains, and direct input from animals. The study aims to provide a detailed investigation of the potential sources of microbes in the CAWS, especially fecal-indicator organisms. The investigation also aims to determine the biogeography of the microbial communities in the CAWS, whether free living or dependant on particular hosts such as humans, pets, or birds. It also aims to determine whether rainfall, temperature, water chemistry, and other factors are associated with the structure and composition of these microbial assemblages. In the long term, these data will provide another layer of information to enable good stewardship and management of this important water resource, as well as to gain insights into how to improve water quality for primary contact recreation uses.

Once completed, this study will document potential changes in microbial communities as the MWRD begins disinfecting its secondary-treated effluents at the O'Brien and Calumet Water Reclamation Plants in 2016 and as phases of the Tunnel and Reservoir Plan (TARP)—the Thornton Composite Reservoir and the first phase of the McCook Reservoir—are completed in 2015 and 2017, respectively. The results from river water and sediment samples taken in the first 3 years (2013–2015), discussed in this report, will serve as a baseline for future years' data; they will be compared to additional samples that will be collected each year until 2019 as the MWRD takes steps to improve the CAWS water quality.

To date, we have analyzed and processed via 16S rRNA amplicon analysis 196 blank (equipment, filter) samples, 24 fish gut samples, 24 fish mucus samples, 261 sediment samples, and 429 water column samples from 17 sites in the Chicago River and related manmade waterways, together with 28 influent sewage, 10 mixed liquor, and 190 secondary-treated final effluent samples from two Water Reclamation Plants (WRPs, O'Brien and Calumet) sampled

during 2013, 2014, and 2015. We have also analyzed via shotgun metagenomics 54 samples from the 2013 and 2014 collection.

The study found differences in the microbial community structure based on the type of sample analyzed and possibly local environmental characteristics. Here we present the results of this baseline study period, demonstrating differential biogeographic patterns, source apportionment, and temporal structure for the microbial assemblages at each sampling site. We report the results of the microbial data as a function of wet/dry events across the three sampling years, and the effects of land-use type on CAWS water- and sediment-associated physiochemical properties.

Combining intensive water quality monitoring, land use and land cover (LULC) analysis at sampling sites, and bacterial source tracking provides useful information on major sources that could be contributing bacteria to the CAWS. We analyzed the distribution of the LULC to assist in characterizing the samples from each site. We determined similarities between sampling sites using the water chemistry data, microbial indicators, and LULC to describe the influence of effluent from WRPs, the distribution of LULC, and discharge from Lake Michigan on the water quality along the CAWS. The water quality sampling sites are not the same as the locations where flow and stage are monitored along the CAWS. Therefore, to extract flow, velocity, and stage data at each water quality sampling site, hydraulic modeling was implemented using the DuFlow model constructed for this project using the 2013 data. In the future, the model will be used to develop a computational interface that will allow us to integrate the microbial data with the hydraulic data; the objective which is to develop, in the future, better analytical and predictive capabilities for different sources of bacteria, in addition to spatial and temporal distribution of microbial hotspots. Data from subsequent years will be added to the model as data becomes available and as the interface is developed.

Results from this baseline analytical period can be summarized as follows:

- The microbial communities in the CAWS were significantly different based on sampling location and sampling medium; however, they were stable between years and monthly sampling events.

- Effluent-associated microorganisms, including human fecal indicators, could be tracked downstream of the secondary-treated effluent, and were typically more abundant close to its discharge location.

- Sequencing of some of the *E. coli* culture plates provided by the MWRD show that the cultured bacteria represented a number of different *E. coli* strains, with a small number of genes present that were related to virulence, disease, and defense functions.

- Land use had a significant effect on water- and sediment-associated physiochemical properties and microbial communities.

- Differences between microbial communities could be attributed to different sources depending on sampling site, but showed similar apportionment between years.

- No significant difference in microbial community structure was observed between wet and dry sampling events.

- Metagenomic analysis trends were similar to amplicon sequence trends. Mapping genes against *E. coli* supported amplicon evidence for a low abundance of this species, including a very low abundance of *E. coli* associated virulence markers.

Proposed work for the upcoming Phase 2 includes (1) repeating the ambient water quality monitoring schedule to observe whether disinfection causes recordable changes in microbial communities, (2) method development studies to determine with more precision the analytical sensitivity of the method in the presence of a large microbial diversity, and (3) the coupling of the current methodologies with methods such as qPCR to attain a better understanding of the quantitative aspects of this research. In 2016, we also plan to develop the modeling interface to link microbial community indicators to hydraulic parameters.

SUMMARY OF PRACTICAL IMPLICATIONS

This report summarizes the research conducted by Argonne National Laboratory (Argonne) on behalf of the Metropolitan Water Reclamation District of Greater Chicago (MWRD) for the first 3 years of a 7-year study to investigate the typical sources and distribution of microbial communities in the Chicago Area Waterway System (CAWS).

Microbial communities are key players in maintaining the health of the CAWS. 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. This study, which started in December 2013, aims to better understand the composition and sources of the microbial communities associated with the CAWS using 16S rRNA gene amplicon- and metagenome-based sequencing. For nearly two decades, 16S rRNA gene sequencing has been used for the accurate and reliable qualitative identification and classification of microorganisms. It remains a powerful technique for investigating microbial relationships and diversity; however, it cannot currently provide quantitative measurements of microorganisms. A metagenome represents the entire genetic material in a given sample. At the right sequencing depth, metagenomics-based sequencing can capture all genes present in all microbes. As a result, this scientific method gives insight into the functional potential of microbes present in that sample. Together with 16S rRNA gene sequencing, molecular methods such as metagenome sequencing can supersede, for qualitative analyses, typical culture-based methods that currently only detect approximately 8% of known microbes. 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., antimicrobial resistance), and they can be used to predict with greater accuracy the common sources of microbes in these waters. These methods can also help us discover which microbes are present in the CAWS, and what these microbes are capable of doing in the CAWS. As noted, these methods are not quantitative and therefore cannot by themselves help us determine how many microbes are present for a specific genus, species, or phylum of bacteria.

This study aims to provide an understanding of several key questions: What are the sources of the CAWS' microbial communities? Are they from specific sources? Are they widespread, or are they constrained to particular sections of the CAWS? Are they free-living microbes, or dependent on particular hosts such as humans, pets, and birds? The study also aims to determine whether environmental parameters such as characteristics of flow, rainfall, temperature, or water chemistry are associated with microbe presence. Potential sources include effluent from water reclamation plants (WRPs), direct stormwater runoff, and combined sewer overflows (CSOs). CSO events occur when stormwater runoff exceeds the capacity of the sewer network, resulting in the discharge of untreated wastewater and storm water into surface waters. CSOs are recognized as significant additional sources for micro-pollutants in surface waters. This study uses synoptic sampling at predetermined locations to collect water and sediment samples for microbial and metagenomics analysis based on set wet or dry weather conditions.

This project will provide information on how to dynamically manage this important water resource for primary contact recreation uses.

Once completed, this study will document potential changes in the microbial communities as the MWRD begins disinfecting its secondary-treated effluents at the O'Brien and Calumet WRPs in 2016 and as phases of the Tunnel and Reservoir Plan (TARP)—the Thornton Composite Reservoir and the first phase of the McCook Reservoir—are completed in 2015 and 2017, respectively. The results from river water and sediment samples taken in the first 3 years (2013–2015), which are discussed in this report, will serve as a baseline for future years' data. Additional samples will be collected each year until 2019 as the MWRD takes steps to further improve the CAWS water quality.

To date, we have processed and analyzed 196 blank (equipment, filter) samples, 24 fish gut samples, 24 fish mucus samples, 278 sediment samples, and 429 water column samples from 17 sites in the CAWS and related man-made waterways, together with 22 influent sewage, 10 mixed liquor, and 190 secondary-treated final effluent samples from two WRPs (O'Brien and Calumet) sampled during 2013, 2014, and 2015. We also analyzed samples of gull and dog droppings to identify microbes that are unique to these sources. These samples allowed us to differentiate the microbial community structure based on local environmental characteristics.

METHODS

For this study, we rely on two different sequencing methods to analyze the samples received from the MWRD. First, we isolate and analyze the genomic DNA of the samples using 16S rRNA gene-based amplicon sequencing. In this analysis, the 16S rRNA genes of bacteria and archaea in these samples are amplified and sequenced using high-throughput sequencing (Illumina Miseq), which provides information about the particular bacterial and archaeal species present (i.e., "Who is there?"). Second, selected samples are also analyzed through shotgun metagenome sequencing, which provides us with valuable information on the gene functions associated with these microorganisms (i.e., "What are they capable of doing?"). Statistical analysis is then utilized to correlate the presence of specific microbial communities with potential sources and with other characteristics of the samples collected, such as chemical parameters or land use in the vicinity of the collection site.

In this report, we present the results of the baseline study period and the statistical analysis of the data generated, demonstrating differential biogeographic patterns, source apportionment, and temporal structure for the microbial assemblages at each sampling site.

A second part of this report presents the results of a parallel hydraulic model development effort, in which we sought to provide accurate model results of flow and other hydraulic parameters. These model outputs will allow us, in the future, to develop a computational interface to integrate the hydraulic data with the microbial community data so that we will be able to predict microbial dynamics as a function of river hydraulics and water quality.

RESULTS

Approximately 16.8 million high-quality 16S rRNA amplicons were generated from 891 CAWS samples collected from 2013 through 2015. This analysis allowed us to determine which were the most common phyla (taxonomic grouping of microorganisms) present in our samples. Acidobacteria, Actinobacteria, Bacteroidetes, Chlorobi, Chloroflexi, Cyanobacteria, Firmicutes, Planctomycetes, Proteobacteria, and Verrucomicrobia were the 10 most abundant phyla, comprising approximately 90% of all reads. Acinetobacter was the most abundant bacterial genus, comprising approximately 4% of all sequence reads. These phyla are found in almost all environments (e.g., soil, sediment, water, marine sponge, wastewater), and include potentially beneficial and potentially pathogenic microbes.

Overall, when analyzing all samples together we observed no significant differences by year between samples collected across 2013, 2014, and 2015; this suggests that the riverine ecosystem is stable. No significant difference was observed in species richness (or the number and relative frequency of bacteria in each sample) at any sampling location (alpha diversity) on a month-by-month or seasonal basis. However, we found significant differences in species richness between sediment and water column samples, with sediment samples showing greater bacterial diversity than water column samples. When assessing species richness in samples taken at the Calumet and O'Brien WRPs, mixed liquor samples from the aerobic tanks were significantly more diverse than influent sewage samples. This shows that the activated sludge reactors host a varied and diverse microbial life, as predicted. Species richness for microbial communities in the final effluent samples was comparable to that observed for the water column samples, including beach, samples upstream of WRP, and tributaries. In addition, no significant differences were found in samples from the Calumet and O'Brien WRPs. This suggests that the sample medium (water versus sediment or mixed liquor or sewage) was the most significant driver of community diversity, compared to sampling month or location; each sample medium or type (ambient water, sediment, effluent, etc.) carries a distinct measure of diversity in terms of number or species represented and their frequency of distribution.

Core microbiomes (bacterial genera shared across 90% of all samples) for water column, sediment, sewage, and secondary-treated effluent samples were each computed. Sewage samples harbored the largest core microbiome as compared to secondary-treated effluent, sediment, and water column samples. Pairwise comparison of Operative Taxonomic Units (OTU) sharing between the different sampling media revealed the greatest OTU overlap between secondary-treated effluent and sewage samples; sediment samples share the fewest OTUs with the other environments.

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 water, but no significant differences in beta diversities were observed by sampling month or year. Once again, this suggests that the type of sampling media (water, sediment, effluent, etc.) has a larger effect on microbial community composition than does sampling month or year. Finally, beta-diversity analysis demonstrated that the influent sewage and secondary-treated final effluent samples clustered closely together, indicating a closer

similarity, and had bacterial community profiles more similar to the water column in the CAWS samples than to sediment. This suggests that sediments are not the highest contributors to microbes in the water column.

Source tracking methods were used to apportion microorganisms to known potential sources. The source tracking analysis to date (2013–2014 data) shows wide variability in the potential origin of bacterial communities across sites. Strikingly, the likelihood of human stool–associated bacteria being present in these samples using the metagenomic method was low. Although source apportionment was shown to change dramatically over seasons and across sites, these sources provide evidence for potential contamination events.

All samples were also analyzed to determine the presence and distribution of human fecal and sewage contamination indicators. Sediment and water column samples comprised members of genera Bifidobacterium and Bacteroides, which are both indicators of human fecal contamination (more detail is found in Figures S6 and S7). Likewise, members of genera Acinetobacter, Arcobacter, Pseudomonas, and Thiothrix, all of which represent sewage contamination, were also identified in sediment and water column samples (Figures S8, S9, S10, and S11). Sampling locations immediately downstream of the two WRPs typically contained higher abundances of these indicators. This is particularly exemplified in the presence of Thiothrix, which was only found at a relatively high abundance in water column samples downstream of the O'Brien WRP. However, most of these indicators, including Arcobacter, Acinetobacter, Bacteroides, Bifidobacterium, and Pseudomonas, were also identified at relatively high abundances upstream of the two WRPs, thus lowering their suitability as indicators of WRP-derived sewage and human fecal contamination.

It is important to emphasize that despite the conclusive evidence of these fecal and sewage indicators, their presence (determined by 16S rRNA gene-based analysis) provides no information about their absolute abundance (quantitative measurements), virulence, pathogenicity, or viability. More information on the occurrence of virulence markers associated with *E. coli* is presented in the section on metagenomes; we are still processing and analyzing these data to determine further profiles of interest, including antibiotic resistance and viral organisms. Quantitative assessment methodologies such as qPCR need to be developed in association with the methods used in this study to relate the results from this analysis to any quantitative method commonly used to determine water quality.

Other attempts at understanding the relationships between typical culture-based methods and modern molecular-based methods included the sequencing of culture plate samples used by the MWRD to count fecal coliform and *E. coli* in effluent samples. Results showed that the cultured bacteria represented a number of different *E. coli* strains, but that only 0.03% of the genes identified in these cultures were related to genes associated to virulence, disease, and defense subsystems. Notwithstanding genome incompleteness, these results suggest that not all *E. coli* strains grown on the culture plates were pathogenic, and that some of the cultures were not of *E. coli* but rather general coliforms.

Using the extensive metadata collected by MWRD for 2013, 2014, and 2015, we investigated the effects of land-use type on CAWS water- and sediment-associated

physiochemical properties. This analysis showed that most water- and sediment-associated properties were significantly related to land-use type (e.g., commercial, recreational, residential).

Finally, our 16S rRNA analysis compared the microbial community diversity during wet and dry sampling events. Results showed that microbial diversity (number, distribution, and profiles of microorganisms) did not change significantly between samples collected during dry weather (dry events) and samples collected after precipitation. This may be due to the lack of changes in the overall community structure. This is represented by no change in the membership of observed microbes; however, this does not preclude changes in the absolute abundances (quantities) of key microbes such as fecal and sewage contamination indicator bacteria. Quantitative methods such as qPCR assays are essential to validate this hypothesis in 2016.

Shotgun metagenomics analysis provided complexity trends similar to those from the 16S rRNA amplicon analysis, showing that water samples had the lowest gene functional diversity and richness compared to sediment. The genera Rhodobacter, Novosphingobium, Synechococcus, Sediminibacterium, and Polynucleobacter were again differentially represented across sewage and ambient water samples. Polynucleobacter (a freshwater ecology indicator) 16S rRNA sequences were resolved to the strain level using oligotyping. Geographic localization (linear distance between sampling sites) had no significant correlation with either OTU or oligotype distribution, which suggests that physicochemical factors, and hence local adaptation, shapes the diversity of Polynucleobacter strains.

Protein-coding genes from shotgun metagenomes from 2013 were cataloged to depict the functional potential and distribution of potential virulence genes for the microbial community in sediment and water across the CAWS. The virulence markers associated with *E. coli* include subsets of functions associated with sites that were most likely to be contaminated with *E. coli* due to their proximity to secondary-treated effluents. Strikingly, the abundance of virulence marker genes for *E. coli* was very low at all sampling locations, including those associated with secondary-treated final effluent locations. Further analysis and more samples will help us determine the temporal variance and spatial heterogeneity of these signals and to catalog the potential origin of existing *E. coli* signatures.

DEVELOPMENT OF A PREDICTIVE MODEL

Because of the intermittent nature of sampling along the CAWS, the spatial and temporal occurrence of microbial communities may not be fully captured. In addition, the analytical sampling sites are not the same as the locations where flow and stage are monitored along the CAWS; however, their hydraulic parameters (e.g., flow stage, and velocity) may be critical drivers of microbial resuspension, growth, and die-off. Therefore, to extract flow, velocity, and stage data at each sampling site, hydraulic modeling becomes the plausible alternative. It enables the integration of microbial data into a modeling framework to provide analytical and predictive assays of spatial and temporal microbial occurrences.

To prepare for the integration of hydraulic and microbial data, we built the model DuFlow from previous efforts of the MWRD to provide hydraulic modeling of the CAWS. We

focused on building the model using the 2013 data as a prototype for further development in future years. Details on the data used for the model are in the main body of the report. Model accuracy was validated using Lemont Chicago Sanitary and Ship Canal Station data. Statistical analysis of the data generated showed that the model was of high quality and could reliably predict hydraulic parameters for the CAWS.

PLANNED AND ONGOING ACTIVITIES FOR 2016

- Continue the analysis of the ambient water monitoring samples of the CAWS to determine the impact of disinfection, and TARP improvements on the microbial communities.

- Develop a better understanding of the sequencing depth needed to reliably determine the presence of regulatory species at the concentrations of interest as a function of sample diversity measures.

- Integrate the sequencing methods with methodologies such as qPCR to provide absolute quantitative measures of diversity.

- Develop the hydraulic-microbial interface using neural network modeling methodologies. Once that has been developed, we will apply the entire modeling framework (DuFlow plus neural network interface) to the entire 2013–2015 period and eventually, when available, to the 2016–2019 period. With the integration of flow and microbial data we will be able develop an integrated microbial predictive model for forecasting and analysis of alternative management scenarios.

PART 1—16S rRNA AND SHOTGUN METAGENOMIC ANALYSIS OF WATER AND SEDIMENT SAMPLES COLLECTED DURING THE CHICAGO AREA WATERWAY SYSTEM MICROBIOME PROJECT

Melissa Dsouza, Naseer Sangwan, Jarrad Hampton-Marcell, Jack A. Gilbert
Argonne National Laboratory and the University of Chicago

1 INTRODUCTION

Microbiological water quality has broad implications for economic, health, and environmental impacts. It is often complicated to monitor pathogens directly to assess water quality due to their low abundance in natural river systems; in addition, they are often difficult to culture and have patchy distributions. To develop better tests for the assessment of ecosystem health, foundational data is required to answer fundamental questions about the presence and distribution of microbial communities. In particular, how do these communities vary over time, across different sites, land-use types, and storm events?

Microbiome studies based on high-throughput deoxyribonucleic acid (DNA) sequencing have advanced to the point where we are now able to determine the effect of environmental conditions on microbial community composition; this is essential if we are to understand how the Chicago Area Waterways System (CAWS) is responding to changes in management practices. Understanding the response of the entire microbial ecosystem to changes in how the water reclamation plants (WRPs) are operated is fundamental to our appreciation of the implications for environmental health. Microbial communities can be described in terms of diversity levels, such as the number of species (e.g., richness, alpha diversity), or the relative abundance and structure of these species (e.g., biogeography and beta diversity). Standard approaches to cataloging the impacts of secondary treated effluent on the environment (such as tracking the abundance of individual human pathogens), fail to capture the dynamic response of the ecosystem to such perturbations. These methods fail to capture, for example, the dynamics of potential pathogens, emergent organisms, viruses, and the ability to track functional genes that could be related to human health. Amplicon and shotgun metagenomic sequencing approaches fill this gap, substantially improving our understanding of the system and providing a platform from which to assess potential risk due to impacts that can destabilize the ecological processes of the river. Both amplicon and shotgun approaches have advantages and disadvantages, which we are working to quantify over the duration of this proposal. Amplicon sequencing provides a broad qualitative assessment of total community composition, but can fail to capture rare organisms, including human pathogens, at low sequencing depths. In addition, amplicon sequencing only really identifies a coarse-level assessment of the taxonomy of the microbial community; shotgun metagenomic improves upon this by providing data to support the characterization of functional potential and viral composition, and to link to known bacterial species virulence markers such as antibiotic resistance cassettes and toxin pathways. Leveraging a full array of tools to assess ecosystem health, and contextualizing these with regard to existing requirements for observational assessment (e.g., abundance of *E. coli*) is the primary aim of this research.

Taxonomic characterization of riverine microbiomes has revealed that seasonal taxonomic shifts could be important in predicting the effect of contamination on microbial communities. For example, a study that sampled the Zenne River in Brussels (Belgium) once per season for one year found seasonal variability in the recovery of bacterial community composition after exposure to sewage effluent (García-Armisen et al. 2014). Another study, limited to the early summer over two years, indicated that land use affected the taxonomic composition of bacterial communities in the Upper Mississippi River in Minnesota across forested, urban, and agricultural sites (Staley et al. 2014).

1.1 FECAL INDICATOR BACTERIA

Fecal indicator bacteria (fecal coliform, *Clostridium perfringens*, *Escherichia coli*, and enterococci) have been historically used to assess human health risk from waterborne pathogens in an environmental testing (Scott et al. 2002; Field and Samadpour 2007). The presence of fecal-associated bacteria is a major cause of water quality degradation in the nation's waterways and coastal regions. Epidemiological studies have already established human health standards based on quantity and exposure of fecal indicator bacteria in drinking, recreational, and shellfish waters (Field & Samadpour 2007). Because the most serious threat to human health is thought to come from human fecal contamination, untreated sewage waters are considered one of the greatest human fecal pollution sources, and so combined sewer overflows (CSOs) and sanitary sewer overflows (SSOs) are believed to be the major source of concern for human-associated pathogens (Rijal et al. 2009, 2011; Newton et al. 2013). However, fecal contamination and pathogens may also enter the waterway from storm water, agricultural runoff, leaking sanitary sewers, and other sources. These additional sources can contain not only human feces–associated microorganisms, but also fecal-associated contaminants from pets, wildlife, agricultural animals, or industrial waste. The potential disease risk of these sources is much less understood, and shows that the sources of potential contamination are numerous and complex, which makes an effective mitigation strategy difficult to devise. One way to overcome this is to define specific source metrics (percent probability of contamination coming from that source) for key locations in a waterway network, and determine the dynamic change over time.

Although conventional fecal indicator detection methods—such as culture-dependent assays of total coliforms, fecal coliforms, *E. coli* and enterococci, and culture-independent methods such as quantitative PCR (qPCR for Bacteroides, Bifidobacterium, etc.)—have been widely employed as proxies for fecal pollution in waterways, they often do not accurately represent the health of the ecosystem or the associated human risk and are unable to determine the potential source of the pollution. There are three key reasons why these techniques are less than favorable: (1) their lack of host specificity as a variety of warm- and cold-blooded animals can shed fecal indicator bacteria (Gordon and Cowling 2003), making it unclear whether *E. coli* or other fecal indicator organisms are contributed from human or animal sources; (2) an adequate fecal indicator should not reproduce outside the host, but organisms like *E. coli* and enterococci are ubiquitous in natural environments, where they can establish populations in lakes and streams, sand, sediments, plant surfaces, and other locations; and (3) an indicator should both be correlated with the presence of pathogens and have a survival profile similar to the survival profile of the pathogens whose presence it indicates (Field and Samadpour 2007). However,

although epidemiological studies have shown a correlation between gastrointestinal illness and elevated fecal indicator organism levels (Wade et al. 2003), the relationship between indicators and other diseases or disease-causing bacteria is not significant (Savichtcheva and Okabe 2006). For example, in most instances *E. coli* and enterococci are not well correlated with pathogenic *Salmonella* spp., *Campylobacter* spp., *Cryptosporidium*, *Giardia* spp., or human enteroviruses. However, moderately positive correlation between enterococci and *Giardia* spp. populations were observed for water column samples collected from inland lakes, rivers, and Lake Michigan than in effluent-dominated waters collected from the CAWS (Dorevitch et al. 2011). Therefore, to estimate human health risk associated with exposure to contaminated waters, it is necessary to diagnose the sources of fecal contamination in water; a procedure often referred to as microbial source tracking (MST). MST relies on the assumption that some characteristics in, or associated with, feces can unequivocally identify a particular feces type or host source, and that this can be detected in water (Field and Samadpour 2007; Roslev and Bukh 2011).

1.2 MICROBIAL SOURCE TRACKING METHODS

Because conventional fecal indicator bacteria are limited in their ability to identify the source of contamination and to accurately diagnose human health risk, several alternative MST methods have been developed; however, the accuracy of many of these methods is not well documented because few of them have undergone comparative testing and/or testing with blind samples. Most MST methods and their advantages and disadvantages have been reviewed in several articles (see Scott et al. 2002; Field and Samadpour 2007; Hagedorn et al. 2011). Those methods are categorized into culture-dependent (e.g., direct culturing of organisms like Bifidobacteria; antibiotic resistance assays; DNA fingerprinting of cultured isolates such as ribotyping, REP-PCR, PFGE), and culture-independent methods (e.g., community fingerprinting [T-RFLP]; chemical methods [e.g., caffeine, fecal sterols detection]; qPCR methods; *E. coli* toxin genes analysis). Importantly, no single source-tracking method alone appeared to be ideal because most of these indicators rely on identifying a taxonomically narrow set of bacteria (e.g., a single species) and most are incapable of discriminating between human sources and some or at least one animal source (Shanks et al. 2009). Therefore, a combination of several methods or the use of several bacterial taxa—instead of a limited number of specific target genes—as source tracking indices is more appropriate to define fecal contamination status. This multi-phasic approach will enhance discrimination and/or could be used to provide confirmation of results.

The advent of high-throughput culture-independent characterization of microbial communities, which can identify thousands of organisms from environmental samples, has enabled a more in-depth characterization of bacterial community structure. These community-based approaches are better able to explore microbial fluctuations due to physical, chemical, and biological influences. Therefore, high-throughput sequencing technologies (e.g., Illumina) are proposed as a promising MST method and have received increasing attention for their use in the characterization of water contamination and in the accurate assessment of human health risk (Unno et al. 2010; McLellan et al. 2010; Newton et al. 2011, 2013; VandeWalle et al. 2012). High-throughput sequencing techniques, through the use of multi-taxon signatures, could help identify many new source-specific targets and improve sensitivity and specificity for tracking fecal pollution sources.

1.3 HIGH-THROUGHPUT SEQUENCING AS A MST TOOL IN SEWAGE AND WATERWAYS

Some of the first studies using high-throughput sequencing tried to disentangle the overlap among human and other animal sources of microbes. These studies focused on characterizing host-specific fecal microbiota by analyzing human and animal feces, in order to then use that indicator information as an MST tool in water environments. Lee and colleagues (2011) sequenced the V2 region of the 16S rRNA gene of humans, chickens, cows, pigs, and geese using a high-throughput sequencing approach; results indicated that although the general compositions of the gut microbiota in humans and other vertebrates were similar (with members of Actinobacteria, Proteobacteria, Firmicutes, and Bacteroidetes appearing in all fecal samples), specific differences related to microbial diversity and the presence of specific microorganisms that might be useful as host-specific biomarkers were identified in each host gut. For example, *Bifidobacterium* represents fecal contamination from humans, *Yania* spp. is a specific indicator for chicken fecal contamination, *Agromyces* spp. for goose, and *Marinicola* spp. for pig fecal contamination. However, this study did not include watershed samples, so the utility of such host-specific microorganisms in water-based environments remains undetermined.

Unno and colleagues (2010) proposed a new parameter to assess fecal contamination in watersheds, based on the percentage of pyrosequencing-derived shared operational taxonomy units (OTUs) between watershed and intestinal microbiota. Using the V1–V3 region of the 16S rRNA gene to characterize the microbiome of human and farm-animal feces (chickens, ducks, beef cattle, dairy cattle, and swine) they determined the percent contribution these sources made to the watershed, showing that the majority of reads in the shared OTUs belonged to the phyla Bacteroidetes, Firmicutes, and Proteobacteria. They also showed that the greatest overlap of shared OTUs between fecal and environmental samples was identified with human and swine fecal samples at an urbanized agricultural area of the Yeongsan River (South Korea) and at an open area with no major industrial activities (these sites showed ≥1600 Colony Forming Units, or CFU/100 mL and 940 CFU/100 mL *E. coli* counts respectively). However, a third site in the river—representing a typical agricultural area and accounting for ≥1600 CFU/100 mL of *E. coli* counts—shared most of its OTUs with geese fecal samples. Only a few or no sequences in each of the fecal samples analyzed were classified as *E. coli*. This is in agreement with previous studies that reported that *E. coli* typically comprised ~1% of the total gut bacteria from these samples (Dowd et al. 2008).

McLellan and colleagues (2010) used 16S rRNA gene pyrosequencing (V6 region) to attempt to identify a set of new alternative fecal indicators that could track human fecal-associated bacteria in sewage and river water (during overflows), as well as to gain insights into the composition of low-abundance and dominant populations in microbial communities released into the environment as a result of sewage overflows. Eight untreated sewage influent samples from two wastewater treatment plants in metropolitan Milwaukee were studied and their community profiles were compared to a river surface water sample and to the microbial community observed in human feces samples from the Dethlefsen et al. (2008) and Turnbaugh et al. (2009) projects. A human fecal signature was identified in the sewage samples (comprised of several taxonomic groups including multiple Bifidobacteriaceae, Coriobacteriaceae, Bacteroidaceae, Lachnospiraceae, and Ruminococcaceae genera); however, a greater proportion

of the sewage tags belonged to taxonomic groups within Gammaproteobacteria (i.e., Acinetobacter, Aeromonas, and Pseudomonas) in addition to Arcobacter (Epsilonproteobacteria) and Trichococcus (Bacilli), reflecting that sewage microbial communities form a unique population structure. Although the fecal signature comprised a small fraction of the taxa present in sewage, these genera were much more prevalent in the sewage influent than standard indicators species, which were extremely rare (*E. coli*, *Enterococcus*, and *Clostridium perfringens* accounted for <0.7% of the total tags, even when including taxa classified to those organisms at a high taxonomic levels).

As a continuum of the previous study, VandeWalle and colleagues (2012) conducted a more detailed study of sewage influent microbes by analyzing the population structure and temporal dynamics within sewage influent through 16S pyrotag sequencing of the V6 region. The three most dominant taxa in sewage were *Acinetobacter* (16.1%), *Aeromonas* (9.8%), and *Trichococcus* (7.7%) although they occurred in low abundance in uncontaminated surface waters. Only a small fraction of pyrotags from influent samples (~15%) matched sequences from human fecal samples. Lactococcus (1.7%) and Enterobacteriaceae (1.6%) were enriched in sewage samples compared to human datasets. Increases in sewage-associated organisms were detected during combined sewer overflow (CSO) events, with this contaminated water containing a 20–30 times higher relative abundances of sewage-specific indicators (Acinetobacter, Trichococcus, and Aeromonas) and fecal taxa (Lachnospiraceae and Bacteroides).

Newton and colleagues (2011) examined water fecal contamination from the Milwaukee harbor with conventional and alternative indicators. They also used pyrosequencing to identify and develop a new quantitative PCR (qPCR) assay for a Lachnospiraceae phylotype (Lachno2) that they found was highly abundant in sewage influent and prevalent in human fecal communities, but not in cow samples. Pyrosequencing data were used to characterize both the water treatment plant influent samples community and the harbor microbial community during dry weather, rain, and combined sewer overflow events. The prevalence of human fecal pollution was also analyzed by conventional methods such us *Bacteroidales* spp. qPCR, and conventional *Escherichia coli* and enterococci plate count. Authors showed that Lachnospiraceae and human Bacteroidales had increased specificity to detect sewage compared to conventional indicators, and the presence and abundance of those organisms were correlated to human adenovirus occurrence, which suggests that these alternative indicators could be useful in improving assessments for human health risks in urban waters.

Newton and colleagues (2013) also conducted a similar study to identify signatures of sewage and fecal pollution derived from the Milwaukee river systems into waters of Lake Michigan by sequencing the V6 and V6–V4 hyper-variable regions. Samples were collected in a variety of weather scenarios to help scientists better understand the fluctuation of microbial community. No rain (dry weather) samples and samples after rain were collected after a 48-h rainfall total of <1.2 cm and >2.5 cm prior collection. Authors also collected CSO samples during or directly following combined or sanitary sewer overflows. In addition, the authors explored the extent of the fecal bacterial footprint imposed by the discharge of Milwaukee system in Lake Michigan. A microbial signature associated with sewer, nonhuman fecal, and human fecal pollution was identified. Acinetobacter, Arcobacter, and Trichococcus sequences were the sewer-associated genera, while Bacteroidaceae, Porphyromonadaceae, Clostridiaceae,

Lachnospiraceae, and Ruminococcaceae served as fecal contamination signature. Differences in the abundances of these indicators were observed during the different weather scenarios: the relative contribution of the sewer and fecal signature increased to >2 % of the measured surface water communities following sewer overflows. The ratio of the human fecal pollution signature to the nonhuman fecal pollution signature in combined sewer overflows was generally close to that of sewage, suggesting a human-associated source at overflow events. However, this ratio decreased during dry weather and rain events, suggesting that nonhuman fecal pollution was dominant during those scenarios. The qPCR detection of the two human fecal indicators used in the previous study (*Bacteroides* qPCR and Lachno2, (Newton et al. 2011) indicated the extent of the urban fecal footprint, offshore lake Michigan.

2 STUDY AIMS AND OBJECTIVES

This list of questions guided our analyses of the microbial communities associated with the CAWS:

1. Does microbial species diversity show differential geographic and temporal structure?
   a. Are the observed differences correlated with sampling medium (sediment vs. water column vs. effluent)?
   b. Are the observed differences correlated with sampling time points (year and month)?
   c. Are the observed differences correlated with sampling site? And in particular, is there an effect of sampling site location (upstream or downstream of a WRP)?

2. What is the relative abundance of fecal indicator organisms (FIOs)?
   a. Does FIO abundance decay with distance from point sources?
   b. What are the functional attributes of potential FIOs?

3. What are the potential sources of microbial organisms at different points in the CAWS?
   a. Does source apportionment for a particular location vary in different seasons or years?
   b. Are sources highly local, or are they more general across the CAWS?

4. How does land use influence microbial community structure?
   a. Does land use influence physicochemical properties in the CAWS?
   b. Do different land types influence source apportionment?

3 MATERIALS AND METHODS

The locations of sampling and physiochemical analysis sites are illustrated in Figure 1 and Table 1. Samples were collected and prepared (via filtration) by MWRD personnel and delivered to Argonne periodically. Quality assurance/quality control (QA/QC) procedures and methods for the storage and processing of the samples are discussed in the Quality Assurance Project Plan (QAPP) dated February 4, 2015.

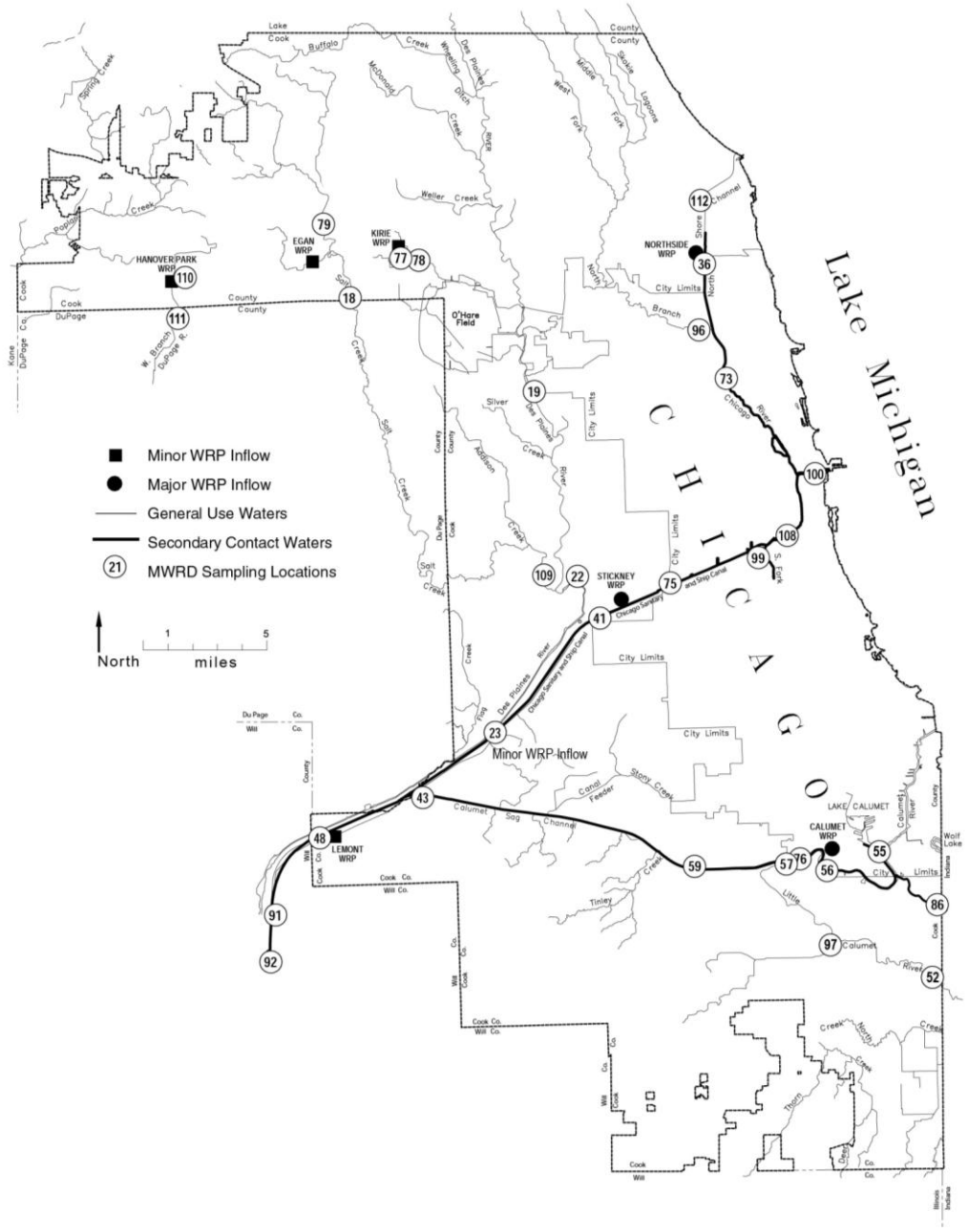

FIGURE 1 Map of sites under investigation during the 2013, 2014, and 2015 sampling seasons.

TABLE 1 Location details for each site.

Site	Waterway System	Street	Reference Distance
A. CAWS North			
112	North Shore Channel (NSC)	Dempster Street	~1.3 miles upstream from O'Brien WRP
36	North Shore Channel	Touhy Ave.	~0.7 miles downstream from O'Brien WRP
73	North Branch Chicago River	Diversey Ave.	~6.7 miles downstream from O'Brien WRP
B. CAWS North Tributary			
96	North Branch Chicago River[a]	Albany Ave.	Tributary river ~3.4 miles from O'Brien WRP (confluence 3.4 miles from WRP, actual station 3.5 miles)
C. CAWS Main Stem			
100	Chicago River Main Stem	Wells St.	Downtown Chicago River ~11 miles from O'Brien WRP (11.1 miles to actual station)
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) ~14.7 miles
E. CAWS Calumet River			
86	Grand Calumet River[a]	Burnham Ave.	Upstream tributary ~4.4 miles from Calumet WRP
55[b]	Calumet River[a]	130th St.	Upstream tributary ~5.6 miles from Calumet WRP
56	Little Calumet River[a]	Indiana Ave.	~1 mile upstream from Calumet WRP
76	Little Calumet River[a]	Halsted St.	~1.2 miles downstream from Calumet WRP
57	Little Calumet River[a]	Ashland Ave.	Tributary River ~1.7 miles from Calumet WRP
52[b]	Little Calumet River[a]	Wentworth Ave.	Tributary River ~1.7 miles from Calumet WRP
97[b]	Thorn Creek[a]	170th St.	Tributary River ~1.7 miles from Calumet WRP
F. CAWS Cal-Sag Channel			
59	Cal-Sag Channel	Cicero Ave.	~6.3 miles downstream from Calumet WRP
43[b]	Cal-Sag Channel	Route #83	~17.2 miles downstream from Calumet WRP

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

[b] Sites sampled in 2014–2015 to document baseline conditions in the Calumet River System in the 2 years preceding completion of the Calumet TARP System's Thornton Composite Reservoir.

3.1 ASSESSING MICROBIAL COMMUNITY STRUCTURE ACROSS THE CAWS OVER 3 YEARS USING 16S rRNA AMPLICON SEQUENCING

We have utilized a 16S rRNA amplicon high-throughput sequencing approach to characterize the microbial communities associated with the CAWS. During this reporting period, we processed 196 blank (equipment, filter) samples, 9 *E. coli* (BioBall®) spiked samples, 27 fish gut samples, 24 fish mucus samples, 278 sediment samples, and 429 water column samples from 17 sites in the Chicago River and artificial canals (Figure 1). In addition, 22 influent sewage, 10 mixed liquor, and 190 secondary treated final effluent samples from two WRPs (O'Brien and Calumet) (Figure 1, Table 1) collected from 2013 through 2015 were also analyzed, in addition to 7 Lake Michigan beach water samples collected in 2015 during river backflow. Additional information on the number of sediment and water column samples that were processed is included in Table 2. A complete list of samples is included in Table S1 (Appendix A).

TABLE 2 Summary of sediment and water column samples by sites on the CAWS.

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

3.1.1 Amplicon Based Microbial Community Sequencing Analysis

Microbial community structure was assessed using standard DNA extraction and 16S rRNA V4 region amplicon sequencing methods (see www.earthmicrobiome.org/emp-standard-protocols; Caporaso et al. 2012). Genomic DNA was extracted using the Powersoil-htp 96 WellDNA isolation kit (MoBio) with a 10-min (65°C) incubation step modification (see http://www.earthmicrobiome.org/emp-standard-protocols/dna-extraction-protocol). For the 16S

rRNA gene amplicon analysis, the V4 16S rRNA region was amplified in triplicate for all samples. The amplification primers were adapted from the Caporaso et al. (2010) protocol to include nine extra bases in the adapter region of the forward amplification primer that support paired-end sequencing on the HiSeq/MiSeq. PCR products were pooled at equimolar concentrations and cleaned using the UltraClean® PCR Clean-Up Kit (MoBio). The 16S rRNA amplicons were sequenced at the IGSB Next Generation Sequencing Core at Argonne National Laboratory using 151 bp paired-end sequencing on an Illumina MiSeq instrument. Paired (forward and reverse) raw sequences were de-multiplexed and quality filtered using Quantitative Insights into Microbial Ecology (QIIME, v.1.9.0) (Caporaso, Kuczynski et al. 2010) and VSEARCH (see https://github.com/torognes/vsearch). OTUs were clustered using cluster_otus in USEARCH (v.8.0) at 97% sequence similarity (Edgar 2010). OTU sequences were aligned using PYNAST (Caporaso, Bittinger et al. 2010). OTU taxonomy was determined using the RDP classifier retrained on the GreenGenes database (97% similarity) (Wang et al. 2007). A tree was constructed after filtering gaps from the aligned set of OTU representative sequences using FastTree (Price et al. 2009). The final OTU table did not contain any chimeric sequences or singletons. Downstream data analysis was conducted using QIIME (Caporaso, Kuczynski et al. 2010) and Phyloseq and Vegan packages in RStudio (v.0.99). Core microbiomes, shared phylotypes, and SourceTracker analyses were performed in QIIME 1.9.0.

3.1.2 Statistical Analysis

Statistical analysis was performed using the Vegan package or in SPSS (v.21.0) and processed in RStudio™. To avoid biases generated by differences in sequencing depth, the OTU table was rarified to an even depth of 5,000 sequences per sample when comparing all the samples from this study. In addition, OTUs represented by fewer than five reads were removed. Principal coordinate analysis (PCoA) plots were utilized to analyze all samples. Microbial alpha diversity between groupings such as sampling media, sampling year, and season were assessed for significance using a nonparametric two-sample t-test over 999 Monte Carlo permutations. Beta diversity of all samples was determined using weighted and unweighted UniFrac distance matrices. Procrustes analysis (least-square orthogonal mapping) was performed in QIIME 1.9.0 to test for the goodness of fit (i.e., to determine whether the same beta-diversity conclusions can be derived regardless of the distance metric used to compare samples). Beta-diversity clustering was analyzed using analysis of similarity (ANOSIM) and permutational multivariate analysis of variance (PERMANOVA) for categorical variables. All water chemistry data utilized for the analysis of land-use effects on CAWS is reported and summarized in Table S2.

3.2 DNA EXTRACTION, ASSEMBLY, AND ANNOTATION OF *E. COLI* GENOMES

E. coli and fecal coliforms were cultured from WRP effluent samples collected from May to April 2013 onto a selective culture medium (mTEC medium for *E. coli* and mFC medium for all other fecal coliforms). One colony per plate was randomly picked and its genomic DNA was extracted using the MoBio Powersoil DNA kit. A library was constructed for each sample using Nextera XT kit, which includes enzymatic fragmentation and simultaneous tagging of DNA followed by a bead-based sample normalization. Libraries from 14 samples were pooled and

sequenced on the Illumina Miseq platform. Sequences were quality trimmed using FASTX-toolkit, and the cutadapt tool was used to remove adaptors from the reads. Read for each samples were assembled using Velvet, a *de novo* genome assembler. The constructed contigs (kmer = 47) were used to predict genes (ORF) through FragGeneScan and Prodigal. We checked for single copy genes (Wu et al. 2012) using the Amphora software, which detects 31 bacterial single-copy marker genes. The RAST online tool was used to annotate those genomes and classify them to their closest neighbor.

3.3 ASSESSING MICROBIAL COMMUNITY STRUCTURE AND FUNCTION ACROSS THE CAWS USING METAGENOMIC SEQUENCE DATA

3.3.1 Quality Filtering, Coverage Estimation, Metagenome Assembly and Annotation

Paired-end metagenome reads were quality trimmed using nesoni (see http://vicbioinformatics.com/nesoni.shtml) with the following parameters: minimum length = 75; quality cutoff = 30; adapter trimming = yes; and ambiguous bases = 0. Taxonomic and functional information was assigned to the individual metagenome reads using MetaPhlAn (Segata et al. 2012) and MGRAST (Meyer et al. 2008), respectively. Individual read-based functional annotations were used for the functional diversity and richness estimation. Quality trimmed metagenomic reads were assembled into contigs using IDBA_UD (Peng et al. 2012) using k-mer length ranging between 31 and 41. Metagenome contigs with lengths less than 300 bp were excluded from further analysis. Metagenome contigs were assigned to various taxonomical levels using the NBC classifier (Rosen et al. 2008). Average metagenome coverage and sequence diversity was computed for each sample using Nonpareil (Rodriguez and Konstantinidis 2014) set at default parameters. AGS (average genome size) was computed for each metagenome sample using MicrobeCensus (Nayfach and Pollard 2015). FragGeneScan (Rho et al. 2010) was also used to predict the protein coding genes across metagenome contigs. Functional annotation of individual metagenome reads and contigs (ORFs) was performed using paladin (https://github.com/twestbrookunh/paladin) and prokka (Seemann 2014), respectively.

3.3.2 Genotype Binning and Population-Level Comparative Genomics

In order to understand population-level dynamics (taxonomical, functional and evolutionary), we focused our further assembly efforts to bin population genomes (genotypes, not individual genomes) for this taxon. Tetranucleotide frequency usage and %G+C values were computed for each metagenome contig using 2TBinning (Saeed et al. 2012). Contigs were clustered into bins using hierarchical agglomerative clustering (HAC) performed with an inter-profile correlation cutoff (R^2) of 0.9. Chimeric contigs (i.e., %G+C profiles >± 1 of sample [bin] mean) were removed from the individual population bin. Population genome bins were further screened (Nmer = 12) for the contaminants (assigned to different taxons) using NBC Classifier (Rosen et al. 2008). Single copy marker gene based CNV (copy number variation) analysis (Kerepesi et al. 2014) was used to estimate the number of species across each bin. To predict the number of species across each site, single copy genes were clustered at 97% identity.

Pseudogenes were predicted across population genomes using GenePRIMP (Pati et al. 2010). Reconstructed population genomes were uploaded to the RAST server (Aziz et al. 2008) for automated genome annotation.

4 RESULTS

4.1 ASSESSING MICROBIAL COMMUNITY STRUCTURE ACROSS THE CAWS OVER 3 YEARS USING 16S rRNA AMPLICON SEQUENCING

QA/QC was conducted by analyzing a total of 196 blank samples comprising equipment blanks, filter blanks, and trip blanks for the three sampling years. Blanks serve as indicators of microbial contamination associated with any equipment or reagent used for sampling and analysis. The large majority of blank samples showed DNA concentrations below 1 ng/µL. Samples containing less than 0.2 ng/µL are typically considered sterile, because they contain DNA quantities that cannot be reliably amplified by a standard PCR reaction. Only seven of all the blank samples we sequenced and analyzed showed DNA concentrations greater than 1 ng/µL. Upon further analysis, we determined that these seven samples clustered with CAWS water column samples. We verified that these seven samples were extracted together with CAWS water samples. This suggests that a cross-contamination event occurred during one particular DNA extraction event. Because of the limited extent of this cross contamination (these blanks were confirmed to contain low quantities of OTUs and low microbial content observed by PicoGreen® DNA quantification Assays), and because all results observed are consistent with previous results observed in the CAWS project and in other riverine systems, the QA/QC for the analyzed samples was deemed acceptable and the samples were maintained in the dataset. We do not believe that this contamination issue represented a significant problem for subsequent biological analysis.

Approximately 16.8 million high-quality 16S rRNA amplicons representing 24,107 unique OTUs were generated from 876 CAWS samples collected from 2013 through 2015. Acidobacteria, Actinobacteria, Bacteroidetes, Chlorobi, Chloroflexi, Cyanobacteria, Firmicutes, Planctomycetes, Proteobacteria, and Verrucomicrobia were the 10 most abundant phyla, comprising approximately 90% of all reads (Figure S1A). All CAWS samples are characterized at the genus level in Figures S1B and S1C, which are organized by sampling medium and sampling site, respectively (the same data is shown in Tables S3A and S3B). Overall, Acinetobacter was the most abundant bacterial genus, comprising approximately 4% of all sequence reads.

Microbial diversity is examined as alpha diversity (the number and relative distribution of microbes within a sample) and beta diversity (the diversity across samples). Microbial community species diversity (alpha diversity) is measured by the number and distribution of species and was estimated using the Chao1, Shannon, and Inverse Simpson metrics. Chao1 is a measure of species richness (i.e., the number of species observed in a habitat). Shannon and Simpson indices are both positively correlated with species richness and evenness (the distributed abundance of those species); Shannon gives more weight to rare species, and Simpson is weighted toward abundant species. Overall, we observed no significant differences between samples as a function of the year of sampling (2013, 2014, 2015) (Figure 2), suggesting that the riverine ecosystem is stable over time. However, all diversity indices showed significant differences in alpha diversity between sediment and water column samples, with sediment samples showing significantly greater bacterial diversity than water column samples ($p < 0.05$,

Figure 2). The mixed liquor samples had significantly more alpha diversity than did influent sewage samples ($p < 0.05$, Figure 3). In addition, alpha diversity for microbial communities in the secondary treated effluent samples was comparable to that observed for the water column samples, including beach samples, samples taken upstream of WRP, and tributary samples. No significant differences were observed between Calumet and O'Brien secondary treated final effluent samples (Figure S2).

Figure 4 summarizes the effect of sampling month on alpha diversity for microbial communities in CAWS samples, including beach, influent sewage, mixed liquor, secondary treated effluent, and water column samples. Water column samples showed no significant differences in alpha diversity based on sampling month. However, significant differences based on sampling month were observed in secondary treated effluent samples; samples collected during the autumn months (September, October, November) were more diverse than those collected in other months ($P < 0.05$). Next, we investigated the alpha diversity of CAWS water column and sediment samples based on sampling month at each sampling site (Figures S3A and S3B). Notwithstanding the considerable variation in Shannon diversity, no significant difference in alpha diversity based on sampling month was observed for water column samples at any sampling sites ($P > 0.05$, Figure S3A). However, alpha diversity for water column samples collected in April was typically lower at each sampling site than those collected in other months. Alpha diversity for sediment samples was relatively stable across all sampling months at each site (Figure S3B). However, we observed considerable differences among sampling locations in alpha diversity for the sediment samples. The location of each water sampling site relative to the two WRPs (upstream of the WRP vs. downstream) had an effect on the alpha diversity at these sites (Figure 5). Typically, CAWS sampling sites that are upstream of the two WRPs (including sites 112, 55, and 56) showed lower alpha diversity compared to those that are downstream of the WRPs.

Core microbes (bacterial genera shared across 90% of all samples) for water column, sediment, sewage, and secondary-treated effluent samples were determined (Table 3). Sewage samples harbored the largest core microbiome compared to secondary-treated effluent, sediment, and water column samples. Shared phylotypes between the different sampling media were computed after consolidation of samples taken from the same sampling medium across all sampling sites and sampling years. Pooled samples were rarified to an even depth. Pairwise comparison of OTU sharing between the different sampling media revealed the greatest phylotype overlap between secondary treated effluent and sewage samples, with sediment samples sharing the least number of OTUs with the other environments (Figure 6). Likewise, we computed shared phylotypes between water samples collected at sampling sites immediately upstream and downstream of the two WRPs at Calumet and O'Brien (Figure 7). At Calumet, sites 55 and 56 (upstream of the WRP) have the least overlap with secondary treated effluent samples and sites 57 and 76 (downstream of the WRP) have the most overlap with effluent samples. Likewise at O'Brien, site 112 (upstream) had the least overlap and sites 36 and 73 (downstream) had the most overlap with secondary-treated effluent samples. Site 96 shared the least number of OTUs with the other O'Brien sites and we hypothesize that the water from the north branch of the Chicago River has a localized effect on the taxonomic composition at this site.

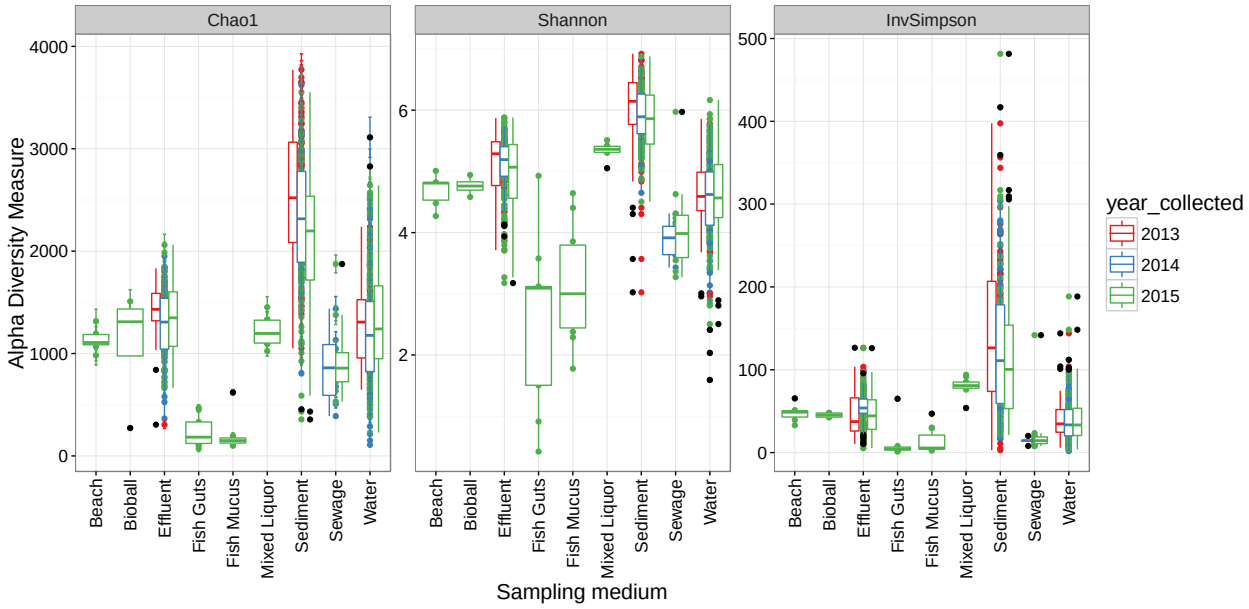

FIGURE 2 Summary of alpha diversity metrics (Chao1, Shannon, Inverse Simpson) for all CAWS samples summarized by sampling year. These are presented as Tukey boxplots wherein the first, second, and third quartiles represent 25, 50, and 75 percentiles, respectively. The upper whisker extends from the hinge to the highest value that is within 1.5 × IQR (inter-quartile range) of the hinge. The lower whisker extends from the hinge to the lowest value within 1.5 × IQR of the hinge. Data beyond the end of the whiskers are outliers and plotted in black. Figure demonstrates stable microbial communities across the three sampling years for each sampling medium. Sediment samples demonstrate the highest alpha diversity of all CAWS sampling media.

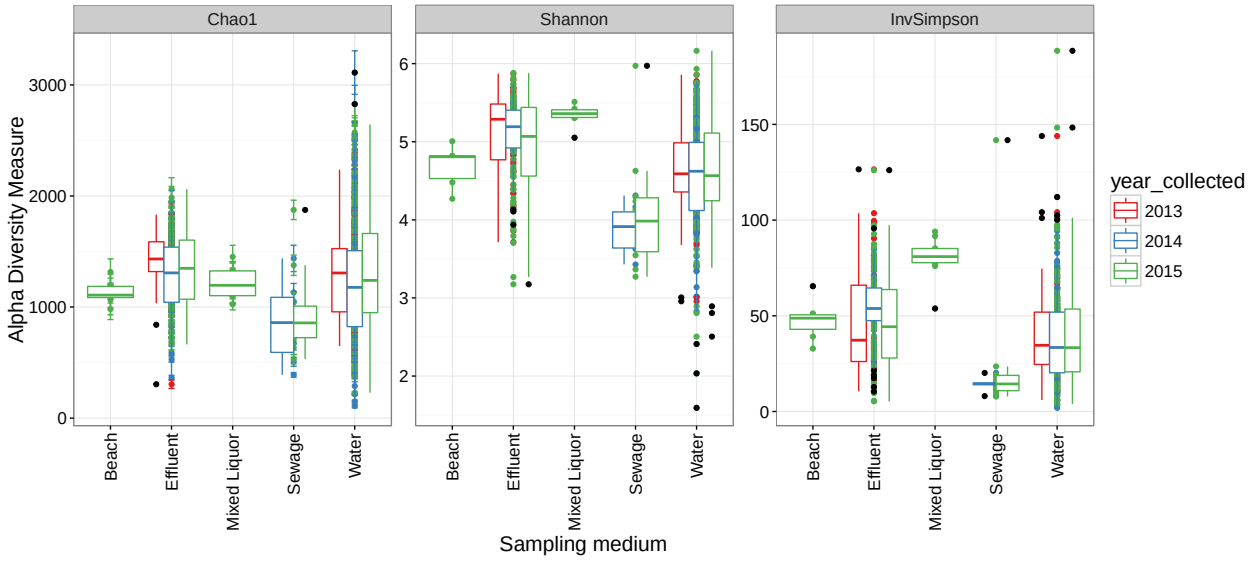

FIGURE 3 Summary of alpha diversity metrics (Chao1, Shannon, Inverse Simpson) for CAWS water-associated samples by sampling year presented as Tukey boxplots. Figure demonstrates stable microbial communities across the three sampling years for each sampling medium.

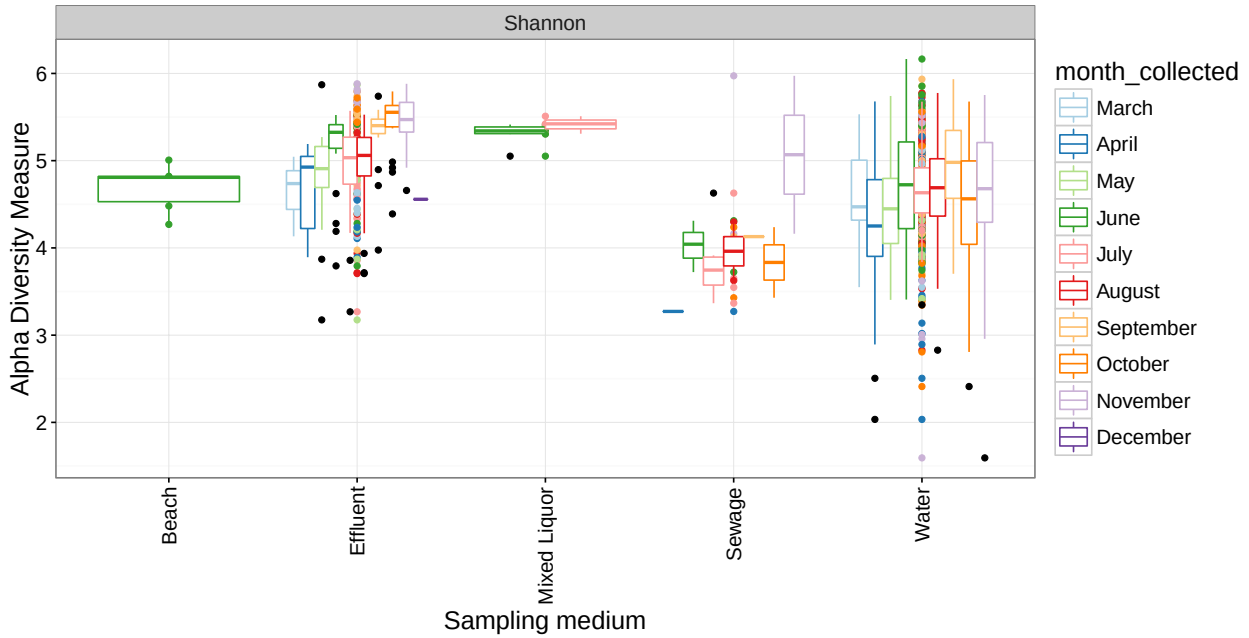

FIGURE 4 Summary of Shannon diversity for CAWS water-associated samples by sampling month presented as Tukey boxplots. Water column samples show no significant differences in alpha diversity by sampling month. Secondary treated effluent samples showed significant differences in the autumn months (September, October, November) as compared to the other sampling months, with these months being more diverse.

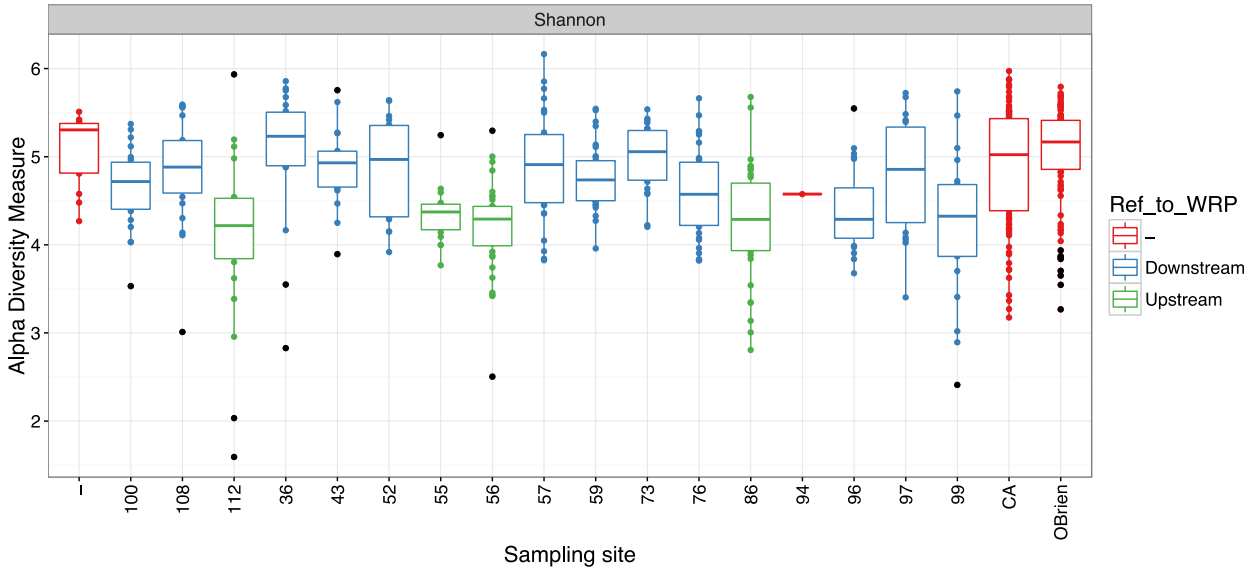

FIGURE 5 Summary of Shannon diversity for secondary treated effluent and CAWS water column samples by sampling site location (relative to the two WRPs at Calumet and O'Brien) presented as a Tukey boxplots. Sampling sites labeled with a hyphen refer to beach water and mixed liquor samples. CA and OBrien refer to Calumet and O'Brien WRPs, respectively. CAWS sampling sites that are upstream of the two WRPs typically showed lower alpha diversity compared to those that are downstream of the WRPs.

TABLE 3 Bacterial genera considered core (found in 90% of all CAWS samples) to specific sampling environments.

Sewage	Secondary treated effluent	Sediment	Water column
Acinetobacter	Acinetobacter	Crenothrix	Acinetobacter
Arcobacter	Agrobacterium	Dechloromonas	Dechloromonas
Bacteroides	Arcobacter	Desulfobulbus	Flavobacterium
Bifidobacterium	Bifidobacterium	Desulfococcus	Hydrogenophaga
Blautia	Candidatus Accumulibacter	Rhodobacter	Polynucleobacter
Cloacibacterium	Cloacibacterium	Sulfuritalea	Rhodobacter
Comamonas	Dechloromonas	Thiobacillus	Sediminibacterium
Dechloromonas	Giesbergeria	Variovorax	
Desulfomicrobium	Hydrogenophaga	WCHB1-05	
Desulfovibrio	Methylibium		
Enhydrobacter	Pseudomonas		
Faecalibacterium	Rhodobacter		
Flavobacterium	Thiothrix		
Giesbergeria	Tolumonas		
Hydrogenophaga	Variovorax		
Paludibacter	Zoogloea		
Parabacteroides			
Prevotella			
Propionivibrio			
Pseudomonas			
Rhodobacter			
Ruminococcus			
Sulfurospirillum			
Thauera			
Tolumonas			
Variovorax			
Vitreoscilla			
Zoogloea			

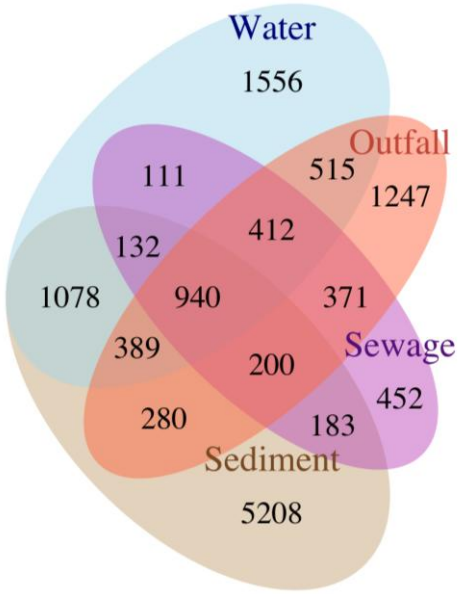

FIGURE 6 Venn diagram of shared OTUs between the different sampled media. Pairwise comparison of OTU sharing between the different sampling media revealed the greatest phylotype overlap between secondary treated effluent and sewage samples. Sediment samples shared the least number of OTUs with the other environments.

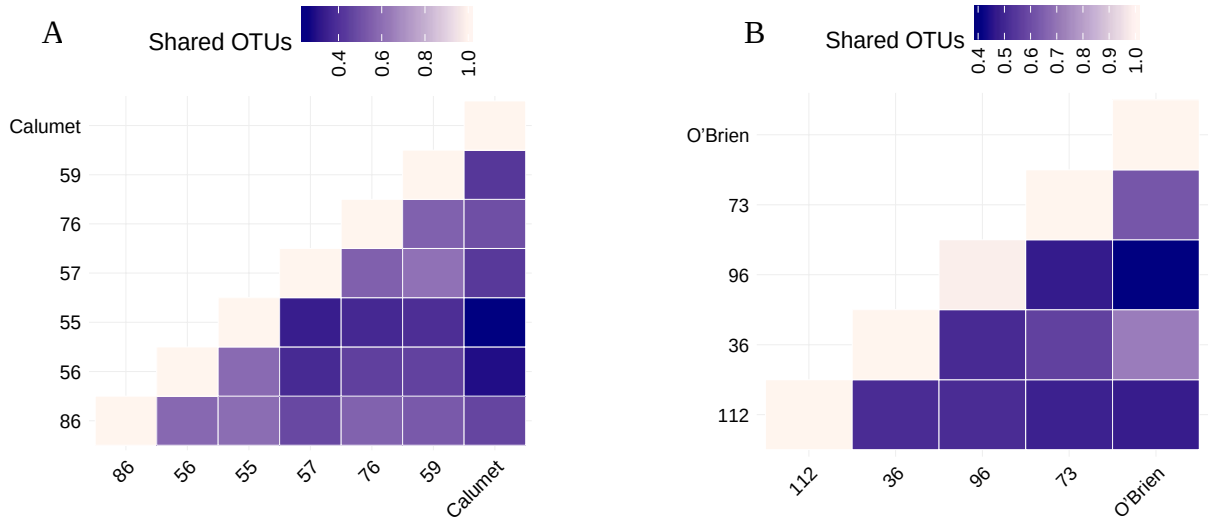

FIGURE 7 Shared OTUs between the different sampling sites located by the two WRPs at (A) Calumet and (B) O'Brien. This is displayed as a heat map wherein the quantity of shared OTUs is colored by a gradient with increased OTUs represented by light purple. In general, sites upstream of the WRP have the least overlap with secondary treated effluent samples and sites immediately downstream of the WRP have the most overlap with effluent samples.

Microbial community clustering (beta diversity) of samples (a measure of the variance in community structure between samples) was investigated using weighted and unweighted UniFrac distance matrices applied in principle coordinate space. In addition, Procrustes analysis (least-square orthogonal mapping) was performed in QIIME to test whether the same beta-diversity conclusions can be derived regardless of the distance metric used to compare samples (Muegge et al. 2011). A distance metric is a function that defines the distance between each data point of the sample set. This analysis attempts to stretch and rotate the points in one matrix, such as points obtained by principal coordinates analysis (PCoA), to be as close as possible to points in the other matrix, thus preserving the relative distances between points within each matrix. The goodness of fit, or M^2 value, of the transformed datasets (weighted and unweighted UniFrac distances matrices) was determined over the first three dimensions. The statistical significance of the computed M^2 value was measured based on 999 Monte Carlo iterations. An M^2 value of 0.483 ($p = 0.00$) was obtained, suggesting that the two distance metrics used were in remarkable agreement with each other.

Cluster comparisons between CAWS samples demonstrated 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 water (ANOSIM and PERMANOVA using weighted and unweighted UniFrac, $p = 0.001$). This was also illustrated in the PCoA plot (across just the first two principal coordinates), wherein unweighted UniFrac beta-diversity comparisons yielded significant clustering by sampling medium (Figure 8A). Similar clustering patterns were observed when sediment samples were removed from unweighted UniFrac analysis (Figure 8B). Weighted UniFrac yielded similar results, with significant clustering by sample type (Figure S4). Using unweighted UniFrac distances, no significant differences in beta-diversities were observed based on sampling month or year (ANOSIM and PERMANOVA using weighted and unweighted UniFrac, $p > 0.001$) (Figures S5). This suggests that sampling media has a larger effect on microbial community composition than sampling month or year.

Finally, beta-diversity analysis (unweighted UniFrac) also demonstrated again that the influent sewage and secondary-treated final effluent samples clustered closely together, and were more similar to the CAWS water column samples than to sediment (Figure 8A). We continue to analyze these similarities in 2016 to confirm whether the overlap in influent sewage, secondary-treated final effluent, and water column samples can be influenced by heavy rainfall or storm events; preliminary analysis is discussed later in this report. In 2016, we also plan to perform a canonical-correlation analysis (CCA) to determine the major environmental parameters responsible for driving the clusters observed in Figures 8A and 8B.

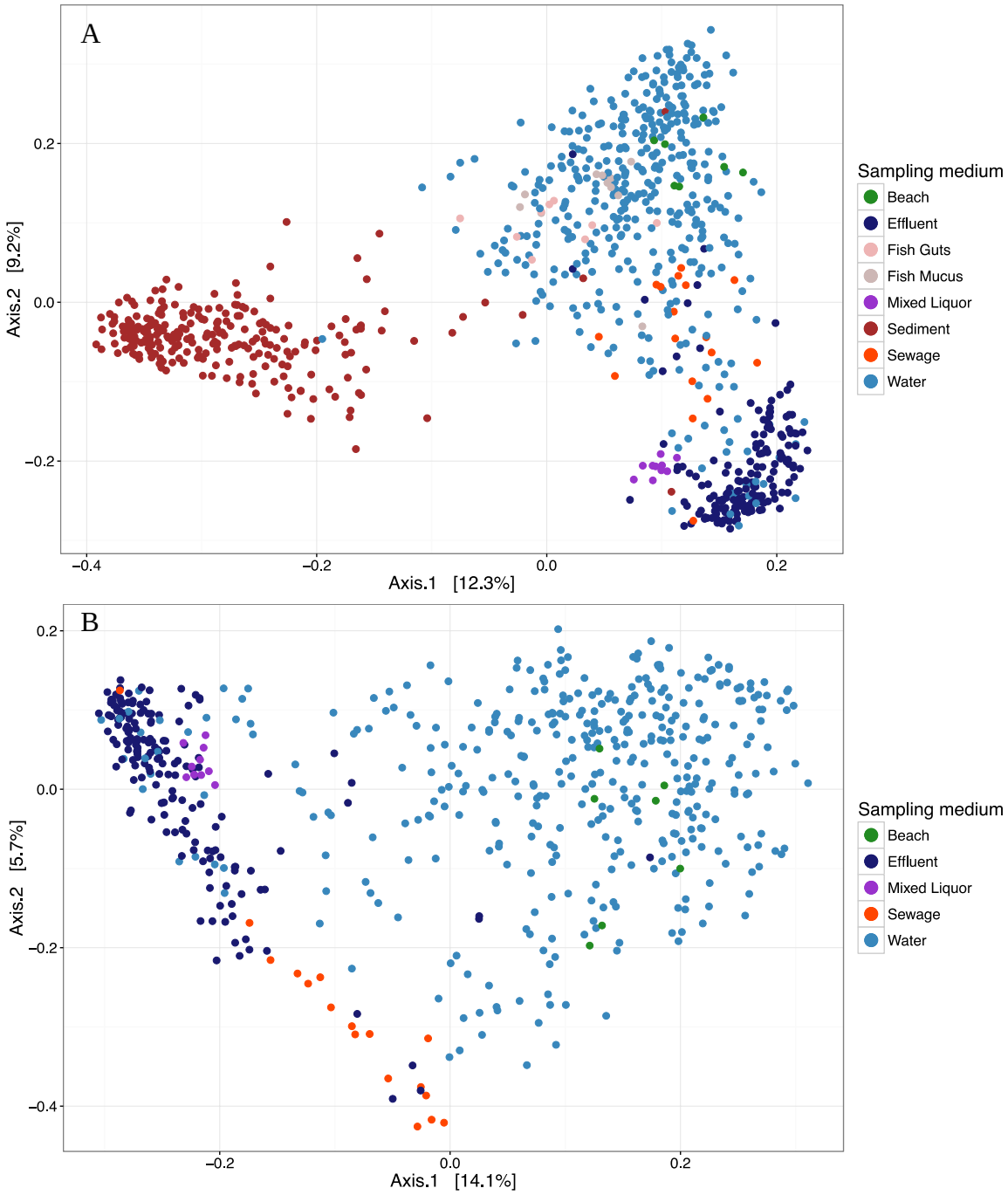

FIGURE 8 Principal coordinate plots showing sample similarities, organized by sample type, using unweighted UniFrac (A) for all CAWS samples, and (B) for only CAWS water-associated samples including beach, influent sewage, mixed liquor, secondary treated effluent, and water column samples. Cluster comparisons between CAWS samples showed significant differences in microbial community composition across the different sampling media.

4.1.1 Assessing Microbial Community Source across the CAWS for Human Fecal and Sewage Contamination over 2 Years

Previously, we used the Bayesian statistical tool Source Tracker to determine the potential source of microbial OTUs associated with each sample location and date. This tool allows us to trace specific sequences back to known originators. The results of this analysis for 2013 and 2014 are shown in Figure 9. To date, the analysis shows wide variability in the potential origin of bacterial OTUs across sites, with variability even between regions. Strikingly, the likelihood of human stool–associated OTUs being present in these samples was substantially lower than expected (Figure 9). Although source apportionment was shown to change dramatically over seasons and across sites, these sources provide evidence for potential contamination events. As shown in Figure 9, a large portion of the sequences are of unknown origin. In 2016, we plan to apply this tool to the 2015 data and cross-compare the results against the 2013–2014 data. We are also building new source databases that take into consideration the updated, relevant Earth Microbiome Project (EMP) data and all of the CAWS source data that we have generated (including influent, effluent, fish samples, etc.) so that the number of unknown sources is reduced.

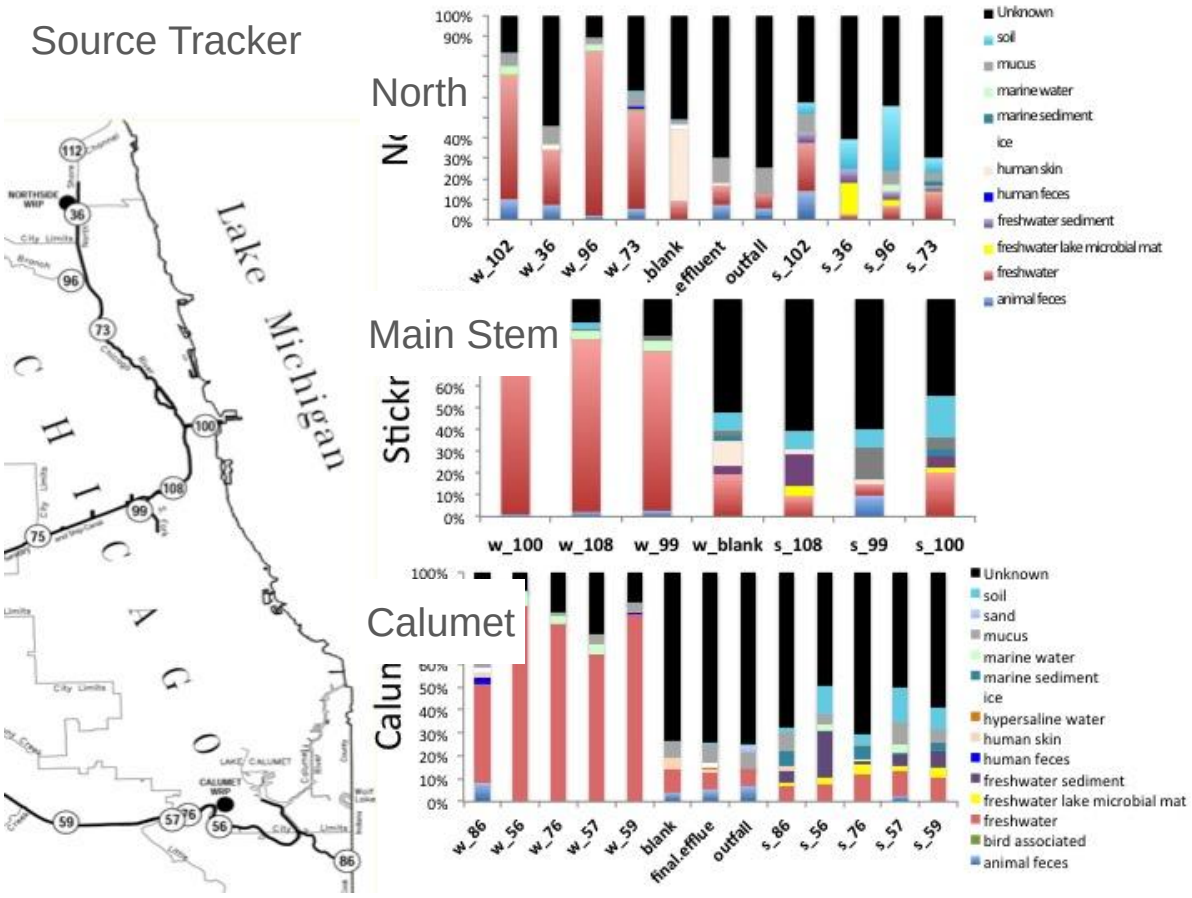

FIGURE 9 Source Tracker Analysis of 2013 and 2014 CAWS sites. Sample sites beginning with W represent water samples, while those beginning with S represent sediment.

All samples were further analyzed to determine the presence and distribution of human fecal and sewage indicators. Members of genera Bifidobacterium and Bacteroides, both representing human fecal contamination, were identified in secondary-treated effluent and water column samples (Figures S6 and S7). Likewise, members of the genera Acinetobacter, Arcobacter, and Thiothrix, all representing sewage contamination, were also identified in sediment and water column samples (Figures S8, S9, and S10). Sampling locations downstream of the two WRPs plants typically contained higher abundances of these indicators. This is exemplified in the presence of Thiothrix, which was only found at a relatively high abundance in water column samples downstream of the O'Brien WRP (site 36, Figure S11). The occurrence of most of these indicators upstream of the Calumet and O'Brien WRPs warrants further investigation into their origin at these sites. In particular, samples from CAWS location 86 (Grand Calumet River at Burnham St.) contained a high proportion of human fecal and sewage indicators. A possible explanation is that this location, which is upstream of the Calumet plant, might be receiving microbial loads from outside the Illinois border, where other large wastewater treatment facilities exist relatively nearby (e.g., the Hammond Sanitary District and the East Chicago wastewater treatment plant). Importantly, despite the conclusive occurrence of these fecal and sewage indicators, their presence (determined by 16S rRNA gene-based analysis) provides no information about their absolute abundance, virulence, pathogenicity, or viability. More information on the occurrence of virulence markers associated with *E. coli* is presented in section 4.1.3. Quantitative assessment methodologies such as qPCR assays are essential to relate the results from this analysis to any quantitative method commonly used for monitoring water quality.

4.1.2 Genomic Characterization of *E. coli* Isolates

As part of the Chicago River system water quality monitoring program, MWRD routinely measures *E. coli* and fecal coliform counts by water filtration onto selective culture medium (mTEC medium for *E. coli* and mFC medium specific to fecal coliforms). Fourteen of the culture plates corresponding to WRP effluent samples collected from May to April 2013 were analyzed through whole genome sequencing (these samples represent the original filtration, as well as 1/10 and 1/100 dilutions of these samples). Seven out of the 14 cultures presented no duplications of single-copy gene markers, suggesting that they represent the genome of a single organism. The others were thought to include more than one organism and therefore were not included in the subsequent analysis. Table 4 summarizes genome annotation results for the seven cultured isolates. Overall, these genomes showed an average of 0.03% genes attributed to the "Virulence, disease and defense" subsystem. Among this 0.03%, most of the genes (~75%) were attributed to "Resistance to antibiotics and toxic compounds," while "Bacteriocins, ribosomally synthesized antibacterial peptide" was the second most abundant function (~15%). This component of the study was performed early on as a basic assessment of what types of *E. coli* were routinely being cultured on these plates. Because of budget constraints and prioritization, we decided not to pursue this aspect of the work. However, we would be willing to revisit this at a later date to determine whether isolates obtained prior to secondary effluent sterilization are different from those obtained after sterilization. We will reanalyze the existing genome constructs to determine whether the assemblies can be improved, and as a result, the taxonomic affiliations.

TABLE 4 Summary of the genome annotation results of seven samples from MWRD culture plates without marker duplication.

Data type	O'Brien, May (mTEC)	O'Brien, May (mTEC, 1/10)	O'Brien, May (mFC, 1/10)	O'Brien, April (mTEC)	O'Brien, April (mTEC, 1/100)	O'Brien, April (mTEC, 1/100)	Calumet, April (mTEC)
Genome Size (bp)	4,502,378	4,758,466	5,023,097	4,589,399	4,791,369	4,722,285	5,535,441
No. of Contigs	192	249	182	150	234	339	2870
No. of Subsystems	579	576	583	581	586	585	496
No. of Coding seqs.	4301	4672	4942	4429	4650	4615	5655
No. of RNAs seqs.	29	31	29	50	57	66	30
Closest Neighbor	*E. coli* 88.1467	*E. coli* 88.0221	*E. coli* AA86	*E. coli* 88.1467	*E. coli* PCN033	*E. coli* 88.1467	*E. coli* PCN033

4.1.3 Determining the Influence of Land Use on Water and Sediment Physicochemical Properties across the CAWS over 3 Years

Using the extensive metadata collected by MWRD for 2013, 2014, and 2015, we investigated the effects of land use type on CAWS water- and sediment-associated physiochemical properties. First, we utilized general linear modeling to test our hypothesis that land-use type has an effect on water- and sediment-associated physiochemical properties. This analysis showed that most water- and sediment-associated properties were significantly affected by land-use type. We then performed principal component analysis (PCA) on water- and sediment-associated properties to determine PC1 and PC2 scores for further analysis. Pearson's correlation analysis was performed on PC scores and all land-use types to determine significant correlations between physiochemical properties and land-use type. Land-use types including road, residential, and open space significantly influenced water-associated properties ($p < 0.05$). Likewise, land-use types including commercial, institution, road, residential, and transport/utility significantly influenced sediment-associated properties. We also performed a Pearson's correlation analysis on these land-use types and water- and sediment-associated properties to identify significant correlations ($p < 0.05$). Water-associated properties including dissolved oxygen (DO) and sulphate (SO_4) were significantly correlated with road, residential, and open-space land-use types ($p < 0.05$). Likewise, sediment-associated characteristics including concentrations of Ag, Cd, Cr, Pb, and Zn were significantly correlated with commercial, institution, road, residential, and transport/utility land-use types (Table S4).

Analysis of variance (ANOVA) was performed to look for significant differences between all CAWS sampling locations using the water- and sediment-associated physiochemical properties for each location across the 3 years. Interestingly, in water-column samples from 2013, CAWS location 86 had significantly greater total organic carbon (TOC) compared to other CAWS locations, other than locations 57, 96, and 99, which did not differ significantly from location 86. Likewise, DO was significantly lower at locations 86 and 99, compared to DO in other CAWS locations. In 2014, similar observations were made with a significant reduction in DO occurring at locations 86, 99, and 57 compared to the remaining CAWS locations. In 2015, sites 86 and 57 showed significantly higher concentrations of SO_4 than any other location. In

2013 and 2015, the sediment at location 100 from that at all other CAWS locations; location 100 was positively correlated with a higher concentration of Cd, which correlated with higher concentrations of some other metals, such as Ag, Cr, Ni, and Pb.

4.1.4 Assessing the Effects of Wet/Dry Events on the CAWS-Associated Microbial Community

To assess the effects of wet/dry events, we classified each sampling date based on MWRD's definitions as follows:

1. Dry weather (<0.1-inch precipitation): Dry weather is defined by antecedent dry conditions for 2 days following a 0.25- to 0.49-inch precipitation event, 4 days following a 0.50- to 0.99-inch event, and 6 days following a >1.0-inch event.

2. Wet weather: Wet weather is defined by amount of precipitation and the occurrence of CSO events as follows:
   - Wet weather without CSOs (0.5–1.0 inches of precipitation). Water sampling to occur within 10 hours of the end of the rain event.
   - Wet weather with CSOs, including 125th Street Pump Station (>1.5 inches of precipitation). Water sampling to occur within 10 hours of the end of the rain event.

A total of 597 water column samples, sediment samples, secondary-treated final effluent samples, and influent sewage samples were analyzed to investigate the effect of wet/dry events on the CAWS-associated microbial community.

Alpha diversity was measured using Chao1, Shannon, and Simpson metrics. Overall, we observed no significant differences in alpha diversity between wet and dry events (Figure S11). Likewise, no significant differences in alpha diversity were observed by sampling year (Figure S12A), sampling month (Figure S12B), sampling site (Figure S13), or sampling medium (Figure 10) as a function of wet/dry events. The box plots show the minimum and maximum values, the mean value and standard deviation. The Y axis represents a combination of factors that are represented in the alpha diversity measure.

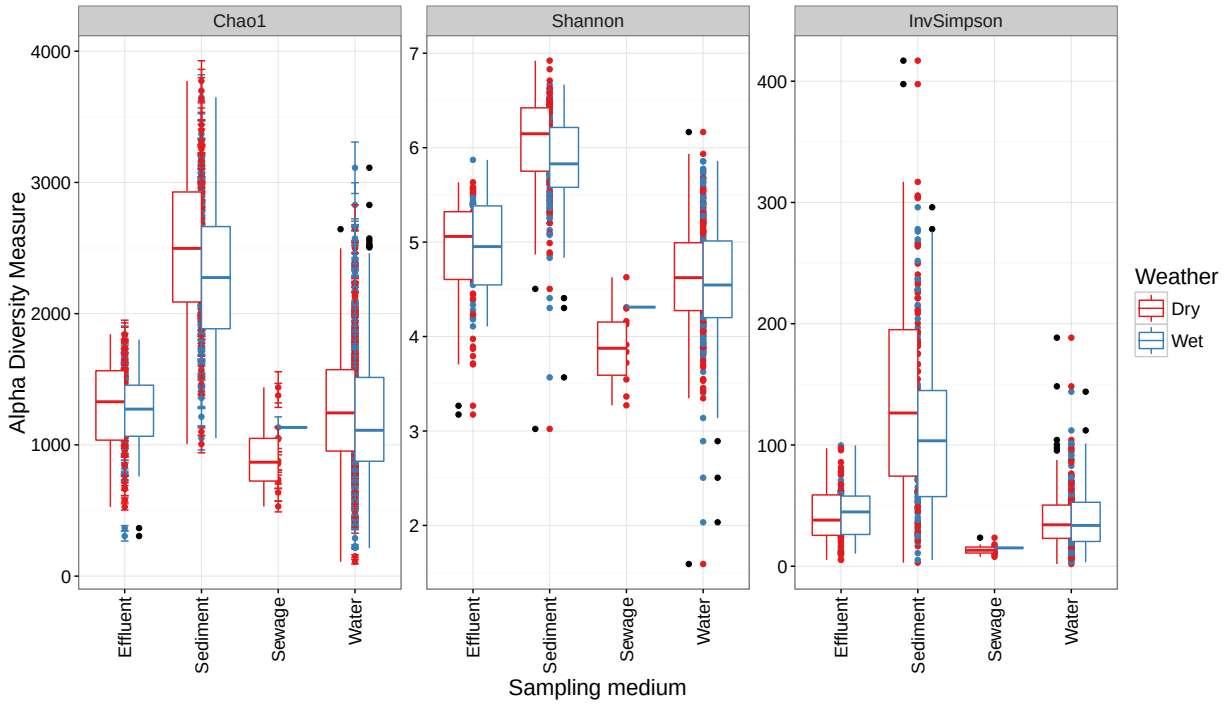

FIGURE 10 Summary of alpha diversity metrics (Chao1, Shannon, InvSimpson) summarized by sampling medium as a function of wet/dry events presented as Tukey boxplots. Figure demonstrates the lack of differences in alpha diversity between wet and dry events across the different sampling media.

Between-sample beta diversity was calculated using weighted and unweighted UniFrac distance applied in principal coordinate space. Procrustes analysis was also conducted to test whether the same beta diversity conclusions can be derived regardless of the distance metric used to compare samples. An M^2 value of 0.464 ($p = 0.00$) was obtained, suggesting that the two distance metrics used were in an agreement with each other. Cluster comparisons between CAWS samples demonstrated that there were no significant differences in qualitative microbial community composition by wet/dry events (ANOSIM and PERMANOVA using weighted and unweighted UniFrac, $p > 0.05$; Figure 11). Likewise, no significant clustering was observed using unweighted UniFrac by sampling year or month as a function of wet/dry events (ANOSIM and PERMANOVA using weighted and unweighted UniFrac, $p > 0.05$; Figure S14A, S14B). Likewise, no significant differences were observed by wet/dry events for alpha- and beta-diversity measures during the re-analysis of just the water column samples (334 of the total 597 samples, Figure S15).

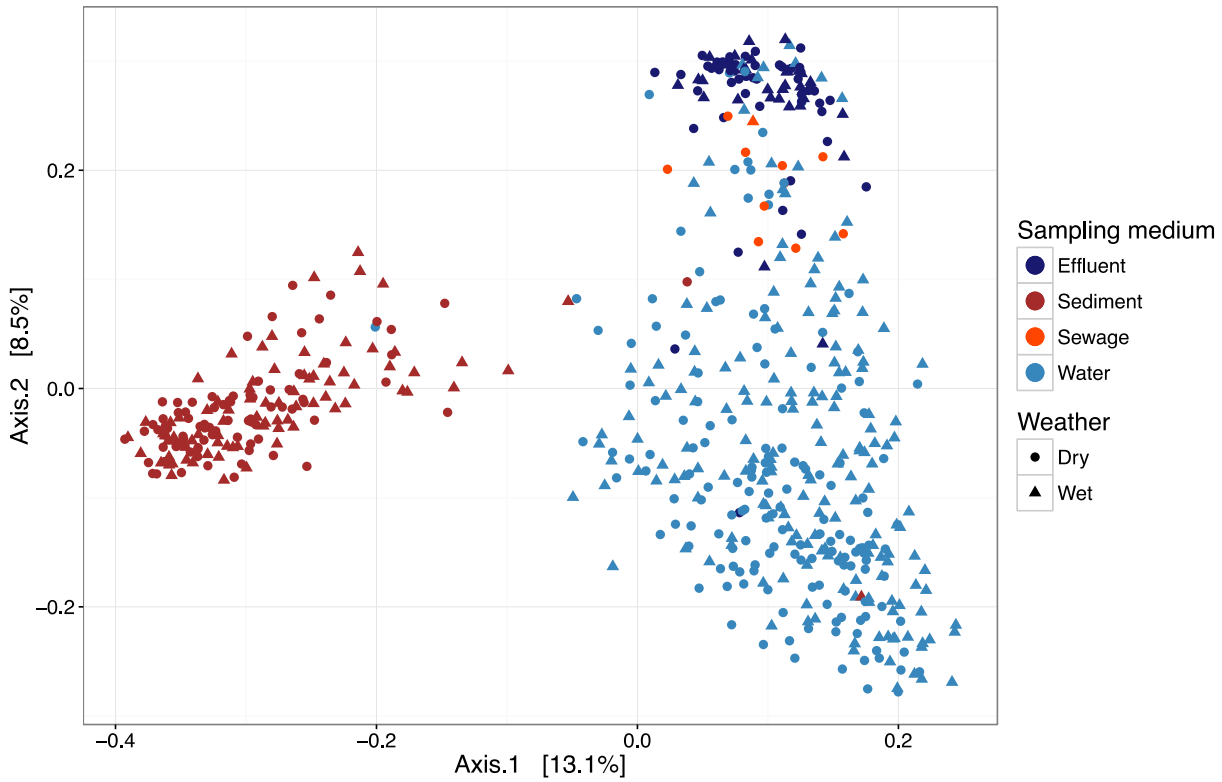

FIGURE 11 Principal coordinate plots showing similarities of samples by sampling medium and by wet/dry events using unweighted UniFrac. Cluster comparisons between CAWS samples demonstrated that there were no significant differences in qualitative microbial community composition by wet/dry events.

4.2 ASSESSING MICROBIAL COMMUNITY STRUCTURE AND FUNCTION ACROSS THE CAWS USING METAGENOMIC SEQUENCE DATA

Shotgun metagenomes for 54 CAWS samples collected during 2013 and 2014 were sequenced; results ranged from 25 to 100 million quality trimmed reads. We will process 2015 samples this year; shotgun metagenomics is time consuming. Samples were selected for shotgun metagenomics analysis by picking out of the three replications per sample for the 16S rRNA analysis the one that appeared the most representative upon analysis. The 16S rRNA gene rarefaction curves and metagenome-based sequence coverage (Rodriguez and Konstantinidis 2014) and assembly analysis produced similar sequence complexity (abundance weighted average coverage patterns) trends across the water and sediment samples. Individual metagenome read-based analysis (genus level) revealed similar alpha diversity and richness patterns, as predicted earlier with 16S rRNA amplicon data (water = 3.6 ± 0.21; sediment = 5.3 ± 2.1; and sewage = 3.6 ± 1.2). Water samples also had the smallest average genome size (2.9 Mb), which could indicate ecologically adapted, but abundant, oligotrophic genotypes (Konstantinidis and Tiedje 2005). The genera Rhodobacter, Novosphingobium, Synechococcus, Sediminibacterium, and Polynucleobacter were differentially abundant across water ecosystems.

Polynucleobacter 16S rRNA sequences were resolved to the strain level using oligotyping (Eren et al. 2013); these oligotypes had a similar pattern of reduced beta diversity within sites from the same region, while similarities decreased between regions. Oligotyping was performed on the dominant Polynucleobacter OTU (OTU32), resulting in six oligotypes. The beta-diversity pattern of these Polynucleobacter oligotypes showed a significant positive correlation with the concentration of ammonia (BIOENV; UniFrac R2 = 0.7; $p < 0.01$), as did the abundance of oligotype 2 (R2 = 0.56; $p < 0.05$). Geographic localization (km, as a pairwise calculation of linear distance between the locations calculated using latitude and longitude) had no significant correlation with either OTU or oligotype distribution, suggesting that physicochemical factors—and hence local adaptation—shapes Polynucleobacter diversity. Contig-based diversity analysis revealed significant (two group t-test; Bonferroni correction, $P < 0.05$) viral diversity trends across water and sediment samples (i.e., Caudovirales water = 3.5% ± 0.02 and sediment = 4.5 ± 0.31). Our future work includes the genotypic characterization of these strains. In addition, evolutionary trends will be analyzed for viral genotypes using nucleotide composition-based binning methods.

Using BLASTX analysis with the PATRIC database, protein-coding genes from 78 shotgun metagenomes were cataloged to depict the functional potential and distribution of potential virulence genes for the microbial community in sediment and water across the CAWS. The virulence markers we identified that are associated with *E. coli* include subsets of functions associated with sites that were most likely to be contaminated with fecal material due to their proximity to secondary-treated final effluents (Table 5; Figure 12). Strikingly, the abundance of virulence marker genes for *E. coli* was very low for all sites, including those associated with secondary-treated final effluent locations. One possible explanation for this is low sequence coverage. However, we do not believe this explanation is adequate; first, this analysis was performed on raw sequence data, as well as assembled data (raw reads assembled into long sequences) and both of the analyses showed similar trends. Second, we analyzed the data using abundance-weighted average coverage analysis to estimate whether the applied sequencing depth was appropriate to sample the observed microbial diversity. Most of the samples were covered to 75% to 90%. However, we are sequencing the metagenomic samples in 2015 at a greater depth of coverage to determine whether the *E. coli* organisms were just at extremely low abundance. Further analysis is needed to determine the temporal variance and spatial heterogeneity of these signals and to catalog the potential origin of existing *E. coli* signatures.

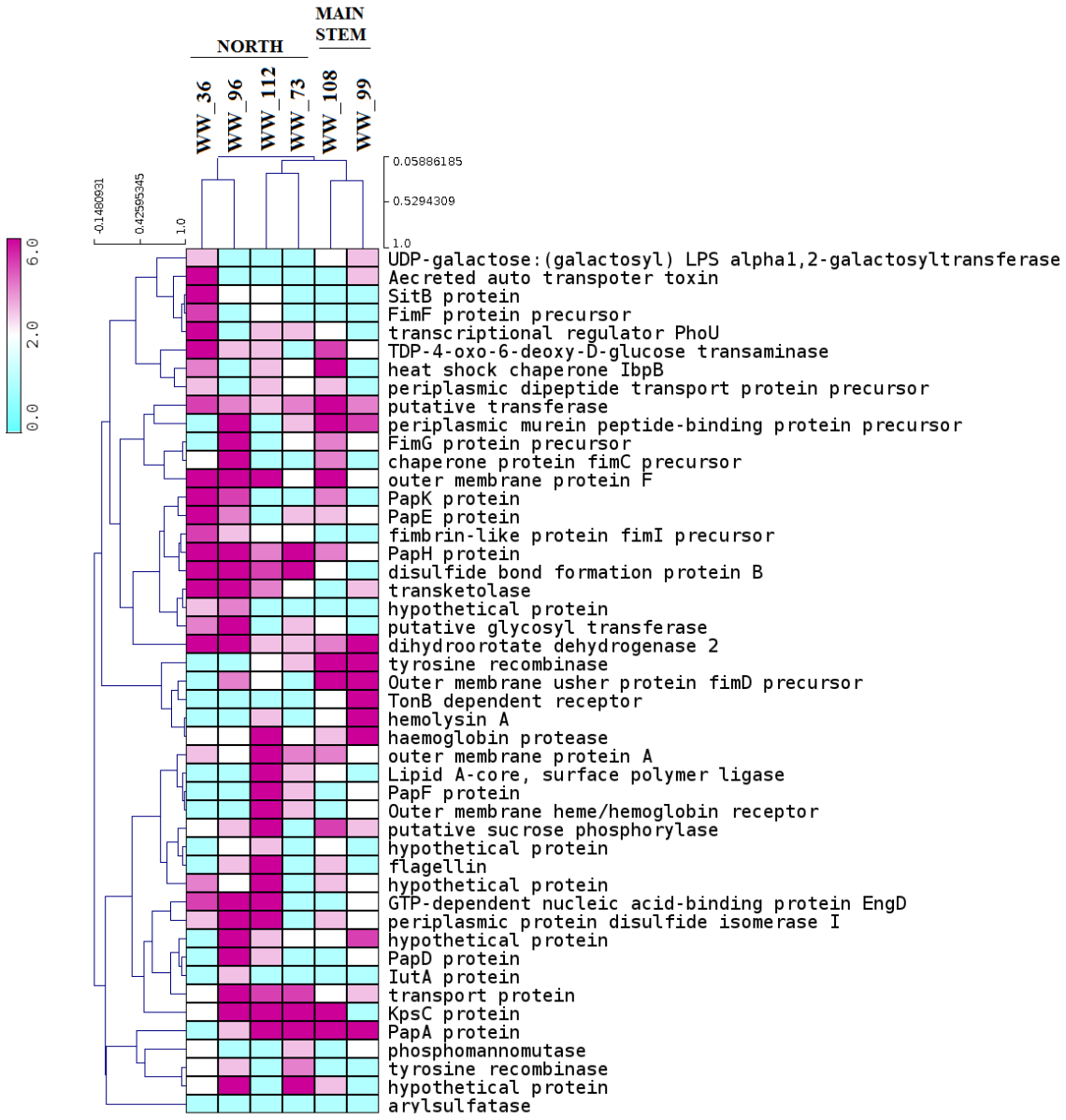

FIGURE 12 Virulence marker heat map for each metagenome from six selected locations closest to WRPs. Virulence genes are associated with *E. coli* in this analysis.

TABLE 5 Number of sequences per metagenome identified as having originated from at least one of 47 *E. coli* virulence markers.[a]

	Site 36	Site 96	Site 73	Site 112	Site 99	Site 108
No. Sequences	124	143	165	152	190	113

[a] Note: these abundances are not significantly different.

5 SUMMARY AND CONCLUSIONS

Based on the results described in the previous sections, we can summarize our results by drawing several conclusions.

First, microbial communities show a distinct distribution pattern across the 17 different sampling locations, and between sampling mediums (water, sediment, effluent, mixed liquor, sewage, etc.). For example, as expected, microbial communities in water are significantly different from those found in sediment. These communities appear to be stable (in their diversity and composition) between years (2013, 2014, and 2015) and between monthly sampling events. This suggests that the microbial communities within the CAWS are inherently stable and, therefore, perturbations that affect that stability will be easy to monitor. This provides us with a baseline for assessing ecosystem stability as a biomarker of system-scale changes in response to management practice change. Our analysis also shows that microorganisms associated with WRP effluent from secondary treatment can be tracked downstream, and typically show increased abundance in proximity to the secondary-treated final effluent location. These include human fecal indicators, such as Bifidobacterium and sewage contamination indicators, such as Acinetobacter. It will be interesting to determine whether this finding changes, and how, with the planned disinfection of the secondary treatment effluent in 2016 and subsequent years, and whether chlorination will provide different results than ultraviolet light disinfection. We also intend to explore the genotypic diversity, for particular indicator organisms. We hypothesize that taxa such as Bifidobacterium and Acinetobacter may still be present following disinfection, but will be represented by different genotypes. We will assess this using shotgun metagenomic genome reassembly, cultured isolate genotyping, and comparative genomic analysis.

Second, land-use types have a significant effect on CAWS water- and sediment-associated physiochemical properties. Road, residential, and open space significantly influenced water-associated properties; likewise, land-use types including commercial, institution, road, residential, and transport/utility significantly influenced sediment-associated properties. These properties will have concomitant influences on the microbial community structure. We will continue to analyze these patterns in 2016, and currently have no recommendations from this analysis.

Third, our analysis showed no significant differences in microbial alpha or beta diversity between wet/dry events. This will be further investigated; it may be caused by the lack of changes in the overall community structure. This is represented by no change in the membership of observed organisms but changes in the absolute abundances of key organisms such as fecal- and sewage contamination-indicator bacteria. To validate this hypothesis, qPCR assays are essential. Further analysis is scheduled for 2016.

Fourth, metagenome analysis revealed similar microbial community trends (alpha diversity and dominant taxa, e.g., Polynucleobacter) as observed with amplicon sequence analysis. Shotgun metagenome analysis also revealed low *E. coli* abundance and low *E. coli* associated virulence marker abundance at all sites, including those associated with final effluent

samples from O'Brien and Calumet WRPs. We will continue to assess these patterns through deeper metagenomic coverage of community patterns in the 2015 data.

Finally, from the results above we can conclude that our analytical methods are promising tools to understand microbial sources that can provide insightful information on the ecology of the CAWS. The work conducted in 2014 and 2015, which analyzed samples from 2013 through 2015, is a unique example of a descriptive baseline of the microbial ecology of a large scale riverine system such as the CAWS, which will enable us to determine important impacts of future management approaches. Many results of our analysis are still poorly understood because of the exploratory nature of this effort, and will be elucidated as more information will become available.

6 PROPOSED ACTIVITIES FOR 2016

To continue this study, we propose that work in the years 2016 to 2019 include the following tasks. Some of these tasks were part of the original work scope devised at the onset of the project, and others are recommended based on the experience and results obtained from the Phase I study period. Proposed activities include:

- Continuing the analysis of the Ambient Water Quality Monitoring program as planned to capture potential impacts of disinfection on the microbial communities in the CAWS.

- Incorporating microbial data, particularly the relative abundance of human fecal and sewage contamination indicators into the DuFlow model. This data will be normalized by alpha diversity to allow comparisons across multiple sites and sampling medium (influent sewage vs. secondary treated final effluent vs. CAWS water column samples), irrespective of the differences in the microbial diversity observed by site or sampling medium.

- Developing reliable markers or indicators for the identification of fecal contamination by animals including dogs, geese, and other wildlife of interest. We have existing databases, but are continuing to collect regionally relevant microbial samples to improve source apportionment for taxa in the CAWS. In addition, we will investigate methods such as oligotyping to differentiate between *E. coli* strains from different hosts. This is important because it will provide insight into the source of these strains, differentiating between human and animals. In addition, we will design qPCR assays to quantify the absolute abundances of these marker organisms, particularly during wet/dry events.

- Determining the nature of the contribution made by sediments to CAWS water during storm events. We are particularly interested in understanding sediment as an additional source of fecal and sewage contamination indicators during storm events due to sediment resuspension into the water column. We have developed a directional statistical approach to help interpret observed patterns across these events, which will be applied here.

- Determining the major environmental parameters that most closely describe the community variance using CCA. We also continue to analyze these similarities to confirm whether the overlap in influent sewage, secondary-treated final effluent, and water column samples can be influenced by heavy rainfall or storm events.

- Determining appropriate sequencing depth, and experimental approach for the reliable and accurate detection of *E. coli* using standard addition experiments. In addition, we are investigating multiple specialized databases to ensure *E. coli* detection and further resolve these short reads to strain-level resolution.

- Applying Source Tracker to data from all sampling years. For this, we are building new source databases that take into consideration the updated EMP data and all of the CAWS source data that we have generated so that the number of unknown sources is reduced.

- Completing the processing and data generation for the 2015 CAWS samples for metagenome sequencing and analysis. Samples representing different sampling media, sampling months, and wet/dry events will be selected.

PART 2—2013–2015 CHICAGO AREA WATERWAYS (CAWS) LAND USE AND LAND COVER ANALYSIS, AMBIENT WATER QUALITY, AND HYDRAULIC MODELING

Herbert Ssegane, Argonne National Laboratory

1 INTRODUCTION, OBJECTIVES, AND TASKS

The main objective of this study is to characterize microbial sources in Chicago Area Waterways (CAWS) and determine their spatial and temporal occurrence. Potential sources include effluent from water reclamation plants (WRPs), direct stormwater runoff, and combined sewer overflows (CSOs). Recognized as significant additional sources for micro-pollutants in surface waters, CSO events occur when stormwater runoff exceeds the capacity of the sewer network, resulting in the discharge of untreated wastewater into surface waters. This study uses synoptic sampling at pre-determined locations to collect water and sediment samples for microbial and metagenomics analysis based on pre-defined wet or dry weather conditions. Because of the intermittent nature of sampling along CAWS, the spatial and temporal occurrence of microbial pollution may not be fully captured. Also, flow and stage monitoring is not carried out at the same sampling sites along CAWS, yet hydraulic parameters (e.g., flow, stage, and velocity) may be critical drivers of microbial resuspension, growth, and die-off. Therefore, to extract flow, velocity, and stage data at each sampling site, hydraulic modeling is needed. Accordingly, hydraulic modeling enables the integration of microbial data into a modeling framework to provide analytical and predictive assays of spatial and temporal microbial occurrences.

This task had several objectives and activities:

1. Define and assess land use categories of interest based on the Chicago Metropolitan Agency for Planning (CMAP) land use inventory.

2. Assess sampling protocol to quantify frequency of wet and dry sampling events.

3. Characterize flow metrics at the microbial sampling sites with hydraulic modeling of CAWS. This objective was met with two activities.

    a. Data management and streamlining from multiple sources, such as the U.S. Geological Survey (USGS) and the Metropolitan Water Reclamation District of Greater Chicago (MWRD)
    b. Development of hydraulic model

2 MATERIALS AND METHODS

2.1 ASSESSMENT OF LAND USE AND LAND COVER DISTRIBUTION

Classification of land use and land cover (LULC) was based on major watersheds in Cook County and the drainage areas contributing to surface runoff at each sampling site. The drainage areas were created from a 10 ft.-digital elevation model (DEM) using the deterministic eight nodes (D8) algorithm as implemented in ArcGIS 10.2. The D8 algorithm estimates a specific catchment area by routing flow to only one dominant downhill direction. On relatively flat areas, however, flow may be proportionally distributed to multiple directions. The DEM was generated from 3 ft.-light detection and ranging (LIDAR) data for Cook County. Because of the sewer network, these drainage areas represent the potential source of surface runoff for each site because actual surface drainage areas are influenced by the sewer network. LULC classes at each site were assessed to quantify similarity of LULC distributions across sampling locations. This work was done to provide an overall assessment of potential differences in LULC that could be correlated with specific microbiome and analytical differences in the samples collected. A hypothesis is that surface runoff will be different based on predominant LULC. The base LULC data used in this analysis is the 2010 CMAP data, available at http://www.cmap.illinois.gov/data/land-use. The CMAP data was reclassified into 11 major categories: agricultural, commercial, construction, industrial, institution, non-parcel road, open space, residential, transport or utility, vacant, and water.

2.2 CHARACTERIZATION OF AMBIENT WATER QUALITY

Ambient water quality data for 2012 and 2013 was grouped into water chemistry and microbial indicator data. Water chemistry data included 15 parameters: temperature, acidity (pH), alkalinity, suspended solids (SS), total oxygen carbon (TOC), dissolved oxygen (DO), total dissolved solids (TDS), total phosphorus (TP), total Kjeldahl nitrogen (TKN), chlorophyll, chlorine (Cl), nitrates (NO_3), ammonia (NH_3), sulfates (SO_4), and fluoride (F). Water chemistry data was relatively complete, while the microbial indicators included two parameters of fecal coliform and *Escherichia coli* (*E. coli*). The objective for analyzing ambient water quality data was to identify sites for which ambient water quality (water chemistry or microbial indicators) were relatively similar based on distribution of the parameters in the two groups.

Sites with similar water quality were identified through classification using the *k-means* algorithm for cluster analysis (Hartigan and Wong 1979, Jain 2010) with no spatial constraint, so that similarity across sites was not restricted to proximity to either the Terrence J. O'Brien (North Branch) WRP or the Calumet WRP. This classification uses the geometric mean of the annual data. For 2013, water chemistry data with 15 parameters created a [12 x 15] data matrix of 12 sampling sites per each of the 15 parameters, thus 12 samples (n) each with 15 dimensions (d). This data structure is subject to the curse of dimensionality problem, $d > n$, for classification purposes (Keogh and Mueen 2010, Chen 2009, Pestov 2000), such that, classification results may not be statistically sound and reliable. The curse of dimensionality is more prevalent when the number of samples (n) is less than the number of variables for each sample (p). It is

considered a curse of dimensionality because the value of *n* needed to train an estimator grows exponentially as *p* increases for one to get a good classification or clustering. If it is ignored, samples may be considered to be close to each other yet they are far in the *d-space*. To reduce the dimensionality of the dataset, a correlation matrix was generated and one of the correlated parameters (Pearson correlation, which is $r \geq 0.7$ or $r \leq -0.7$) was excluded from subsequent classification. This process reduced 15 parameters to 10. The water chemistry parameters excluded from this classification included alkalinity, TOC, TKN, Cl, and SO_4 because they were all highly correlated to TDS and NH_3.

2.3 2013–2015 WEATHER CLASSIFICATION OF SITE SAMPLING DATES

Hourly rainfall data for the 25 rain gauges included in the Cook County Precipitation Network were acquired from the Illinois State Water Survey. Thiessen polygons were generated to determine the closest rain gauge for each sampling site. Weather classification was based on MWRD's definitions.

1. Dry weather (<0.1 inch precipitation): Dry weather is defined by antecedent dry conditions for 2 days following a 0.25–0.49 inch event, 4 days following a 0.50–0.99 inch event, and 6 days following a >1.0 inch event.

2. Wet weather:
   - Wet weather without CSOs (0.5–1.0 inch precipitation). Water sampling to occur within 10 hours of the end of the rain event.
   - Wet weather with CSOs, including 125th Street Pump Station (>1.5 inch precipitation). Water sampling to occur within 10 hours of the end of the rain event.

2.4 DUFLOW MODELING

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

Accordingly, the DuFlow network was modified to represent the new boundary conditions. Hydraulic modeling accounted for the system's major inflows and outflows. Major inflows are influenced by control structures, pumping stations, tributaries, and CSO discharges. The control structures include Wilmette Pumping Station, Chicago River Controlling Works, and T.J. O'Brien Lock and Dam. Pumping stations include the Racine Avenue Pumping Station and the 95th, 122nd, and 125th pumping stations. Major tributaries include the North Branch Chicago River (NBCR) and Little Calumet (LC), while minor tributaries include Tinley, Midlothian, Mill, Navajo, Natalie, and East and West Stony creeks.

Missing stage and flow data were calculated based on data from neighboring gauged tributaries and adjusted for the correlation strength between the target gauge and the neighboring gauged tributaries (Figure 14). Local temporal relationships were established using data recorded immediately before and after the missing data using multivariate adaptive regression splines (MARS) (Friedman 1991, Sánchez-Borrego 2011). MARS were implemented in R using the "Earth" package. Post-processing to minimize data anomalies was achieved by visual inspection of the calculated and filled-in data and the graphs.

Upstream and downstream boundary conditions included the North Shore Channel at Wilmette (05536101), Chicago River Main Stem at Columbus Drive (05536123), Calumet River at the T.J. O'Brien Lock and Dam (05536358), Racine Avenue Pump Station (RAPS), Little Calumet River at South Holland (05536290), and CSSC near Lemont (05536890). Refer to Figure 13b for the respective spatial locations. Stage (H) and discharge data (Q) were provided by MWRD, as was CSO event data. The approximately 105 unique CSO events in 2013 across CAWS were represented by a system of 44 discharge points. Flow at the locks was estimated by aggregating discharge hourly data due to navigation, blockages, leakages, and discretionary diversions. All stage data was referenced to the City of Chicago Datum of 579.48 ft.

2.5 MODEL VALIDATION

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

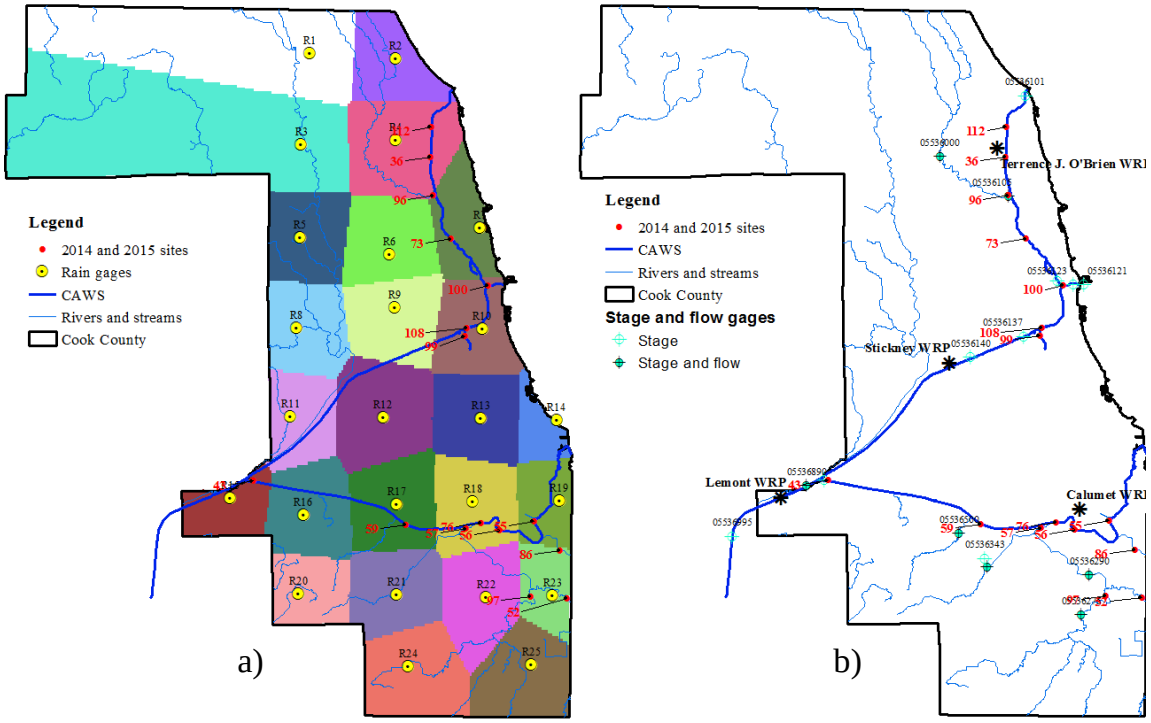

FIGURE 13 (a) Thiessen polygons showing the sampling sites and the closest rain gage and (b) location of USGS gages for stage and flow data.

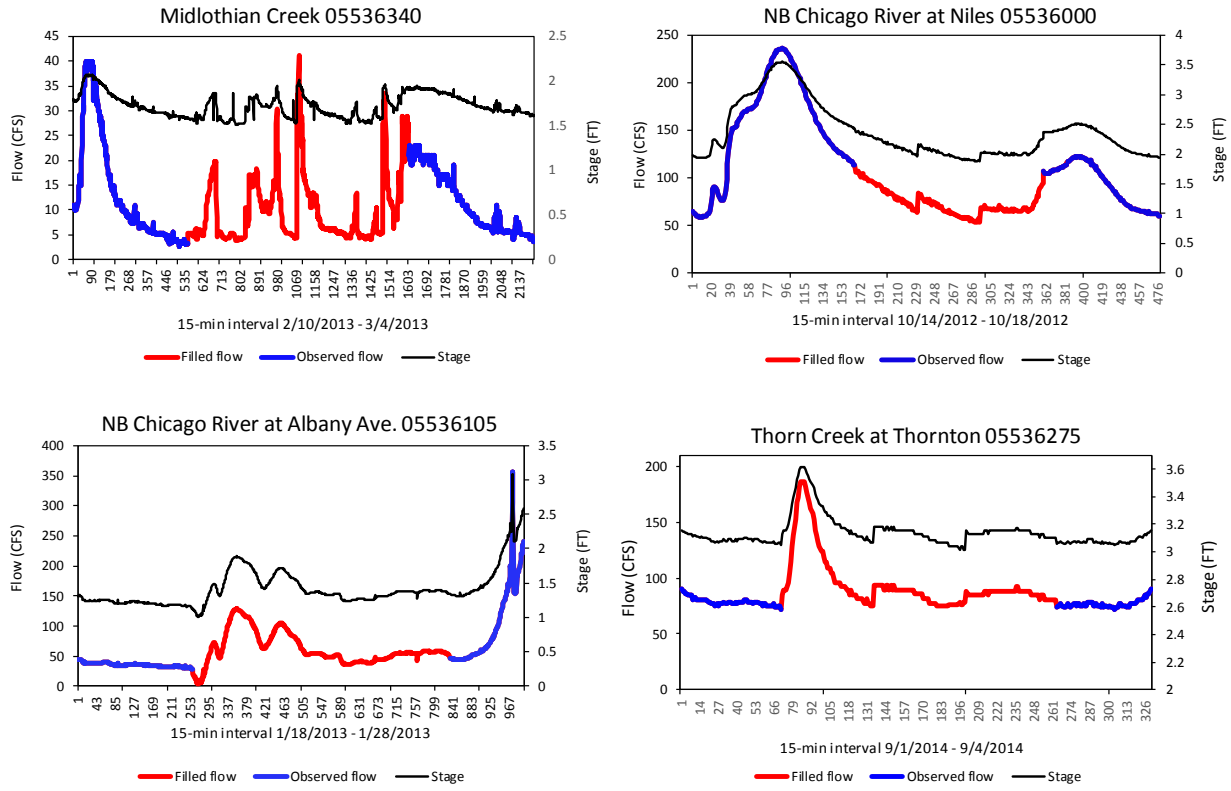

FIGURE 14 Examples of calculated and filled data at a 15-minute time interval for 2013–2014.

3 RESULTS

3.1 LAND USE AND LAND COVER DISTRIBUTION BY SAMPLING LOCATION

Figure 15 shows the delineated drainage area for each sampling location. The sampling sites at the outlets of the two tributaries of NBCR and LC (sites 96 and 57) have the largest drainage areas while the Chicago River main stem (site 100) has the smallest drainage area. The built environment dominates the land use at each site except site 86, which is dominated by open space with vegetation.

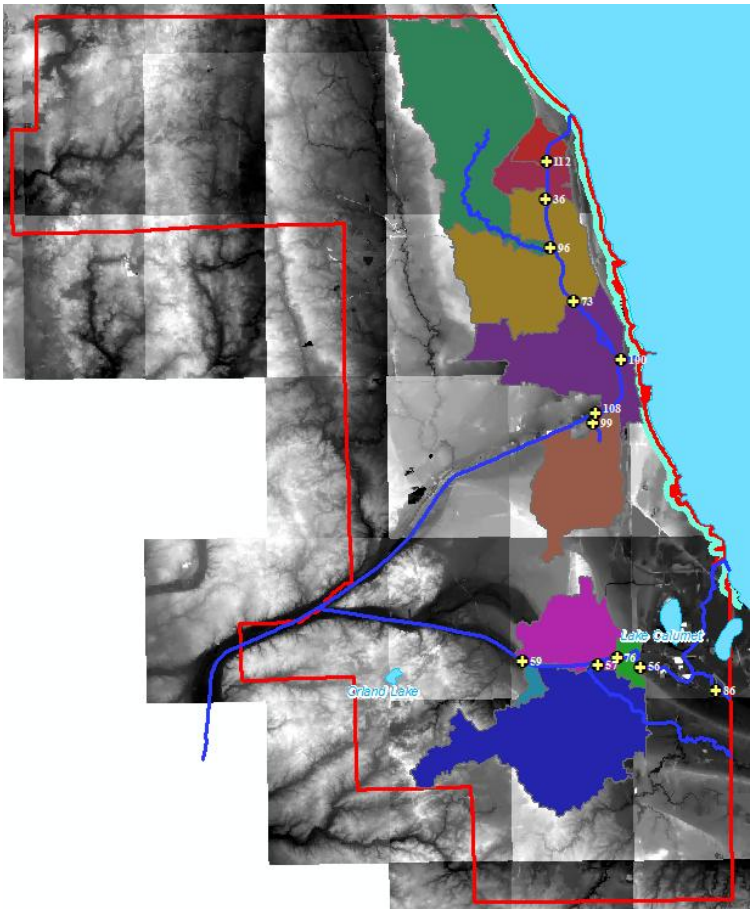

FIGURE 15 Boundaries of drainage areas contributing surface runoff at sampling sites in 2013.

3.2 INFLUENCE OF MAIN INFLOWS ON THE SPATIAL VARIATION OF WATER CHEMISTRY AND MICROBIAL DATA ALONG CAWS

Figure 16 shows sites that had a similar water chemistry for 2013, which was a wet year compared to 2012. Similar chemistries are given the same site color. Based on classification results, the water chemistry of the two major tributaries NBCR (site 96) and LC (site 57) are similar. Compared to other sites, they have the lowest water temperature and NH_3 and the highest levels of SS, pH, chlorophyll, and TDS. The water chemistry of the site along the Grand Calumet (site 86) and the site near the Racine Avenue Pumping Station (site 99) are similar with the lowest levels of DO, pH, SS, and the highest levels of NH_3. Irrespective of the water chemistry upstream of the WRPs, it changes after inflows from both WRPs and is similar downstream both WRPs. Therefore, inflows from the two tributaries NBCR and LC do not appear to dramatically influence the water chemistry. Effluent from the WRPs appears to influence downstream chemistry such that the chlorophyll levels were lowered to the minimum, while both NO_3 and TP increased to the highest levels. Sites 112 and 56 show lake water effects and have the highest temperature and lowest levels of TDS, NO_3, TP, and Fl.

Figure 17 shows which sites have similar microbial indicators. Site 36, which is close to the T.J. O'Brien WRP had the highest loadings of both fecal coliform and *E. coli*. The microbial indicators at this site were significantly higher than at all other sites. Analysis of samples post-disinfection will be compared to the results to-date to determine the impact of disinfection on water quality. As with the water chemistry at sites 112 and 56, the corresponding microbial indicators of both sites were similar. Sites near the lake (112 and 56), NBCR (96), and CRCW (100) had the lowest levels of microbial indicators. Site 73 had relatively lower levels compared to site 36, which is indicative of dilution effects of inflows from NBCR and die-off along the river. Future work will allow us to model how far downstream there could be detectable levels of fecal indicators based on flow conditions, temperature, and other parameters. Microbial indicators of inflow from LC and effluent from Calumet WRP are relatively lower than microbial loadings at downstream site 59. This observation is indicative of additional microbial inputs possibly from surface runoff or CSOs.

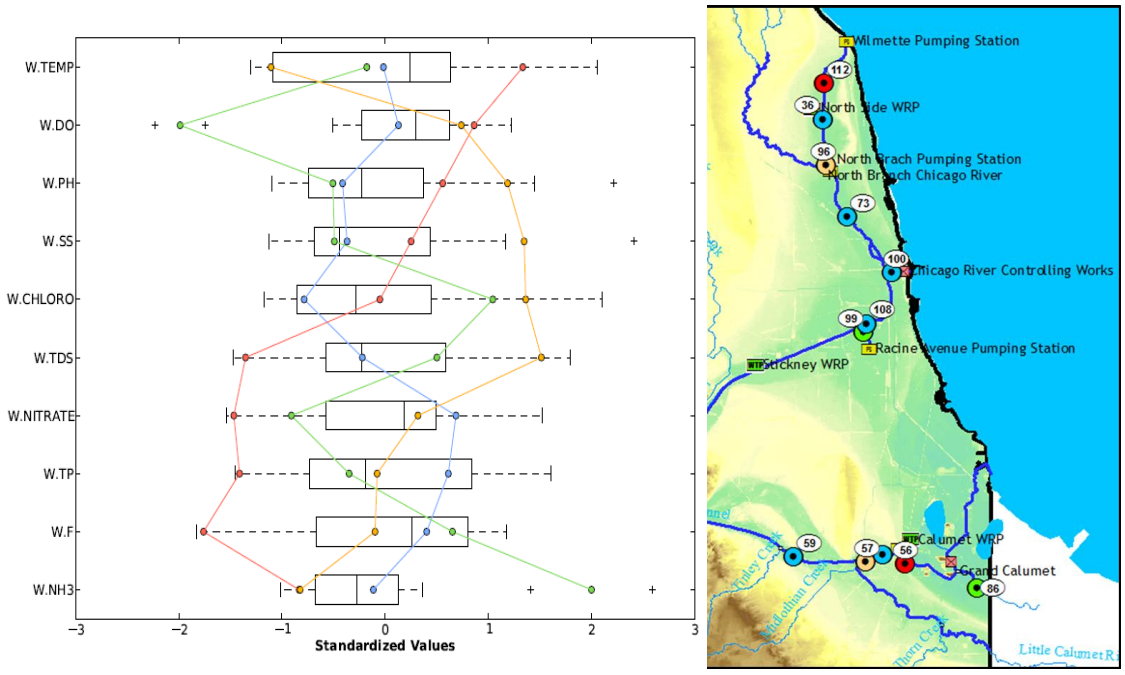

FIGURE 16 Classification results for sites where water chemistry is relatively similar. The left figure shows standardized values of water chemistry parameters and corresponding trends that define the similarity for each group. Sites with the same color indicate similar water chemistry. The line colors, left, and point colors, right, match the classified groups.

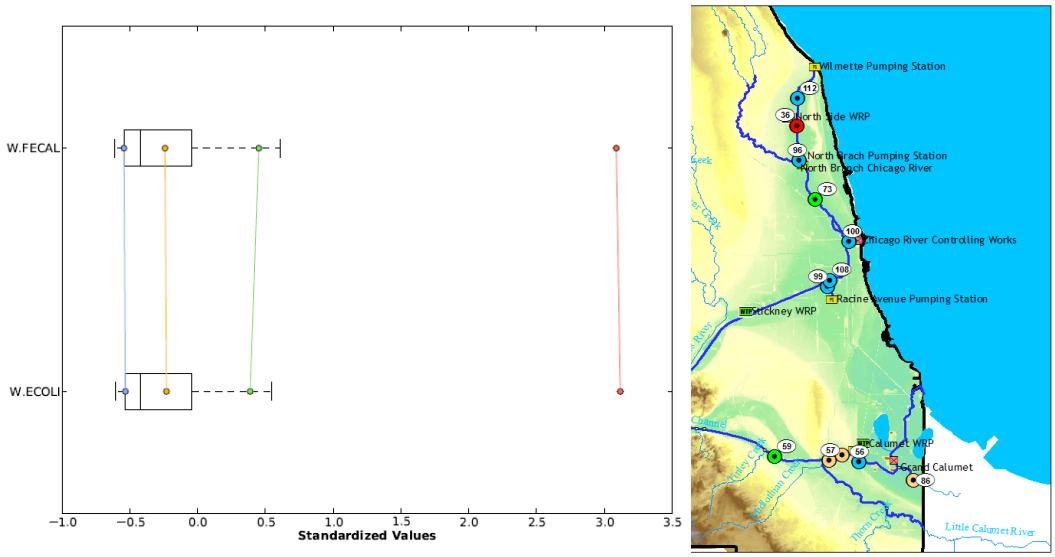

FIGURE 17 Classification results for sites where water microbial indicators are relatively similar. Sites with the same color indicate similar microbial indicators. The left figure shows standardized values and corresponding trends that define the similarity for each group. The line colors, left, and point colors, right, match the classified groups.

3.2.1 Site Incidences of Dry and Wet Sampling Events

Table 6 shows the number of sampling events classified by wet and dry weather conditions for each site. Refer to Appendix A for detailed classification for each sampling date and site. There were more wet sampling events in 2014 for sites in the Calumet region. However, the wet and dry sampling events were evenly split in 2015. There were no samples collected in 2013 for sites 43, 52, 55, and 97, which were added in 2014. Refer to Section 2.3 for weather classification categories 0–10.

TABLE 6 Summary of weather classification on sampling dates.

	2013		2014		2015[a]	
Site	WET	DRY	WET	DRY	WET	DRY
112	4	3	4	1	3	3
108	3	4	0	4	3	4
100	3	4	0	4	3	4
99	3	4	0	2	3	4
96	3	4	1	4	2	4
86	3	4	10	2	8	6
76	3	4	9	2	8	6
73	3	4	1	4	2	5
59	3	4	8	3	7	7
57	3	4	8	4	8	6
56	2	3	8	2	8	6
36	4	3	4	1	3	4
43	---	---	6	4	4	3
52	---	---	10	2	5	2
55	---	---	9	2	6	1
97	---	---	10	2	5	2

[a] The 2015 data summary consists of sampling dates before October.

3.3 2013 AND 2014 COMBINED SEWER OVERFLOW (CSO) EVENTS

CSO events for 2013 and 2014 were analyzed to guide the selection of representative CSOs in the hydraulic model because inclusion of all CSOs does not guarantee model accuracy and is computationally expensive. Analysis focused on the representative CSOs in the Calumet, Stickney, and O'Brien (North Shore) regions for both years with emphasis on the frequency of events (number of events per year by a single CSO location). This analysis excluded data in the Des Plaines watershed. Refer to Figure 18 for the spatial distribution of CSO locations and the corresponding 2013 number of events. Figure 18 illustrates that more CSO events occurred around the Stickney and O'Brien regions than the Calumet region. Table 7 summarizes 2013 and 2014 CSO events.

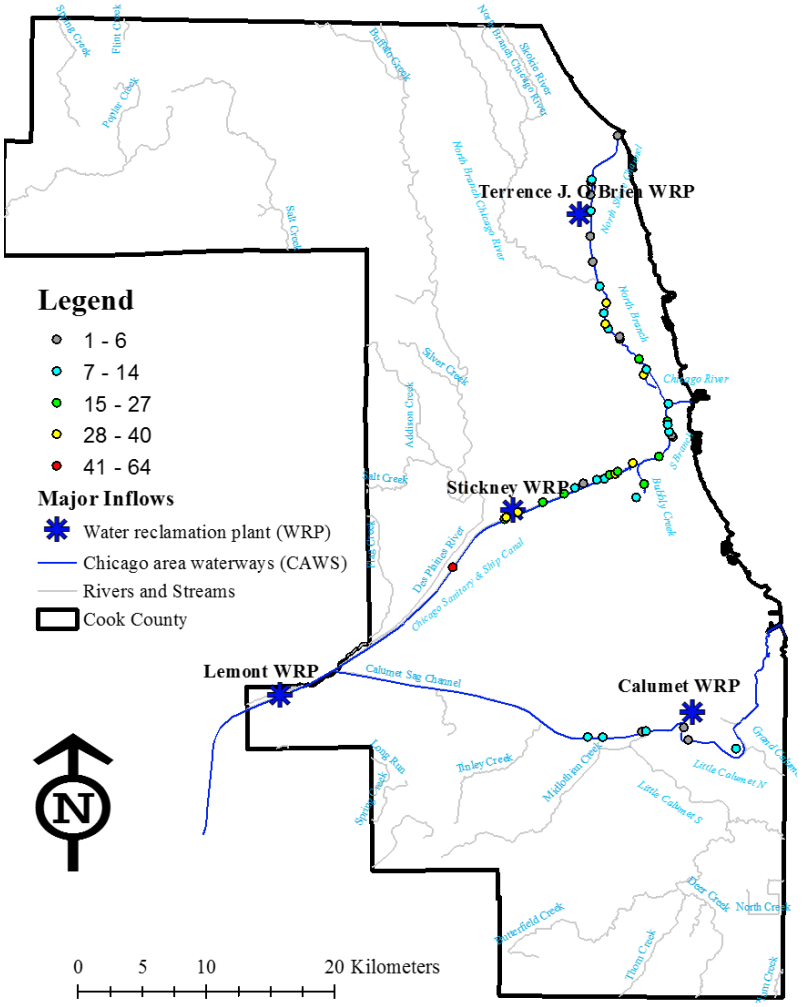

FIGURE 18 CSO event frequency for 2013.

TABLE 7 2013 and 2014 CSO event summary.

Parameters	2013	2014	Comment (2013/2014)
Unique CSO discharge points	105	63	Excluding Des Plaines
Max # of events	64	48	DS-M79 / CD-S21
Max event duration (hrs)	40.7	20.5	CD-S39 / CD-S2 & CD-S4
Minimum event duration (min)	5	2	CD-S43 / CD-S43

3.4 VALIDATION OF DUFLOW STREAMFLOW SIMULATIONS

Figure 19 visually compares streamflow observations (USGS data) and DuFlow simulations for a gage on the CSSC near Lemont, IL for 2013. The graphs and model performance metrics (R^2=0.73, NSE=0.72, and PBIAS=-6.0%) are indicative of DuFlow's ability

to capture both the magnitude and sequence of flows at this gaged station. A coefficient of determination of 0.73 (R^2=0.73) indicates that the model captured 73% of the hourly streamflow variability. A PBIAS of -6.0% indicates that, on average, the model over-estimates hourly streamflow. However, a PBIAS within ±10% is indicative of an excellent model (Moriasi 2007). Figure 20 illustrates the accumulation of flow along CAWS on a day with confirmed CSO events throughout the stream network.

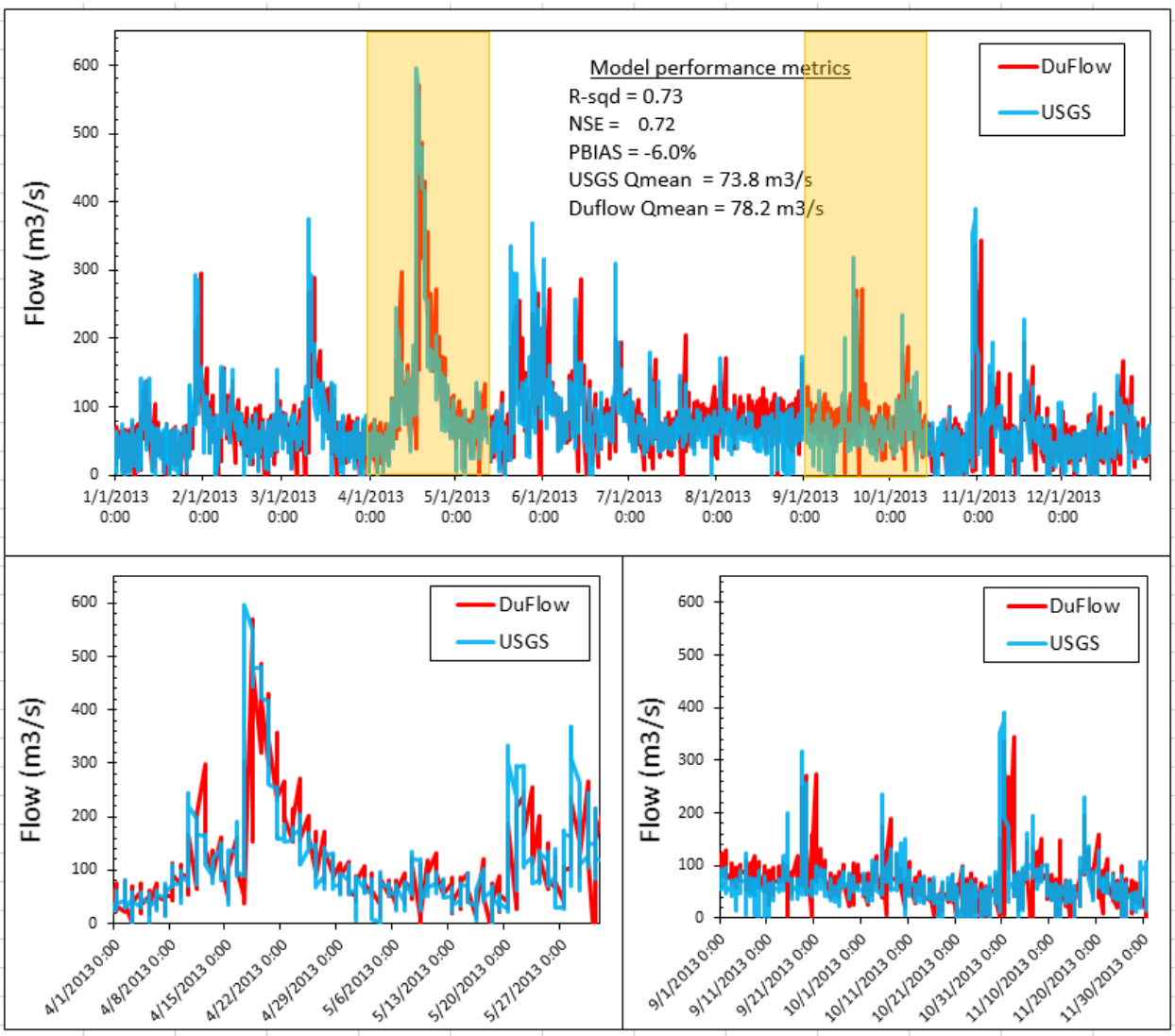

FIGURE 19 Observed (USGS) and simulated (DuFlow) streamflow on CSSC near Lemont (USGS gaging station 05536890). The simulation results are at an hourly time step. The bottom two graphs are magnifications of the two shaded regions in the top graph.

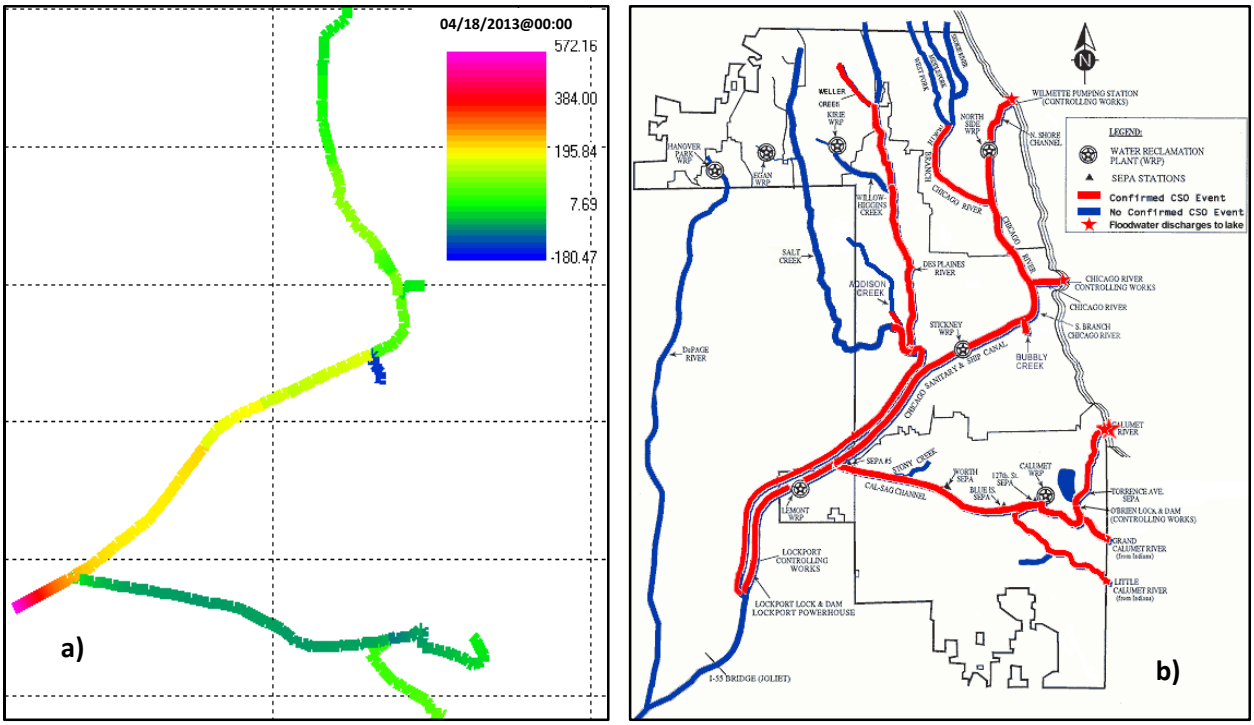

FIGURE 20 A snapshot of spatial difference in streamflow (in m^3/s) (a) along CAWS on April 18, 2013 at midnight and (b) CAWS sections with confirmed CSO events by MWRD.

3.4.1 Streamflow Simulations at Sampling Sites

Appendix A1 contains the average and maximum streamflow for each site for the 2013 sampling dates. The values were based on hourly outputs from DuFlow. Other hydraulic parameters such as stage and velocity are not presented. They will only be considered depending on their conditional relevance in the presence of streamflow for microbial predictive analytics.

4 ACTIVITIES PLANNED FOR 2016

4.1 DUFLOW MODELING AND PREDICTIVE ANALYTICS

Current and planned activities include the development of the interface between the 2013 model data and a predictive neural network-based model that incorporates the data on selected microbial genera with hydraulic data to obtain a preliminary, prototype microbial predictive model for forecasting and analysis of alternative management scenarios. Once that is developed and tested, planned activities include the expansion of the entire predictive model to also include the years 2014-2019.

4.2 THREE-STEP APPROACH TO INCORPORATE MICROBIAL AND HYDRAULIC DATA

1. Concurrent hydraulic and water quality modeling, DNA extraction, quantitative PCR assays and quantification
    a. This work will continue to generate microbial community data for the years 2016-2019

2. Extraction of hydrologic and water quality parameters at sampling points as explanatory variables for predictive microbial model
    a. For the year 2013, data on flow, stage and CSO will be extracted from the DuFlow model to reflect likely conditions at the ambient water quality monitoring sampling points throughout the CAWS
    b. In future years, this will be repeated for the subsequent years when the model will be available for those years.

3. Integration of microbial predictive models into Hydraulics and Hydrology for forecasting and analysis of alternative management scenarios
    a. To integrate the microbial data with the hydraulic data a neural network model will be generated which will use the variables shown in Figure 21 as possible explanatory variables for microbial variances in the probabilistic model.

Figure 21 illustrates the proposed conceptual framework for generating the conditional probabilities between specific microbial abundance and environmental variables, and Figure 22 illustrates the artificial neural network to capture the interactions between microbial taxa and their environment. This modeling approach is stochastic, meaning it accounts for uncertainty in the interactions, and assumes that microbial community patterns share mathematically describable relationships with environmental conditions.

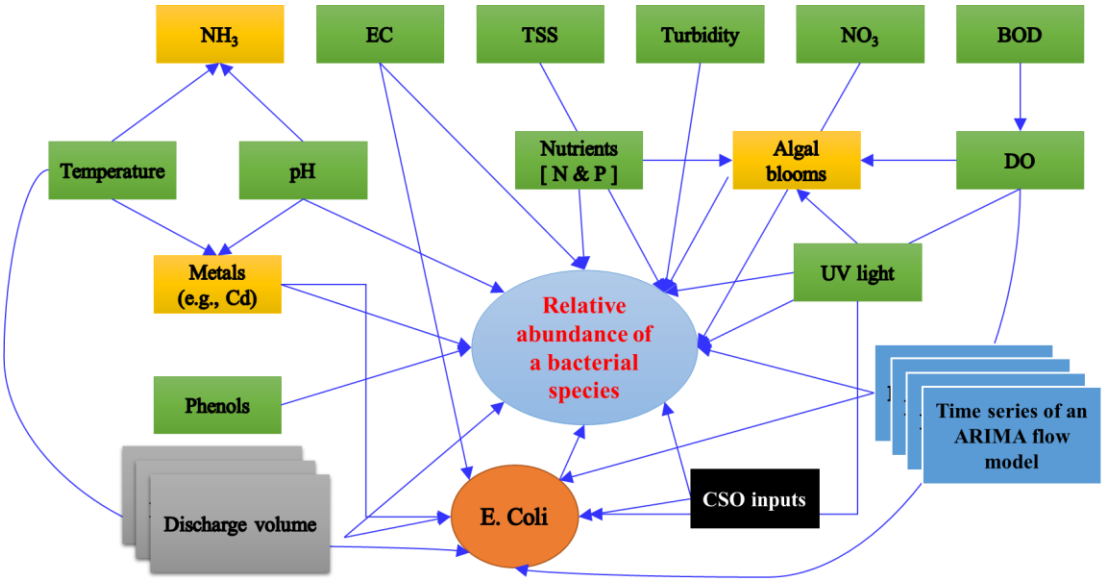

FIGURE 21 Conceptual Bayesian network for generating probabilities between specific microbial abundance and environmental variables.

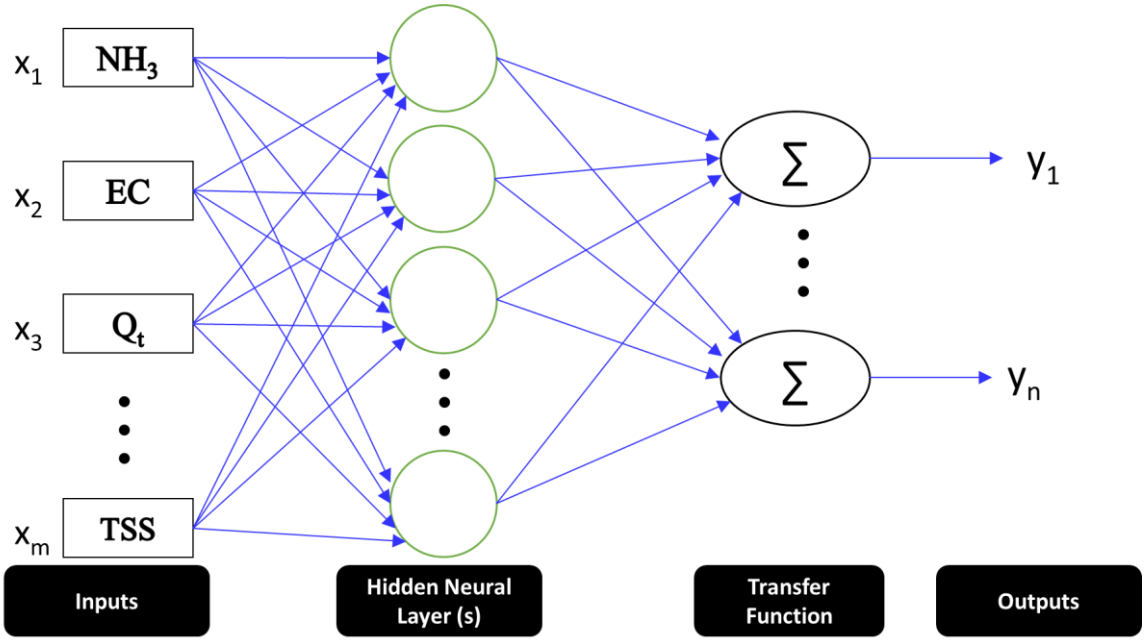

FIGURE 22 Microbial Assemblage Predictive neural network structure.

5 REFERENCES

I. Sánchez-Borrego, M. del Mar Rueda and J.F. Munoz, “Imputation and Inference with Multivariate Adaptive Regression Splines.” In *Modern Mathematical Tools and Techniques in Capturing Complexity*, 341–353. Springer, 2011.

APPENDICES

APPENDIX A1

TABLE S1 2013 sampling dates, weather classification, average and maximum flow (Q_{avg} and Q_{max}) at monitoring sites.

Site	Date	MWRD weather classification	Q_{avg} (m^3/s)	Q_{max} (m^3/s)	Site	Date	MWRD weather classification	Q_{avg} (m^3/s)	Q_{max} (m^3/s)
	4/8/2013	WET	-0.268	0.206		4/22/2013	DRY	0.185	0.266
	5/13/2013	DRY	0.035	0.659		5/28/2013	WET	0.677	4.644
	6/10/2013	WET	-0.020	0.656		6/24/2013	WET	0.001	0.019
112	7/15/2013	DRY	0.033	0.714	**86**	7/29/2013	DRY	0.002	0.010
	8/12/2013	WET	-0.068	0.345		8/26/2013	DRY	0.000	0.004
	9/16/2013	WET	0.128	1.913		9/30/2013	WET	0.153	0.212
	10/14/2013	DRY	0.041	0.755		10/28/2013	DRY	0.170	0.425
	4/15/2013	WET	15.0	23.4		4/22/2013	DRY	27.8	29.6
	5/20/2013	WET	25.8	89.9		5/28/2013	WET	20.9	31.7
	6/17/2013	DRY	19.6	36.5		6/24/2013	WET	14.8	21.6
108	7/22/2013	WET	29.1	39.2	**76**	7/29/2013	DRY	28.6	31.3
	8/19/2013	DRY	28.3	33.9		8/26/2013	DRY	25.0	28.1
	9/23/2013	DRY	28.0	36.5		9/30/2013	WET	18.2	22.5
	10/21/2013	DRY	6.1	16.8		10/28/2013	DRY	7.3	16.6
	4/15/2013	WET	16.3	22.1		4/8/2013	WET	17.7	20.6
	5/20/2013	WET	14.8	31.2		5/13/2013	DRY	9.4	11.8
	6/17/2013	DRY	19.6	32.8		6/10/2013	WET	13.1	16.3
100	7/22/2013	WET	28.6	39.5	**73**	7/15/2013	DRY	10.3	13.1
	8/19/2013	DRY	28.0	34.5		8/12/2013	DRY	10.0	12.3
	9/23/2013	DRY	27.2	33.2		9/16/2013	WET	11.7	17.1
	10/21/2013	DRY	7.1	17.8		10/14/2013	DRY	8.6	11.1
	4/15/2013	WET	0.007	0.127		4/22/2013	DRY	77.5	92.1
	5/20/2013	WET	-9.036	0.083		5/28/2013	WET	36.3	51.5
	6/17/2013	DRY	0.008	0.157		6/24/2013	WET	18.0	26.2
99	7/22/2013	WET	-0.004	0.042	**59**	7/29/2013	DRY	29.6	35.3
	8/19/2013	DRY	-0.005	0.049		8/26/2013	DRY	25.8	31.8
	9/23/2013	DRY	-0.006	0.047		9/30/2013	WET	19.7	27.1
	10/21/2013	DRY	0.009	0.096		10/28/2013	DRY	7.6	23.4
	4/8/2013	WET	5.3	6.4		4/22/2013	DRY	39.1	52.6
	5/13/2013	DRY	1.4	1.4		5/28/2013	WET	10.4	16.1
	6/10/2013	WET	2.3	2.5		6/24/2013	WET	2.5	2.7
96	7/15/2013	DRY	1.3	1.6	**57**	7/29/2013	DRY	0.7	0.9
	8/12/2013	DRY	1.4	2.3		8/26/2013	DRY	0.7	1.0
	9/16/2013	WET	3.7	6.3		9/30/2013	WET	1.2	1.3
	10/14/2013	DRY	1.0	1.1		10/28/2013	DRY	0.7	1.2

Site	Date	MWRD weather classification	Q_{avg} (m^3/s)	Q_{max} (m^3/s)	Site	Date	MWRD weather classification	Q_{avg} (m^3/s)	Q_{max} (m^3/s)
	4/8/2013	WET	12.7	13.5		6/24/2013	WET	4.9	12.5
	5/13/2013	DRY	8.3	9.3		7/29/2013	DRY	21.0	23.3
	6/10/2013	WET	10.9	12.1		8/26/2013	DRY	18.4	21.0
36	7/15/2013	DRY	9.1	10.1	**56**	9/30/2013	WET	11.3	14.9
	8/12/2013	WET	8.9	9.6		10/28/2013	DRY	0.6	11.1
	9/16/2013	WET	8.0	10.8					
	10/14/2013	DRY	7.5	8.6					

APPENDIX A2

TABLE S2 2014 sampling dates and weather classification at monitoring sites.

Site	Date	MWRD weather classification	Site	Date	MWRD weather classification
112	4/14/2014	WET	**43**	5/21/2014	WET
	5/12/2014	WET		5/27/2014	DRY
	6/9/2014	WET		6/23/2014	WET
	7/14/2014	WET		7/1/2014	WET
	11/10/2014	DRY		7/22/2014	DRY
108	5/19/2014	DRY		7/28/2014	DRY
	7/21/2014	DRY		8/5/2014	WET
	8/18/2014	DRY		8/22/2014	WET
	11/17/2014	DRY		8/25/2014	WET
100	5/19/2014	DRY		11/24/2014	DRY
	7/21/2014	DRY	**52**	4/15/2014	WET
	8/18/2014	DRY		4/28/2014	WET
	11/17/2014	DRY		5/21/2014	WET
99	8/18/2014	DRY		5/27/2014	WET
	11/17/2014	DRY		6/23/2014	WET
96	4/14/2014	WET		7/1/2014	WET
	5/12/2014	DRY		7/22/2014	WET
	6/9/2014	DRY		7/28/2014	DRY
	7/14/2014	DRY		8/5/2014	WET
	11/10/2014	DRY		8/22/2014	WET
86	4/15/2014	WET		8/25/2014	WET
	4/28/2014	WET		11/24/2014	DRY
	5/21/2014	WET	**55**	4/15/2014	WET
	5/27/2014	WET		4/28/2014	WET
	6/23/2014	WET		5/21/2014	WET
	7/1/2014	WET		5/27/2014	WET
	7/22/2014	WET		6/23/2014	WET
	7/28/2014	DRY		7/1/2014	WET
	8/5/2014	WET		7/22/2014	WET
	8/25/2014	WET		7/28/2014	DRY
	8/22/2014	WET		8/5/2014	WET
	11/24/2014	DRY		8/25/2014	WET
				11/24/2014	DRY

Site	Date	MWRD weather classification
76	4/28/2014	WET
	5/21/2014	WET
	5/27/2014	WET
	6/23/2014	WET
	7/1/2014	WET
	7/22/2014	WET
	7/28/2014	DRY
	8/5/2014	WET
	8/25/2014	WET
	8/22/2014	WET
	11/24/2014	DRY
73	4/14/2014	WET
	5/12/2014	DRY
	6/9/2014	DRY
	7/14/2014	DRY
	11/10/2014	DRY
59	4/28/2014	WET
	5/21/2014	WET
	5/27/2014	WET
	6/23/2014	DRY
	7/1/2014	WET
	7/22/2014	WET
	7/28/2014	DRY
	8/5/2014	WET
	8/25/2014	WET
	8/22/2014	WET
	11/24/2014	DRY
57	4/15/2014	WET
	4/28/2014	WET
	5/21/2014	WET
	5/27/2014	WET
	6/23/2014	WET
	7/1/2014	WET
	7/22/2014	WET
	7/28/2014	DRY
	8/5/2014	WET
	8/25/2014	WET
	8/22/2014	WET
	11/24/2014	DRY
97	4/15/2014	WET
	4/28/2014	WET
	5/21/2014	WET
	5/27/2014	WET
	6/23/2014	WET
	7/1/2014	WET
	7/22/2014	WET
	7/28/2014	DRY
	8/5/2014	WET
	8/22/2014	WET
	8/25/2014	WET
	11/24/2014	DRY
56	4/28/2014	WET
	5/21/2014	WET
	6/23/2014	WET
	7/1/2014	WET
	7/22/2014	WET
	7/28/2014	DRY
	8/5/2014	WET
	8/25/2014	WET
	8/22/2014	WET
	11/24/2014	DRY
36	4/14/2014	WET
	5/12/2014	WET
	6/9/2014	WET
	7/14/2014	WET
	11/10/2014	DRY

APPENDIX A3

TABLE S3 2015 sampling dates and weather classification at monitoring sites.

Site	Date	MWRD weather classification	Site	Date	MWRD weather classification
112	4/13/2015	WET	**43**	4/10/2015	WET
	5/11/2015	DRY		5/21/2015	DRY
	6/8/2015	WET		6/11/2015	DRY
	7/13/2015	WET		6/16/2015	WET
	8/10/2015	DRY		7/14/2015	WET
	9/14/2015	DRY		7/17/2015	WET
	10/12/2015			8/14/2015	DRY
	11/9/2015			4/10/2015	WET
108	3/16/2015	DRY	**52**	5/21/2015	DRY
	4/20/2015	WET		6/11/2015	WET
	5/18/2015	DRY		6/16/2015	WET
	6/15/2015	WET		7/14/2015	WET
	7/20/2015	DRY		7/17/2015	WET
	8/17/2015	WET		8/14/2015	DRY
	9/21/2015	DRY		4/10/2015	WET
	10/19/2015			5/21/2015	DRY
	11/17/2015			6/11/2015	WET
100	3/16/2015	DRY	**55**	6/16/2015	WET
	4/20/2015	WET		7/14/2015	WET
	5/18/2015	DRY		7/17/2015	WET
	6/15/2015	WET		8/14/2015	WET
	7/20/2015	DRY		4/10/2015	WET
	8/17/2015	WET		5/21/2015	DRY
	9/21/2015	DRY		6/11/2015	WET
	10/19/2015		**97**	6/16/2015	WET
	11/17/2015			7/14/2015	WET
99	3/16/2015	DRY		7/17/2015	WET
	4/20/2015	WET		8/14/2015	DRY
	5/18/2015	DRY		3/23/2015	WET
	6/15/2015	WET		4/10/2015	WET
	7/20/2015	DRY		4/27/2015	DRY
	8/17/2015	WET		5/21/2015	DRY
	9/21/2015	DRY		5/26/2015	WET
	11/17/2015			6/11/2015	WET
96	4/13/2015	DRY	**56**	6/16/2015	WET
	5/11/2015	WET		6/22/2015	DRY
	6/8/2015	WET		7/14/2015	WET
	7/13/2015	DRY		7/17/2015	WET
	8/10/2015	DRY		7/27/2015	DRY
	9/14/2015	DRY		8/14/2015	WET
	10/12/2015			8/24/2015	DRY
	11/9/2015			9/28/2015	DRY
				10/26/2015	
				11/23/2015	

Site	Date	MWRD weather classification	Site	Date	MWRD weather classification
86	3/23/2015	WET	**59**	3/23/2015	WET
	4/10/2015	WET		4/10/2015	WET
	4/27/2015	DRY		4/27/2015	DRY
	5/21/2015	DRY		5/21/2015	DRY
	5/26/2015	WET		5/26/2015	WET
	6/11/2015	WET		6/11/2015	WET
	6/16/2015	WET		6/16/2015	WET
	6/22/2015	DRY		6/22/2015	DRY
	7/14/2015	WET		7/14/2015	WET
	7/17/2015	WET		7/17/2015	WET
	7/27/2015	DRY		7/27/2015	DRY
	8/14/2015	DRY		8/14/2015	DRY
	8/24/2015	WET		8/24/2015	DRY
	9/28/2015	DRY		9/28/2015	DRY
	10/26/2015			10/26/2015	
	11/23/2015			11/23/2015	
76	3/23/2015	WET	**57**	3/23/2015	WET
	4/10/2015	WET		4/10/2015	WET
	4/27/2015	DRY		4/27/2015	DRY
	5/21/2015	DRY		5/21/2015	DRY
	5/26/2015	WET		5/26/2015	WET
	6/11/2015	WET		6/11/2015	WET
	6/16/2015	WET		6/16/2015	WET
	6/22/2015	DRY		6/22/2015	DRY
	7/14/2015	WET		7/14/2015	WET
	7/17/2015	WET		7/17/2015	WET
	7/27/2015	DRY		7/27/2015	DRY
	8/14/2015	WET		8/14/2015	WET
	8/24/2015	DRY		8/24/2015	DRY
	9/28/2015	DRY		9/28/2015	DRY
	10/26/2015			10/26/2015	
	11/23/2015			11/23/2015	
73	3/9/2015	DRY	**36**	3/9/2015	DRY
	4/13/2015	DRY		4/13/2015	WET
	5/11/2015	WET		5/11/2015	DRY
	6/8/2015	WET		6/8/2015	WET
	7/13/2015	DRY		7/13/2015	WET
	8/10/2015	DRY		8/10/2015	DRY
	9/14/2015	DRY		9/14/2015	DRY
	10/12/2015			10/12/2015	
	11/9/2015			11/9/2015	

SUPPORTING INFORMATION

Energy Systems Division
9700 South Cass Avenue, Bldg. 362
Argonne, IL 60439-4854

www.anl.gov