ABSTRACT

Title of Dissertation:	MONITORING AND PREDICTING THE MICROBIAL WATER QUALITY IN IRRIGATION PONDS
	Matthew Daniel Stocker Doctor of Philosophy, 2022
Dissertation directed by:	Professor Robert L. Hill, Department of Environmental Science and Technology

Small- to medium-sized farm ponds are a popular source of irrigation water and provide a substantial volume of water for crop growth in the United States. The microbial quality of irrigation waters is assessed by measuring concentrations of the fecal indicator bacteria *Escherichia coli* (*E. coli*). Minimal guidance currently exists on the use of surface irrigation waters to minimize consumer health risks. The overall objective of this work was to provide science-based guidance for microbial water quality monitoring of irrigation ponds. Spatial and temporal patterns of *E. coli* were evaluated in two Maryland irrigation ponds over three years of observations. Patterns of *E. coli* were stable over the three years and found to be significantly correlated to patterns of water parameters such as temperature, dissolved oxygen, turbidity, and pH. The EPA Environmental Fluid Dynamics Code model was used to evaluate the spatial 3D heterogeneity of *E. coli* concentrations within the ponds. Significant

differences in *E. coli* concentrations by sampling depth were found. Spatial heterogeneity of *E. coli* within the pond also resulted in substantial temporal variation at the irrigation pump, which was dependent on the intake location. Diurnal variation of *E. coli* concentrations was assessed for three farm ponds. *E. coli* concentrations declined from 9:00 to 15:00 for each pond, but statistically significant declines were only observed in two of the three ponds. Dissolved oxygen, pH, and electrical conductance were found to be the most influential environmental variables affecting *E. coli* concentrations. To better describe the relationships between *E. coli* and the environmental variables, four machine learning algorithms were used to estimate *E. coli* concentrations using water quality parameters as predictors. The random forest algorithm provided the highest predictive accuracy with $R^2 = 0.750$ and $R^2 = 0.745$ for Ponds 1 and 2, respectively, in the multi-year dataset containing 12 predictors. Temperature, electrical conductance, and organic matter content were identified as the most influential predictors. It is anticipated that the recommendations contained in this dissertation will be used to improve microbial monitoring strategies and protect public health.

MONITORING AND PREDICTING THE MICROBIAL WATER QUALITY IN IRRIGATION PONDS

by

Matthew Daniel Stocker

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Robert L. Hill, Chair
Professor Allen Davis
Dr. Yakov Pachepsky
Dr. Adel Shirmohammadi
Dr. Stephanie Yarwood

Dedication

This work is dedicated to my family and above all else, my Mother, Joyce Stocker who provided my sister and I with an incredible childhood despite being a single parent. Her constant support has been crucial in my life and is simply immeasurable.

Acknowledgements

The works presented in this dissertation were only possible through the combined assistance of all mentors, supervisors, advisors, coworkers, interns, farm owners and operators. In a successful confluence of efforts, I was able to perform my research and improve the ways in which we assess the microbial quality of irrigation waters.

Within this confluence, I firstly must thank my USDA-ARS supervisor, Dr. Yakov Pachepsky. In 2013, Yakov took a chance by hiring someone with little previous experience within the field of microbial water quality. I am grateful that such an accomplished scientist took interest in me and helped to guide me throughout my graduate career. Yakov provided constant financial support throughout my Masters and PhD studies for which I am truly grateful.

I would like to secondly acknowledge the assistance of my academic advisor, Dr. Robert Hill, who I first encountered during undergraduate coursework in 2008. Little did we know then we would start a collaboration that still continues today. Dr. Hill has provided countless hours of advice and input regarding my research, coursework, and navigating my graduate program.

Appreciation is extended to my Program of Study Committee, including Doctors Allen Davis, Adel Shirmohammadi, and Stephanie Yarwood for the time and effort they have placed into seeing me advance as a doctoral research scientist. Thank you for your continued support and also for creating challenging comprehensive exam questions designed to enhance my critical thinking skills and broadening the way I view water quality issues. A special thank you to Dr. Yarwood who taught me about soil microbiology in my first class as a graduate student in Fall 2014 and provided input during my graduate research career.

I want to acknowledge Doctors Lynda Keifer and Ryan Blaustein for instructing me in the laboratory and for continuing to support me even though they moved on from the USDA many years ago.

Appreciation is extended to my coworkers within the Environmental Microbial and Food Safety Laboratory (EMFSL). The microbial water quality research using farm ponds is very different from typical research performed within the lab, but we still manage to help each other when possible. Specifically, I want to thank Laura Del Collo for being such a good friend over the years and for helping guide me through contracting issues. I want to thank both Jackie Sonnier and Nadja George for their help in the lab and for being there to help troubleshoot. Drs. Jo Ann van Kessel and Jeff Karns were equally helpful in providing microbiological insight and even resources to help accomplish my goals. I want to thank Drs. Seonwoo Kim and Serajus Salaheen for their support and friendship. Within EMFSL, I would also like to thank Drs. Bradd Haley, Manan Sharma, Jitu Patel, and Monica Santin-Duran. Each of these people helped in some way as I have developed as a scientist. Thanks to Doctors Insuck Baek and Moon Kim for supporting the research and for designing the imaging platform utilized in Chapter 4. Thanks to George Meyers for helping us locate Pond 3 and for providing us with the details we needed regarding the surrounding fields.

I would like to especially acknowledge the help of my immediate coworkers Miss Jaclyn Smith and Miss Billie Morgan. Together, we have been able to efficiently collect very interesting data over the past 5 years by working as a cohesive team. I would like to also thank all the interns who participated in our research in the last 5 years including Mauricio Peña, Cesar Hernandez, Ashely Vázquez, Nhu Le, Nailah Michelle Washington, and Lauren Wyatt-Brown.

A special thank you to Doctors Kenneth Staver and Kate Everts for allowing us to work on the Wye Research and Education Center (WREC) in Wye Island, Maryland. Thank you to Mike Newell for coordinating with us in the field and supporting our research needs. A final thank you to the private farm owners and operators who allowed us to perform research on their properties and who would frequently drop what they were doing to support us.

Table of Contents

List of Tables

Chapter 2

Chapter 3

Chapter 5

Chapter 6

List of Figures

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 1: Introduction

1.1 Background

Ensuring that agricultural waters used for irrigation meet certain quality standards is paramount to reducing the incidences of foodborne illness. A major subsection of the overall health and financial burdens associated with foodborne illnesses are the result of consuming produce which has been irrigated with waters containing bacterial, viral, or parasitic pathogens. Pathogenic microorganisms which are transferred to crops and later ingested by humans have been responsible for countless illnesses, hospitalizations, and deaths as well as billions of dollars a year in medical treatment expenses, loss of productivity, and produce recalls (Painter et al., 2013; Hussain and Dawson, 2013; Hoffman et al., 2015; CDC 2021; WHO, 2020).

To reduce the incidences of foodborne outbreaks linked to microbial pathogens, regulatory bodies in the U.S. and worldwide have required testing of irrigation waters (FDA, 2015; Allende and Monaghan, 2015). The microbial quality of these waters is commonly determined by measuring the concentrations of the fecal indicator bacteria (FIB) *Escherichia coli* (*E. coli*) and comparing the results to numeric criteria. Failure to meet the mandated criteria can result in the need to perform corrective measures such as increasing the time between irrigation and harvest, treating the water with disinfecting chemicals, or discontinuing the use of an irrigation water for fresh produce (Banach and van der Fels-Klerx, 2020).

Concentrations of both *E. coli* and pathogenic microorganisms are extremely variable in environmental waters. This variability is often broken down into the

spatial and temporal components. Many studies have been performed with the aim to elucidate the spatiotemporal variability of FIB and pathogens in irrigation waters. However, the temporal component is often evaluated in finer detail than the spatial component. For example, many samples may be collected in a study which were sampled weekly, biweekly, or monthly, but often only a single sample is taken from a waterbody. These sampling disparities may at least partially be explained by the sampling requirements in regulative documents such as the Food and Drug Administration's Food Safety Modernization Act's (FSMA). In this act, irrigation water sources must be sampled 20 times over a 2- to 4-year period, but no guidance is given on sampling protocols other than "as close as practicable to harvest" (FSMA, 2015). If substantial spatial variability exists across a waterbody, then the choice of sampling location may result in the over- or under-estimation of FIB concentrations within the waterbody due to the propensity of sampling convenient locations (e.g. nearshore and at the water surface). With a multitude of possibilities of where to sample within irrigation sources, it appears evident that new guidance is needed to determine sampling locations which best represent the pond-wide average. Identification of such locations would also help reduce the need to sample extensively across a waterbody.

Sampling at depth is not routinely performed in microbial water quality studies of irrigation water sources. However, the largest volume of water within any waterbody is that which is contained below the surface and is what is likely withdrawn during irrigation events. If there exists substantial vertical gradients of *E. coli* or pathogen densities in subsurface waters, then samples collected at the surface

(e.g. 0 to 30 cm) are not representative of the concentrations within the pond. Research is needed to assess if concentration gradients can form in small irrigation ponds which are sometimes assumed to be homogenously distributed and to what effect these gradients have on concentrations of *E. coli* in the withdrawn water.

Estimating *E. coli* concentrations from water quality parameters appears to be a beneficial way of increasing the efficiency of microbial water quality monitoring. Although relationships between microbial concentrations and parameters which describe the aquatic habit are expected to exist, quantifying these relationships has been difficult due to their complexity. In recent years, the field of machine learning (ML) has become attractive for processing complex datasets and discerning non-linear or threshold based relationships between inputs (e.g. environmental or economic variables) and the output which is being predicted (e.g. stock futures, changes in home pricing, concentration of contaminants etc.). Machine learning is gaining a lot of attention within the field of microbial water quality studies as aquatic environments are typically highly heterogeneous and can be characterized with countless measurable variables which may influence the observed concentrations of FIB or pathogens in water (Quetglas et al., 2011). However, the use of ML in microbial water quality studies of irrigation sources, and specifically in farm ponds, is extremely scarce in the literature. ML shows promise for expediting microbial water quality determinations using environmental variable data which can be rapidly or instantaneously measured such as the levels of turbidity, pH, and dissolved oxygen. Therefore evaluation of ML within microbial water quality studies is warranted. With over several hundred ML algorithms in existence, comparisons between general types

of algorithms is needed to assess their applicability for predicting *E. coli* concentrations and for determining the most influential variables for making predictions.

In summation, there exist substantial inadequacies in the monitoring of microbial water quality of irrigation sources and numerous knowledge gaps must be addressed. This dissertation aims to enhance our understanding of the inherent spatial and temporal variability of *E. coli* concentrations in agricultural irrigation ponds. To accomplish these tasks, several methodologies will be evaluated which are new to this field, but have successfully applied in other fields of research. These new methods will be expected to 1) reduce sampling burdens, 2) improve the accuracy of microbial assessments, and 3) expedite these assessments using rapidly sensed water quality variables. Altogether these new methodologies hold the promise to facilitate and improve our ability to monitor the microbial quality of irrigation sources and by extension reduce the incidences of foodborne outbreaks in the United States and abroad.

1.2 Objectives

The overarching objective of this work is to develop effective microbial water quality monitoring strategies which will improve our understanding of the inherent spatiotemporal variability of FIB in irrigation waters and by extension reduce the incidences of foodborne outbreaks caused by microbial pathogens. To accomplish this objective, a series of monitoring experiments were conducted to assess the variability of *E. coli* in farm ponds which are commonly used in the U.S. and elsewhere to supply water for irrigation.

The specific objectives of this work were:

Objective 1: Examine the spatiotemporal variability of *E. coli* concentrations in two working irrigation ponds

Sub objective 1.1: Evaluate the applicability of the Mean Relative Difference (MRD) analysis for assessing variability and identifying locations which best represent the pond-wide relative average *E. coli* concentrations.

Sub objective 1.2: Evaluate the empirical orthogonal function (EOF) analysis for separating the spatial and temporal components of the variability in *E. coli* concentrations as well as for determining locations which deviate the least from the pond-wide average.

Sub objective 1.3: Assess relationships between *E. coli* concentrations and environmental covariates in the studied ponds.

Objective 2: Assess 3-dimensional variability of *E. coli* in a working irrigation pond based on measurements taken at multiple depths.

Sub objective 2.1: Investigate the effects of 3-dimensional variability of *E. coli* concentrations on water withdrawn for irrigation through use of the Environmental Protection Agency's Environmental Fluid Dynamic Code (EFDC).

Sub objective 2.2: Evaluate the impacts of changing the location of the irrigation intake pipe on the resultant concentrations of *E. coli* within the water removed from the pond.

Objective 3: Determine if the relative time of day for sample collection significantly impacts the results of microbial water quality surveys of irrigation ponds.

Sub objective 3.1: Calculate and summarize the microbial die-off rates for *E. coli* samples collected at different times of the day.

Sub objective 3.2: Determine the most influential water quality variables related to the changes in *E. coli* concentrations throughout the day.

Objective 4: Apply machine learning algorithms to predict *E. coli* concentrations using physiochemical water quality variables as predictors.

Sub Objective 4.1: Evaluate and compare several popular machine learning algorithms for the prediction of *E. coli* concentrations and compare the predictions to results obtained using a conventional statistical approach.

Sub Objective 4.2: Evaluate different predictor sets ranging from simple to complex and assess if model accuracy is improved with larger predictor sets.

Sub Objective 4.3: Document and rank the most important variables for the prediction of *E. coli* concentrations across ponds and years of observations.

1.3 Layout of the Dissertation

The organization of this dissertation presents first a literature review (Chapter 2) which summarizes the previous work that has been performed in microbial water quality surveys of lentic irrigation water sources. Public health and economic burdens associated with foodborne outbreaks are discussed as well as irrigation water as a vector for contamination in pre-harvest environments. Key legislation related to the microbial quality of irrigation waters is discussed and the spatial and temporal dynamics of *E. coli* concentrations reported in irrigation ponds, lakes, and reservoirs from numerous studies is documented. Sampling methodologies from 20 irrigation water quality studies conducted between 2005 to 2021 is summarized and the potential inadequacies in sampling regimes are highlighted. Finally, a brief overview of machine learning is introduced and the applicability of its use within microbial water quality studies is discussed.

Chapter 3 addresses the first objective of this dissertation. The study focuses on documenting spatial and temporal variability of *E. coli* concentrations in two working irrigation ponds between 2016 and 2018. Mean relative difference (MRD) and empirical orthogonal function (EOF) analyses for the detection of sampling locations which are consistently representative of the average concentrations across the ponds.

The MRD portion of this work was accepted for publication in *Water, Air, and Soil Pollution* in December 2021 and the portion utilizing EOFs has been accepted in the *Journal of Hydrology*.

Chapter 4 addresses the second objective of this dissertation. This study summarizes the results of 3-dimensional sampling of an irrigation pond. In this chapter, *E. coli* concentrations measured at different depths are statistically compared. Using the 3D concentration dataset along with fine-scale bathymetry data, the Environmental Fluid Dynamic Code (EFDC) is used to simulate the withdrawal of water from the irrigation pond and the temporal dynamics of the *E. coli* concentrations are presented over the course of an irrigation event. The effects of wind speed and direction on the *E. coli* concentrations withdrawn are evaluated. Material from this chapter was published in *Water* in the summer of 2020.

Chapter 5 addresses the third objective in this dissertation. It investigates the diurnal variations of *E. coli* concentrations in three irrigation ponds. Microbial die-off rates are calculated between different sample collection times and concentrations are statistically compared. A regression tree analysis is performed which revealed the most influential predictors of the *E. coli* concentrations within each pond. The contents from this chapter was submitted to *The Journal of Environmental Quality* in December 2021 and was accepted for publication in March 2022.

Chapter 6 describes results that utilizes and compare machine learning algorithms for the prediction of *E. coli* concentrations in waters of two ponds using multi-year observations. Predictor set sizes are varied and the effects on the accuracy of the models is assessed. The most influential predictors are compared between the

two ponds in this study. The material presented in this chapter was published in the *Food, Agriculture, and Water* section of the journal *Frontiers in Artificial Intelligence* in December 2021.

Chapter 7 summarizes the conclusions presented in the earlier chapters. Sampling recommendations are made based on the work performed described in the individual chapters and some future avenues of potential research are discussed.

Included in Appendix A is a short communication focusing on the dependence of *E. coli* concentrations on the water sampling depth. The contents of Appendix A contain and summarize the preliminary results of sampling three irrigation ponds at multiple depths between 2019 and 2021. In addition to depth, time of day (i.e. 9:00, 12:00, and 15:00) was investigated as a factor which may influence the observed *E. coli* concentrations. While these factors were investigated independently in Chapters 4 and 5 for depth and time of day, respectively, the goal of this short communication was to evaluate the combined effects of water sampling depth and time of day. Appendices B and C contain supplemental tables and figures, respectively, from each of the four research chapters.

Chapter 2: Literature Review

2.1 Public Health and Economic Burden

Foodborne illness is a worldwide issue caused by the consumption of foods contaminated with bacteria, viruses, parasites, or chemical substances. These illnesses are associated with over 200 diseases ranging from abdominal cramps to the development of cancers and it is estimated that 600 million people a year, roughly 1 in 10 people worldwide, fall ill after eating contaminated foods and of those 420,000 people will die (WHO, 2020). In the United States, there are roughly 48 million cases of foodborne illnesses every year resulting in 128,000 hospitalizations and 3,000 deaths while in Canada approximately 4 million cases of foodborne illness occur annually resulting in over 11,500 hospitalizations and 238 deaths (CDC, 2018; Thomas et al., 2015).

A major portion of the overall burden of foodborne illness is caused by the consumption of fresh produce which has been contaminated by pathogenic microorganisms. An analysis of outbreaks performed by Painter et al. (2013) estimated that between 1998 and 2008 foodborne pathogens in the United States were responsible for over 9 million illnesses with consumption of contaminated fruits and vegetables alone being responsible for the most illness (46%), the second most hospitalizations (38.1 %), and the 5th most frequent cause of death (22.9%). Another study which summarized outbreaks between 1999 and 2019 reported similar results as the previous study and attributed fresh produce as being responsible for 47.4 % of outbreaks in the U.S. and 34.1 % in the EU (Aiyedun et al., 2021). The consumption

of raw or minimally processed produce including tomatoes, peppers, spinach, lettuce, and similar products has been acknowledged as a major vehicle for foodborne outbreaks (Goodburn and Wallace, 2013; Weller et al., 2015; Bennet et al., 2018; Gutierrez-Rodriguez and Adhikari, 2018).

The estimated economic costs of foodborne outbreaks are substantial. Health-related costs are calculated as the sum of medical costs (hospital and physician services and drugs), quality-of-life losses (deaths, pain and suffering, and functional disability) as well as costs to others in society, e.g. costs to insurance companies that pay medical expenses and work productivity losses (Scharff, 2010). In a publication which summarized these economic burdens using data available in 2013, it was reported that foodborne illnesses impose over $15.5 billion in economic burden on the U.S. public each year (Hoffman et a., 2015). In a subsequent report, the same authors documented a $2 billion-dollar increase in the estimated annual economic burden between 2013 and 2018 (Hoffman and Ahn, 2021). Another study estimated the total annual economic loss from consumption of contaminated foods as being as high as $152 billion with $39 billion being attributed to directly to produce (Scharff, 2010). It has been acknowledged that statistical reporting associated with foodborne illness likely underestimates the true health and economic burdens due to poor or inconsistent reporting and our developing understanding of chronic issues associated with foodborne illness (Colombara et al., 2016; WHO, 2020).

The economic effects of foodborne outbreaks extend well beyond medical costs. Recalls represent an extensive revenue loss for food processors and distributors which can result in millions of pounds of recalled foods (Simpson et al., 2020).

Individual recalls often result in estimated economic losses in the hundreds of millions of dollars. It was estimated in 2013 that the total annual economic burden associated with recalls was nearly $7 billion (Hussain and Dawson, 2013). There are also costs associated with consumer hesitancy in buying previously recalled products that affects future sales (Weller et al., 2018).

Fresh produce is an essential staple in the modern diet which is promoted heavily in healthy eating campaigns. Global consumption has consistently increased for several decades (Uyttendaele et al., 2015) as agriculture has been tasked to meet the ever-growing needs of a steadily increasing worldwide population. The U.S. Department of Agriculture recommends consumers double their daily intake of fresh produce within a two thousand calorie diet (USDA, 2015). A greater interest in healthy eating and the year-round availability of most produce due to extensive global import and export increases the likelihood for outbreaks to occur (Suslow et al., 2003; FDA, 2009; Carstens et al., 2019).

The number of foodborne outbreaks has steadily increased between 1998 and 2018 and is projected to continue to climb worldwide (Smith and Fazil, 2019; Franz et al., 2018; CDC, 2021). Due to the overwhelming financial and public health burdens associated with foodborne illnesses in the U.S. and abroad, it is paramount to develop effective strategies which lead to the minimization of risks at all points of the farm-to-fork continuum. With the increased global concern associated with climate change, food scarcity, shrinking arable lands, and limited freshwater resources, farmers and producers will be asked to increase production in the face of mounting environmental challenges.

2.2 Major Pathogens, Sources of Irrigation Water, and Routes of Contamination

2.2.1 Pathogenic Microorganisms and Food Safety

Major pathogens associated with produce include pathogenic *E. coli*, *Salmonella* spp., *Campylobacter* spp., *L. monocytogenes, Shigella* spp., *Staphylococcus* spp., *Yersinia enterocolitica, Vibrio* spp. and Norovirus, A 2015 study reported that these pathogens alone accounted for 95 % of cases where the etiological agent was identified and revealed that nearly 80 % of cases go without identification (Hoffman et al., 2015). The same report attributed five pathogens (*Salmonella, Toxoplasma gondii, Listeria monocytogenes, Campylobacter, and* Norovirus) to more than 90% of the $15 billion annual healthcare burden associated with foodborne illness in the U.S.

Bacterial pathogens are the most common cause of foodborne outbreaks (Bintsis, 2017; Olaimat and Holley, 2012; Warriner et al., 2009; Carsten et al., 2019; Abebe et al., 2020). Three bacterial pathogens (*S. enterica*, *E. coli* O157:H7, and *L. monocytogenes*) were responsible for 83 multi-state and produce-associated foodborne outbreaks in the U.S. between 2010 and 2017 with affected produce including lettuce, sprouts, cantaloupe, tomatoes, cucumbers, melons, and hot peppers (Carstens et al., 2019). These outbreaks alone led to 4,658 illnesses, 1187 hospitalizations, and 55 deaths. Fresh vegetables were implicated in 67.3% of the cases with the rest linked to fruits (Carstens et al., 2019).

The majority of the foodborne pathogens are fecal in origin with optimal environments for survival being the intestinal tracts of animals (FDA, 2008). The spread of these pathogens is typically via the fecal to oral route through the consumption of food or water that has interacted in some way with fecal material (Warriner et al., 2009) . Additionally, while fecal sources are most common, certain pathogens such as *L. monocytogenes* are ubiquitous in the soil environment and can effectively contaminate produce grown near or under the ground surface (Warriner et al., 2009; Vivant et al., 2013; Carstens et al., 2019).

2.2.2 Irrigation in the U.S. and Farm Ponds as an Irrigation Water Source

Irrigation accounts for a major proportion of the total freshwater withdrawals in the United States. Approximately 118 Mgal of water are used for irrigation daily which represents 42% of the total annual freshwater withdrawal (Dieter et al., 2018). Surface-water sources account for 52% of the withdrawal with 48% attributed to groundwater sources (USGS, 2019). Water withdrawal for irrigation is disproportionally higher in the mid- and Western States than the East Coast with California, Idaho, Arkansas, Montana, and Colorado accounting for 54% of the nationwide withdrawal (Dieter et al., 2018). Increased temperatures, lack of precipitation, and a greater proportion of the overall agricultural land used for produce are responsible for the increased withdrawal in the West and mid-West as compared with the Eastern U.S (Wang, 2019; Dieter et al., 2018).

Surface water sources can be categorized as either being lentic or lotic. Lotic sources are characterized by rapidly flowing water and include rivers, creeks, streams, brooks, and springs while lentic sources are characterized by still or standing waters

with low water velocities and include ponds, lakes, and reservoirs (O'Keefe et al., 2021). The use of ponds for irrigation is especially popular due to the relative ease of construction within agricultural settings. There is no universal standard on the difference in area between ponds and lakes, but the former has been estimated to be between 1 m^2 and ~50,000 m^2 (Chopyk et al., 2018). Within the U.S. in 1980, there were more than 2.1 million farm ponds constructed on private lands to provide water for irrigation (NRCS, 1997). A 2005 report which analyzed satellite imagery data estimated that there may be between 8 and 9 million ponds in the U.S. with the largest density between the Great Plains and East Coast with notably less in the Western States (Renwick et al., 2005). The same study concluded that most of the identified ponds were man-made. With the relatively large number of agricultural irrigation ponds in the U.S., it is evident that more work is needed to document the spatial and temporal variability in the microbial water quality of these irrigation sources.

2.2.3 Irrigation as an important route for contamination

Contamination can occur in pre- or post-harvest environments. In the pre-harvest environment, pathogens can be introduced onto fresh produce via contact with irrigation water, splash from contaminated rainwater, or interactions with runoff, soil, fertilizers, manure, insects, and dust, and the direct deposition from animals (Carstens et al., 2019; Goodburn and Wallace, 2012; Bennet et al., 2018). In post-harvest environments, contamination can occur through interaction with contaminated wash waters, surfaces, cutting tools, personnel who are sick or have poor hygiene, inadequate storage or transport conditions, and handling in the home (Olaimat and

Holley, 2012). In the pre-harvest environment, irrigation appears to be a major vector for contamination and the irrigation water quality can be used as an indicator of produce safety (Suslow et al. 2003; FDA, 2008; Uyttendaele et al., 2015; Bennet et al., 2018; Lee et al., 2018).

Irrigating with contaminated water has been shown to effectively transmit pathogens to fresh produce (Guo et al., 2002; Hintz et al., 2010) and several investigations have linked contamination events to the water that was used for irrigation (Ackers et al., 1998; Kozak et al., 2013; Nygård et al., 2008; Gelting et al., 2015). Sources of contamination which may reach waters used for irrigation include sewage sludge and livestock wastes applied to land, animal feces deposited on land or directly into water by grazing or wild animals, and discharges from septic tanks or combined sewer overflows (Tyrrel and Quinton, 2003; Gerba and Smith, 2005). Agricultural practices can impair surface water quality by acting as both point and non-point sources of fecal microorganisms (Markland et al., 2017; U.S EPA, 2015). Precipitation or irrigation events which occur on manure- or biosolid amended lands facilitate the release and transport of large numbers of fecal bacteria with overland flow or tile drainage which can reach and contaminate surface water sources (Blaustein et al., 2015; Stocker et al., 2017; Givens et al., 2016). Likewise, pastured lands can also degrade the microbial water quality of adjacent and downstream waters (Nagels et al., 2002; Muirhead and Monaghan, 2012). Inverse correlations between distance to pastureland and higher fecal microbe levels in surface waters have been commonly reported (Conboy and Goss, 2000; Shelton et al., 2011; Benjamin et al., 2013; Poulsen et al., 2018). Similarly, microbial water quality of rivers and streams

has been shown to improve with increasing distance from urban areas (Kovačić et al., 2011; Derlet et al., 2012). The U.S. EPA currently recognizes agriculture as the most probable source for the most miles of impaired streams and rivers in the U.S and lists microbial pathogens as the leading cause of impairment (US EPA, 2021). For Lakes, reservoirs, and ponds, agriculture is 3rd most probable cause of impairment with pathogens accounting for 503,200 acres of threatened or impaired waters of these types. With so many miles or acres of threatened or impaired surface waters due to microbial pathogens, it is evident that a better understanding of the fate and transport processes associated with these pathogens is needed to reduce the incidences of foodborne outbreaks.

2.2.4 Persistence of pathogens in irrigation waters and after transfer to crops

Survival of microbial pathogens in waters intended for irrigation is controlled by many factors. Temperature effects on microbial pathogens in surface waters are well noted. A meta-analysis of 470 *E. coli* datasets which investigated the effects of temperature in surface waters found that increased temperatures led to higher inactivation rates in all water types evaluated which included rivers, wastewaters, lakes, groundwater, pristine sources, and agricultural waters (Blaustein et al. 2013). In a subsequent meta-analysis, the same authors found the dependencies to hold true for both *Salmonella* and enterococcus (Pachepsky et al., 2014). The strength of the temperature dependencies has been shown to differ between water types including rivers/streams and lakes (Faust et al., 1975; Stocker et al., 2014). In addition to temperature, other factors that affect the survival of pathogens in waters are nutrient

availability, osmotic stress, UV radiation, unfavorable pH (e.g. < 6.5 or > 8.5), predation, and competition, but the extent of these factors may be expected to differ between microorganisms as some are better suited for survival in non-host environments (Barcina et al., 1986; Winfield and Groisman, 2003; Warriner et al., 2009; Maraccini et al., 2016).

Fecal microorganisms which have been transported to waterways will remain in the water column and eventually become inactivated, will be used for irrigation or drinking water supplies, and/or will settle into the bottom sediments. The sediments of waterbodies have been shown to be large reservoirs of fecal bacteria in the environment as they provide a substrate for growth which is oftentimes rich in nutrients, offers protection from UV radiation and predation by other microorganisms, and negates some of the effects of diurnal temperature oscillations (Wilkinson et al., 1995; Jamieson et al., 2005; Muirhead et al., 2006; Stocker et al., 2018). If sediments are disturbed through events such as bioturbations, such as when cattle cross a stream, or when high-flow events occur, massive numbers of fecal bacteria may be re-suspended into the water column (Cho et al., 2010; Stocker et al., 2018; O'Callaghan et al., 2018). Transport of fecal microorganisms in the sediment can also occur during baseflow conditions via in-channel active (e.g. chemotaxis) and/or passive (e.g. hyporheic exchange) transport processes between the sediment and water column (Ginn et al., 2002; Grant et al., 2011; Pachepsky et al., 2017; Park et al., 2017). Positive correlations between *E. coli* concentrations in sediment and the water column across sampling locations have been reported in several studies which

may demonstrate an exchange between the two mediums (Wu et al., 2009; Piorkowski et al., 2014; Stocker et al., 2018; 2019).

Once contact has been established, pathogenic microorganisms have been reported to create biofilms on plant surfaces which reduce environmental stressors and enhance survival (Olaimat and Holley, 2012). In one study, it was shown that *E. coli* O157:H7 could survive for at least 25 days on leaf surfaces once introduced onto the plant tissues (Zhang et al. 2009). In another study, *E. coli* O157:H7 and *Salmonella* were observed to survive for 177 and 231 days, respectively, on parsley leaves (Islam et al., 2004a,b). Microorganism survival on fresh produce is mediated by nutrient availability, UV radiation, toxic compounds released by the plant, competition with other microorganisms, and desiccation (Whipps et al., 2008; Olaimat and Holley, 2012). Meteorological factors, particularly moisture and temperature have been described to exert the most control over survival in the environment (Suslow et al., 2003; Hoagland et al., 2018).

Microbial colonization of plants may also occur via infiltration into plant tissues which promotes survival in pre- and post-harvest environments. The quickest and most effective route for this to occur is via damage to the external plant tissue which expose internal sections (Eblen et al., 2004; FDA, 2009). Another way is via microorganism activity at the plant surfaces. For example. *E. coli* O157:H7 was shown to produce specialized cells which caused stomatal openings in leafy greens and allowed the bacteria into internal plant compartments via chemotaxis, motility, and quorum sensing (Saldaña et al., 2011). *Salmonella* internalization has been widely documented as well in produce such as tomatoes, leafy greens, herbs, oranges,

and sprouts (Zheng et al., 2013; Goldberg et al., 2011; Eblen et al., 2004; Solomon et al., 2002). Once internalized, standard sanitation practices become less effective as removal and inactivation of pathogens becomes more difficult (Wright et al., 2017; Olimat and Holley, 2012).

Foodborne pathogens present in irrigation waters can also be infiltrated into agricultural soils where they may persist for extended periods of time relative to being exposed at the surface. For pathogens at the surface, elevated UV light exposure, diurnal temperature cycles, and drying are the primary factors driving inactivation (Rokitko et al., 2003; 2004; van Kessel et al., 2007; FDA, 2009; Stocker et al., 2015). Other potentially pertinent factors include soil organic matter content, soil moisture, the extent of eutrophication and availability of substrate, soil pH, oxidation-reduction potential, and microbial interactions (Oliveira et al., 2012). Microbial pathogens which enter the soil either through incorporation or with water may exhibit extended persistence as these factors have a less effective reach within the subsurface environment (Hutchison et al., 2004). Survival in the soil environment has been shown to occur for a number of days, weeks, months, and even years (Avery et al., 2004; Entry et al., 2005; Muirhead, 2009). The latter long-term persistence was documented by Brennen et al. (2010) who measured *E. coli* concentrations in manure amended soils for 9-years after application which led the authors to conclude autochthonous populations had become naturalized in the soil.

Once in the soil, splash action of water caused by precipitation or irrigation events has been shown to contaminate adjacent plants and those up to a half meter away (Monaghan and Hutchison et al., 2012; Moyne et al., 2011; Oliveira et al.

2012). If pathogens are able to survive in the soil for long periods of time, then it is possible that each precipitation or irrigation event may result in the spread of the microorganism to the crop (Allende et al., 2018). The possibility of contamination of crops from splash action on soil surfaces may undermine current FSMA PSR corrective measures such as allowing for microbial die-off to occur in the field.

Overall, irrigation has been shown to be conducive for the transfer of microbial pathogens from water to crops and soils. Soil ecosystems can support FIB and pathogens survival for extended periods of time. Pathogenic microorganisms have been shown to develop biofilms on plant surfaces which also enhance survival. Internalization of deposited pathogens into plant tissues is a serious concern as this internalization may make post-harvest sanitation practices ineffective. Altogether, it is clear that the microbial quality of waters used for irrigation is a significant factor for food safety.

2.3 Food Safety Regulations and Fecal Indicator Bacteria

2.3.1 Key Legislation

Due to the severity in economic and public health effects of foodborne illnesses, several regulatory guidelines have been passed within legislation that focuses on reducing the contamination of produce. In 1998, the FDA released the "Guidance for Industry: Guide to Minimize Microbial Food Safety Hazards for Fresh Fruits and Vegetables" which detailed the "scientific agreement" that irrigation practices that expose the edible portion of plants to contact with contaminated water

may increase the risk of foodborne illnesses (FDA, 1998). This document provided guidance on minimizing risks of contamination and advocated adhering to good agricultural practices (GAPS) where practical such as avoiding irrigation close to harvest and using low-volume sprays, drips, furrow, or underground application methods as part of an irrigation program. The document also contained a section regarding microbial testing of agricultural waters including measurement of fecal coliforms and generic *Escherichia coli* (*E. coli*), but provided no numerical criteria to define acceptable levels. In 2008, the FDA updated the guidance document to reflect the FDA's then current thinking on the topic of food safety in the U.S. The update expanded the recommendations on safety controls and good agricultural practices ("GAPS") as well as provided new information regarding the number of outbreaks since the publication of the original document. Of note, the updated document stated that "prevention of microbial contamination of fresh produce is favored over reliance on corrective actions once contamination has occurred". This principle set the foundation for adopting and expanding GAPS and assessing potential on-farm contamination routes with the aim to preemptively minimize risks so that foodborne illnesses and the associated economic costs would pose a lesser threat.

Concurrent with the updated guidance from the FDA, the California Leafy Greens Marketing Agreement (LGMA), a collective who currently produce 90% of the leafy greens grown in the U.S., established numerical criteria in 2008 for the testing of irrigation waters in response to a 2007 *E. coli* O157:H7 outbreak (LGMA, 2015). The criteria were based on that proposed a few decades earlier by the U.S. EPA for recreational waters which were based on the levels of generic *E. coli* present

in a 100-mL sample. The specific numbers were i) for overhead irrigation of foliar surfaces, no water sample can contain > 235 most probable number (MPN) of *E. coli* per 100-mL (ii) for drip irrigation of roots, no sample can contain >576 MPN *E. coli* per 100-mL, and (iii) for either overhead or drip irrigation, *E. coli* concentrations cannot exceed an average of 126 CFU or MPN per 100-mL among five samples taken over 30-days (Shelton et al., 2011). In the same time period, the Arizona-LGMA committee was formed which is considered a sister program to the California based agreement containing very similar guidelines (LGMA, 2015)

Three years after the FDA released the 2008 guidance document, the 2011 Food Safety Modernization Act (FSMA) was signed into law by President Barack Obama. FSMA was a comprehensive reform bill that authorized and mandated the FDA to require science-based preventative controls across the food supply chain including growing, harvesting, packing, and holding/transport (Ribera et al., 2016). Initially FSMA did not contain numerical guidelines for the levels of fecal bacteria that were acceptable without triggering re-testing or abatements. In 2016, FSMA was amended and the Final Rule on Produce Safety (also known as the "produce safety rule" (PSR) established standards for the "growing, harvesting, packing, and holding of produce for human consumption" which was put into effect in January of 2016 (FDA, 2015). Under the PSR, growers are to periodically test surface waters used for irrigation in order to develop a "microbial profile" to discern the average concentrations present as well as the potential variability in those concentrations. Specifically, an initial survey of 20 samples were to be taken over the course of 2- to 4-years as close as practicable to harvest. From these, a geometric mean (GM) and

statistical threshold (STV) would be calculated. The geometric mean represents the logarithmic average *E. coli* concentrations and the STV would be a value such that 90% of the samples would be below the STV value. The values for the GM and STV were 126 CFU 100 mL^{-1} (2.1 log CFU 100 mL^{-1}) and 410 CFU 100 mL^{-1} (2.61 log CFU 100 mL^{-1}), respectively (FDA, 2016). After the initial survey, the microbial profile of the irrigation water would be updated by taking 5 samples a year to create a rolling average. In the case of non-compliance, corrective measures may be taken including allowing die-off in the field to occur post-irrigation, allowing time for die-off between harvest and end of storage which would include commercial activities such as washing, and/or to treat the water with chemicals or antimicrobial substances (FDA, 2017).

The FSMA PSR compliance dates accommodated changes needed for farmers operating at different production scales. These dates were January 26, 2018 for farms earning more than $500,000 in a 3-year period, January 28th 2019 for farms earning more than $250,000 but less than $500,000, and January 27, 2020 for farms earning more than $25,000 but no more than $250,000. At the time of the creation of this document, the FSMA PSR is still in effect as are the California and Arizona LGMAs (LGMA, 2021; FSMA, 2021) and compliance dates for varying FSMA produce safety regulations continue through 2024 (FSMA, 2021).

2.3.2 Escherichia coli as an indicator of fecal contamination

The 2008 California and Arizona LGMAs and the 2016 FDA PSR established the measurement of generic *E. coli* for regulatory compliance and set criteria based on

the concentrations measured in waters used for irrigation. The criteria were adopted from the EPA's 1986 recreational water quality criteria which used generic *E. coli* and enterococci microbial water quality determinations. The datasets and resulting criteria were collected as part of EPA's epidemiological studies conducted during the late 1970's and early 1980's when levels of *E. coli* and enterococci in surface waters were compared with illness rates of waterborne pathogens (U.S. EPA, 2012). The findings of these studies indicated that *E. coli* and enterococci had the highest correlations to disease incidence in freshwater systems and marine beaches, respectively, and therefore enumeration of these organisms would be used to assess fecal loading and potential health impacts (Ishii and Sadowsky, 2008).

Studies into the applicability of *E. coli* as an indicator were based on knowledge that *E. coli* is a relatively abundant microorganism present in animal feces and therefore elevated levels of this microorganism may represent occurrences of fecal contamination in water. In all, there are 7 criteria that a fecal indicator organism (FIB) must exhibit which includes: i) it should be a member of the intestinal microflora of warm-blooded animals, ii) it should be present when pathogens are present and absent in uncontaminated samples, iii) it should be present in greater numbers than pathogenic microorganisms, iv) it should be at least equally resistant as the pathogen to environmental stresses and to disinfection processes, v) it should not multiply in the environment, vi) it should be detectable easily, rapidly, and inexpensively, and vii) it should be non-pathogenic (Bitton, 2005).

E. coli are a proteo-bacteria of the family Enterobacteriaceae of genus *Escherichia* (Odonkor and Ampofo, 2013). *E. coli* are a gram-negative rod-shaped

bacterium which reside in the lower intestines of warm-blooded organisms. The microorganism is a facultative anaerobe of which most strains are harmless and make up less than 1% of the intestinal microbiota (Delmas et al., 2015). *E. coli* experience optimal growth conditions at temperatures that mimic the host environment (37 °C or 98.6 °F) but are considered thermo-tolerant as they can survive in temperatures as high as 50 °C (Odonkor and Ampofo, 2013). The bacterium has also been shown to survive in freezing temperatures (< 0 °C) (Smith, 1995; Berry and Foegeding, 1997).

Several pathogenic strains of *E. coli* carry a combination of virulence genes which can be utilized to cause intestinal and extra-intestinal infections (Delmas et al., 2015). The virotypes of *E. coli* include enterotoxigenic (ETEC), enteropathogenic (EPEC), enteroinvasive (EIEC), enterohemorrhagic (EHEC) and enteroaggregative (EAEC) *E. coli* which are classified based on the infection mechanisms, biochemical properties, and resulting pathologies in infected organisms (Odonkor and Ampofo, 2013). The most common symptoms of infection are profuse and often bloodied diarrhea as well as high fever. Other more serious symptoms include hemolytic-uremic syndrome, sudden kidney failure, and inflammatory responses (Bien et al., 2012).

The efficacy of *E. coli* as an FIB has been shown to vary widely across whereby in some cases *E. coli* concentrations correlate with those of pathogenic microorganisms and in other cases there is no relationship. In a review article which analyzed the co-measurement of indicator bacteria and pathogenic microorganisms in surface waters between 1970 and 2009 it was reported that 29 of 40 studies found no correlation of *E. coli* and pathogens while 11 of 40 did (Wu et al., 2011). The authors

attributed the lack of correlation to insufficient sample sizes as well as higher levels of other indicators such as total or fecal coliforms in the environment (and after fecal deposition) than *E. coli*. The same report also described variability in correlations due to geographic and methodological differences across studies as well as differences attributed to water type (e.g. freshwater, saline or brackish, wastewater etc.).

Associations between *E. coli* and pathogens have also been observed to be pathogen specific. For instance, *Salmonella* has been shown to have significant correlation with *E. coli* concentrations in surface waters in more studies than not, but the opposite is true for pathogens such as *Campylobacter* and *Cryptosporidium* (Pachepsky et al., 2016; Wu et al., 2009; Ceuppens et al., 2015).

Additional reasons for poor relationships may be that *E. coli* is a bacterium which is used to assess the contamination of bacterial but also non-bacterial pathogens such as protozoa or viruses which have been shown to exhibit differences in fate and transport within the environment (Brookes et al., 2004). The growth of *E. coli* in non-host environments also contradicts the criteria listed in Bitton (2005) for an ideal indicator organism (Stocker et al., 2015; Ishii and Sadowsky, 2011). In this case, *E. coli* measured in soils or waters do not represent recent fecal contamination, but rather are representative of persistent or naturalized populations. Finally, it is likely that there may be threshold values of indicator organism concentrations which correspond with elevated levels of pathogenic microorganisms. A couple of studies have reported average prevalence (%) of pathogens and pathogenic markers in Pennsylvania waters as having been 0 % with *E. coli* levels under 300 CFU 100 mL^{-1} however, above this threshold prevalence increased to between 20 and 50 % (Duris et

al., 2013; Pachepsky et al., 2016). A similar 2015 study which analyzed 1037 water samples from 45 farms in five countries determined that *E. coli* was an effective FIB when concentrations were above 100 CFU 100 mL^{-1} . The efficacy to *Salmonella* and Shiga-toxin producing *E. coli* were present whereas weaker findings were reported for *Campylobacter* (Ceuppens et al., 2015). A recent study in Spain which examined irrigation waters found a specific threshold of 2.35 log CFU 100 mL^{-1} (224 CFU) below which there was a 90% probability that samples were not contaminated with pathogens (Truchado et al. 2018). Numerous other researchers have also shown that the probability of pathogen presence increases with the measured *E. coli* concentration in surface waters (McEgan et al., 2013; Partyka, 2016; Partyka, 2018; Truitt et al., 2018; Havelaar et al., 2017).

Several other indicator microorganisms have been evaluated including coliform bacteria, fecal streptococci (enterococci), anaerobic bacteria including *Clostridum perfringens*, *Bacteroides spp.*, and bifidobacteria, and bacteriophages such as somatic coliphages, F+ coliphages, and phages infecting *Bacteroides fragilis* (Bitton, 2005). However, only *E. coli* and enterococci have been widely adopted for the characterization of the microbial quality of water for recreation by the EPA and for irrigation by the FDA, respectively. The LGMAs do not recommend the measurement of enterococci which has been described a better indicator in marine waters (Wade, 2003). In the 2012 Recreational Water Quality Criteria (RWQC) the EPA removed measurement of total coliforms from assessing fecal contamination, but recommended the measurement continue for drinking water standards (U.S. EPA, 2013).

There is no obvious perfect indicator organism for the assessment of fecal contamination of recreational and agricultural waters. Efficacy of any specific indicator organism is dictated by season, region, and local fecal inputs as well as the extent of differences between indicators and pathogens in terms of fate and transport, sampling volumes and methodology, and responses to environmental conditions (Wu et al., 2009; Pachepsky et al., 2014; Truchado et al., 2018). Efficacy is likely to improve when more samples are taken and the author agrees with a conclusion in Savill et al. (2001) which stated, "*a low Spearman's r_s value does not allow 'no correlation' to be inferred but merely denotes that there may not be sufficient evidence to conclude that a relationship exists*". It is also plausible that sampling designs are responsible for the lack observed relationships. As noted in Benjamin et al. (2013), "*grab sampling involves collecting a relatively small quantity of water at a particular point in time and might lead to false negative results because E. coli O157:H7 can be present at low levels in surface water or might not be homogenously distributed in the water source*".

2.4 Monitoring the Microbial Quality of Irrigation Waters

2.4.1 Environmental Monitoring and Assessment

In addition to detecting when caution or corrective measures should be taken when using an irrigation source due to microbial exceedances, the focus of many research and monitoring programs is to better understand the variability and factors

that control the populations of FIB or pathogens in the environment. High spatial and temporal variability of microbial water quality metrics is common. Unfortunately, in most studies either the spatial or temporal component of the variability is favored while the other is not well elucidated. Commonly, the temporal component is given greater focus and more measurements are taken over time than over space which has revealed interesting inter-annual FIB dynamics.

2.4.2 Spatial variability of microbial water quality metrics

Dense spatial sampling of *E. coli* in freshwater ponds and small lakes or reservoirs is rarely performed due to considerations of time, availability of funds, and/or difficulties in accessing potential sampling locations. Table 2.1 documents the apparent lack of accounting for spatial differences in irrigation water sources. In the overwhelming majority of studies, only a single sample taken from an irrigation pond, reservoir, or lake was collected and used for microbial water quality assessments.

The limitations on resources for sampling have led some researchers to sample as close as possible to the irrigation water intake location and/or at a few locations along the perimeter of the pond (Gu et al., 2012; Luo et al., 2015; Lee et al., 2018) while others have elected to include analysis of water from irrigation distribution systems and/or irrigation water that has reached the crops (Antaki et al., 2016; Pagadala et al., 2015). Lee et al., (2018) documented large spatial differences of *E. coli* and *Salmonella* concentrations among 3 locations around the perimeter of an irrigation pond measured from March to September. The authors related the spatial differences to potential microhabitats present within the pond along with physicochemical parameter preferences of the organisms. They also found the

samples taken near the intake (pond interior) contained lower concentrations of *E. coli* than those sampled near the bank. Conversely, Harris et al., (2018) reported higher *E. coli* concentrations in interior sampling locations than bank locations in two GA ponds. Pagadala et al., (2015) measured total aerobic bacteria, *E. coli*, and total coliforms in irrigation well and pond water samples within the source as well as at the end of drip lines and found no significant differences between concentrations of all three groups. In another study, 10-fold or greater differences were observed between pond water at the source and what is released at the end of drip lines or pivot sprinkler heads (Antaki et al., 2016). Another group of authors sampled 5 locations along the flow path of 6 different California reservoirs and found that the microbial water quality in downstream irrigation supplies was not statistically associated with conditions in the reservoirs, rather there was a greater similarity to water quality just upstream of the reservoirs (Partyka et al., 2018)

Table 2.1. Summary of sampling information in 20 microbial water quality studies performed in lentic waterbodies.

Microorganism	Location	Water Type	Sites	Time of Day	Sampling Frequency	Sample depth (m)	Samples per visit	Sampling locations	Total samples	Reference
EC, S spp.	Southern Georgia	P	2	N.A.	Intermittently	0 - 1	1 – 2	1	83	Antaki et al. 2016
EC, O157, S spp.	California Central Coast	P	3	N.A.	Intermittently	N.A.	N.A.	1	46	Benjamin et al., 2013
EC, STEC, S spp.	Belgium	P	5	N.A.	N.A.	Surface	1 – 2	1	53	Delbeke et al., 2015
APC, TC, FC, EC, EN	Southern Pennsylvania	P	1	N.A.	N.A.	N.A.	1	1	55	Draper et al., 2016
EC	Northern Texas	L	6	N.A.	Monthly	0.5	3	3	216	Durham et al., 2016
Campy spp.	Southern Georgia	P	10	N.A.	Monthly	Surface	2	1	280	Gu et al., 2012
L. mono., S. ent, EC	Eastern Shore of Virginia	P	4	N.A.	Weekly	Surface	1	1	12 - 49	Gu et al., 2021
EC, S spp.	Southern Georgia	P	2	N.A.	Precipitation Driven	Surface	2	3	24	Harris et al., 2017
EC; Crypto. Spp.	Western Canada	R	1	N.A.	Biweekly	Surface	1	1	16	Jokinen et al., 2019
EC, S. spp.	Southern Georgia	P	5	N.A.	Monthly	Surface	3	3	507	Lee et al. 2018
EC, TC, APC, S spp.	Eastern Shore Maryland	P	9	N.A.	Various	Surface	1	1	N.A.	Marine et al., 2015
EC, TC, S spp.	Central Florida	P	1	Before 12 pm	Monthly	0 – 0.2	1	1	12	McEgan et al., 2013
S. spp.	Mid-Alantic U.S.	P	11	N.A.	Intermittently	Surface	1	1	34	Micallef et al., 2012

EC, TC, FC, S spp.	Eastern Shore Maryland	P	1	N.A.	Biweekly	Surface	3	3	87	Pahl et al., 2013
EC, FC, S. STEC, O157	Eastern California	R	6	N.A.	Monthly	Multiple	1	3	N.A.	Partyka et al., 2018
EC, TC, ENT, Aero	Eastern Shore Maryland	P	2	N.A.	Various	0.15 to 0.30	1	1	69	Solaiman et al., 2020
EC, TC, FC, EN	Southern Ontario	P	27	N.A.	Intermittently	Surface	2	1	501	Steele et al., 2005
EC, TC, EN, S spp. STEC	Western Florida	P	6	Before 12 pm	Weekly	Surface	1	1	540	Topalcengiz et al., 2017
APC, EC, TC, S spp.	Eastern Shore Virginia	P	20	Before 11 am	Biweekly	0.5	1	1	400	Truitt et al., 2018
EC, TC	Ohio	R	4	N.A.	Various	0.5	1	1	160	Won et al. 2013

P = pond, R = reservoir, L = Lake. N.A. = information not available. EC = generic *E. coli*, TC = total coliforms, EN = enterococci, FC = fecal coliformsS spp. = *Salmonella* spp., STEC = shiga-toxin producing *E. coli*, Aero = *Aeromonas* spp., Crypto spp. = *Cryptosporidium* spp., O157 = *E. coli* O157:H7, Campy spp. = *Campylobacter* spp., S. ent = *Salmonella enterica*, L. mono = *Listeria monocytogenes.* Surface = reported a "surface" sample from unspecified depth, Multiple = samples collected from epilimnion, thermocline, and hypolimnion of multiple sites but specific depths were not listed. N.A. = no information available.

While spatial variability of microbial water quality metrics is largely under-reported, even less is known about variations in concentrations at different water depths. Concentration changes with depth can be expected to exist due factors such as the development of thermal and/or dissolved oxygen gradients, closer proximity to sediments, and the reduced effects of solar radiation on FIB inactivation. Kleinheinz et al., (2006) reported significantly different concentrations of *E. coli* in lake waters among sampling depths of 30, 60, and 120 cm. Hosetti and Patil (1987) observed maximum bacteria counts in bottom layers of wastewater ponds and related this observation to the depth-dependent light penetration. In a storm water pond, He et al. (2015) documented uniformly distributed levels of *E. coli* and fecal coliforms up to the 1.2-1.5 m depths with increased levels below this layer. The authors also concluded that thermal gradients inhibited the vertical movement of microorganisms and that stratification and de-stratification of bacterial concentrations followed diurnal cycles. Antaki et al., (2016) sampled two irrigation ponds in Southern Georgia and reported conflicting results. In one pond the levels of *E. coli* and *Salmonella* were roughly 31 and 4 times greater, respectively, at a 1-m depth compared to the surface water. At the second pond, *E. coli* and *Salmonella* were 0.5 and 6 times lower, respectively, at the surface than at the 1-meter depth. Partyka et al. (2018) measured *E. coli,* fecal coliforms, *Salmonella*, STEC, and *E. coli* O157:H7 at six California reservoirs and at 3 different depths which included the epilimnion, the thermocline, and the hypolimnion. The authors reported overall no significant differences by water depth when averaged across reservoirs, but two reservoirs did have significantly higher *E. coli* and fecal coliform concentrations within the epilimnion layers

compared to both the thermocline and hypolimnion layers. The authors also found a trend of higher detection of all pathogens with increased water depths. All in all, the spatial variability of FIB and pathogens in irrigation sources is largely unknown due to measurements performed in only a single sampling location. This shortcoming is sometimes acknowledged in studies such as that of Cooley et al. (2007) who stated that statistical analysis of pathogen data could not be performed due to low sample numbers.

2.4.3 Temporal variability of microbial water quality metrics

Temporal variability of *E. coli* concentrations has been well documented in stream waters and at beaches (U.S. EPA, 2010). Less information is available for irrigation water sources. During a 19-month study, Lee et al., (2018) measured *E. coli* and *Salmonella* in 5 different irrigation ponds in Georgia. The authors reported that *E. coli* concentrations peaked with peak temperatures in early- to mid-summer, however, *Salmonella* concentrations peaked in the late summer or early fall. Durham et al., (2016) studied the level of *E. coli* in 6 small urban lakes in Texas and found that 5 of 6 lakes contained higher concentrations in the summer than the winter. Results from their 36-month study also demonstrated substantial inter-annual differences in *E. coli* in all studied ponds which was attributed to varying water temperatures and the number of geese present each year (i.e. correspondingly the quantity of geese feces). Similarly, Jokinen et al. (2019) documented significant inter-annual differences in *E. coli* and *campylobacter* in an irrigation reservoir. In year 1 of their study, microbial water quality exceedances were attributed to runoff from surrounding agricultural

lands which occurred earlier in the season whereas in year 2 most of the exceedances occurred later in the season when there was little precipitation and were attributed to feces from wild birds and cattle.

It is important to point out that there is a strong effect of seasonality on the incidences of foodborne outbreaks. A meta-analysis on outbreaks in the U.S. between 1996 and 2017 compared the seasonality in reported cases and found that the majority were tightly clustered in late-July at both the state and national level (Simpson et al., 2020). Another study that documented seasonal changes in outbreaks reported that 50% of outbreaks were associated with vegetable row crops, fruits, and seeded vegetables and occurred during the April to July periods with a resurgence in October (Bennet et al., 2018). Similarly, in a systematic review of the cases of campylobacteriosis, salmonelellosis, and vero cytotoxin-related illnesses between 1960 and 2010, it was shown that in general there was increased prevalence of these bacteria-related diseases in the summer months while cryptosporidiosis and giardiasis, the only parasite-related illnesses in the study, showed peaks in cooler seasons such as the spring and fall (Lal et al., 2012). The authors attributed the observations to seasonal land use patterns between animals and humans and changes in temperature and precipitation throughout the year. The higher prevalence of pathogens in the summer months corresponds with the growing season and when irrigation withdrawals are the greatest which may explain the larger number of illnesses reported in this season.

The time of day an irrigation source is sampled may also have significant impacts on the results of microbial water quality monitoring. Strong diurnal changes

occur in FIB concentrations, whereby concentrations are highest during the evening or early morning and lowest in the afternoon, and have been documented in streams, canals, and coastal environments (Boehm et al., 2002; U.S. EPA, 2010; Stocker et al., 2016; Lothrop et al., 2018), however, these findings have not been investigated for small irrigation ponds. Brissaud et al., (2002) studied the level of fecal coliforms in a stabilization pond and reported diurnal fluctuations between mixed and stratified conditions which substantially affected fecal coliform populations measured at the pond outlet. In their study, concentrations were highest in the 6:00 to 10:00 am period and began to decline from 10:00 am to 14:00 pm. The primary drivers of mixing and stratification were reported to be daily temperature variations and density fields as well as wind friction at the surface of the pond. Table 2.1 shows that the time of day in which samples were taken is largely unreported in irrigation microbial water quality studies. Where documented, most researchers report taking samples in the morning hours prior to solar noon (Topalcengiz et al., 2017; Truit et al., 2018). Without accounting for diurnal variability of indicators in irrigation water quality, results from study to study or even within the same study may inadvertently over- or under-estimate the true risk of using an irrigation water source.

The greatest temporal differences in indicator densities are attributed to precipitation events. Jenkins et al. (2015) documented two orders of magnitude increase in FIB in three ponds during stormflow events observed in their study. These increased levels were attributed to runoff and sediment disturbances. These authors also demonstrated that it took about 5 days for levels of *E. coli* and enterococci to return to pre-storm levels. Interestingly, the levels of *Salmonella* and *E. coli* O157

remained significantly higher 5 days after the event as compared with the pre-event levels. Harris et al. (2017) measured the levels of *E. coli* and *Salmonella* before and after rainfall events in two irrigation ponds in GA and reported significantly higher concentrations of both organisms after precipitation events which agrees with the results of several other studies that involved measurements conducted during precipitation events (Steele et al., 2005; Chen and Chang, 2014; Kleinheinz et al., 2010; Stocker et al., 2016)

It is important to note that when creating Table 2.1 it was discovered that in many publications there was insufficient information to characterize the monitoring designs. Many reports did not provide the number of samples, the number of sampling locations, the time of day, and/or the depth at which samples were taken (Strawn et al., 2013; Marine et al., 2015; Benjamin et al., 2015; Partyka et al., 2016; Pagadala et al., 2015; Gorksi et al., 2011; Cooley et al., 2014; Cooley et al., 2007). Extracting pertinent spatial and temporal information from these studies was also made difficult or impossible due to researchers lumping irrigation sources together (ponds, streams, creeks, reservoirs) without differentiating the characteristics between them.

2.5 Modeling FIB and Pathogens Concentrations

2.5.1 Physicochemical, meteorological, and hydrological covariates

The measurement of *E. coli* in surface waters is currently limited by culturing methods which require 18 to 24 hours before enumeration can begin. Similarly,

determining pathogen densities is also laborious and often takes several days for results. Due to such delays, the U.S. EPA has recommended the use of predictive models to improve the timeliness and accuracy of microbial water quality assessments (U.S. EPA 2012; Francy et al., 2013; Francy et al., 2020). These models utilize environmental variables with expected mechanistic relationships with the target organism being predicted. Table 2.2 shows a summary of the environmental covariates used in the studies reported in Table 2.1. Examination of Table 2.2 shows that water temperature, pH, specific conductance, turbidity, air temperature, and precipitation were the most commonly reported covariates measured in these studies.

Table 2.2. Summary of environmental covariates measured across studies presented in Table 2.1.

Parameter	Reference	Number of Appearances	Percent (%)
Water Temperature (°C)	1,2,3,4,5,6,7,8,10,12,13,14,17	12	60
pH	1,2,3,4,5,6,7,8,10,14,17	10	50
SPC (uS cm^{-1})	1,2,3,4,5,6,7,8,10	8	40
Turbidity (NTU)	1,2,3,4,5,6,7,8,10	8	40
Air Temperature (°C)	1,2,3,4,6,8,9,14	7	35
Precipitation[a]	2,3,6,7,9,13,16,17	7	35
Tcol (CFU100 mL^{-1})	1,2,3,7,9,13,16	6	30
ORP (mV)	1,2,4,5,7,10	5	25
DO (mg L^{-1})	1,6,7,8,10,17	5	25
TSS (mg L^{-1})	4,5,11,17	3	15
NO_3 (mg L^{-1})	4,10,17	2	10
Ent (CFU 100 mL^{-1})	2,3,13	2	10
APC (CFU100 mL^{-1})	1,3,9	2	10
Fcol (CFU100 mL^{-1})	3,4,13	2	10
Salinity (%)	10,17	1	5
OP (mg L^{-1})	4,17	1	5
Solar radiation (W m^{-2})	1,12	1	5
Humidity (mmHg)	12	1	5
NH_3 (mg L^{-1})	17	1	5
Chloride (mg L^{-1})	10	1	5
Wind Speed (mph)	12	1	5
Wind Direction	12	1	5
DOC (mg L^{-1})	4	1	5
DN (mg L^{-1})	4	1	5
Farming system (Org/Conv)	15	1	5
Growing Season (Spring/Fall)	15	1	5
Within Season (start/middle/end)	15	1	5
Region (MD,NJ,DE)	15	1	5
Number of Geese	17	1	5

a = amount or days since last precipitation event, SPC = specific conductance, Tcol = total coliforms, ORP = oxidation reduction potential, DO = dissolved oxygen, TSS = total suspended solids, NO_3 = nitrate, Ent = enterococci, APC = aerobic plate count, Fcol = fecal coliforms, OP = orthophosphate, NH_3 = ammonia, DOC = dissolved organic carbon, DN = dissolved nitrogen. References: [1]McEgan et al. (2013), [2]Topalcengiz et al., (2017),[3]Draper et al., (2016), [4]Gu et al., (2013), [5]Lee et al., (2018),[6]Antaki et al. (2016),[7]Gu et al. (2021),[8]Partyka et al. (2018),[9]Truitt et al., (2018),[10]Solaiman et al. (2020),[11]Harris et al. (2018),[12]Jokinen et al. (2019),[13]Steele et al. (2005),[14]Pahl et al. (2013),[15]Marine et al. (2015),[16]Won et al. (2013),[17]Durham et al. (2016),[18]Benjamin et al. (2014),[19]Micalleff et al. (2012),[20]Delbeke et al. (2015).

Total coliforms and other microbial populations such as enterococci or total heterotrophic bacteria, ORP, dissolved oxygen, salinity, and nutrient concentrations were also measured in multiple studies to assist in the prediction of *E. coli* and pathogen densities. The choice of parameter to include in a study is often site-specific. For example, studies focusing on microbial water quality at freshwater beaches often include day of the year, wave height, antecedent rainfall, wind speed and direction, stream inflow, solar radiation, tide stage, and changes in water levels (Francy et al., 2020) whereas some of these variables are not measurable in other systems (e.g. rivers or ponds). In general, exploratory variables in microbial water quality studies can be divided into meteorological (e.g. precipitation, radiation, wind), hydrological (e.g. river flow, sea level, or tide), and direct water quality measurements such as turbidity, temperature, salinity, pH, nutrients, and past microbial concentrations (Sokolova et al., 2022; Francy et al., 2020). Several studies considered geographical, physical, and remotely sensed covariates such as region, elevation, basin/channel slope, proximity to cropland, pastures, impervious surfaces, and composition of soils and sediments in the waterbody or surrounding areas (Laureano-Rosario et al., 2019; Weller et al., 2021; Álvarez-Cabria et al., 2016; Woo, et al., 2019).

2.5.2 Conventional Statistical Analyses and Microbial Water Quality

The levels of environmental covariates utilized in microbial water quality studies are reported and related to the concentrations of the target microorganism using statistical models. This procedure is usually followed because these parameters

effectively describe conditions that control the habitat and fate and transport processes of the microorganism. In the 20 studies summarized in Table 2.1, 11 of 20 used correlation analysis, 5 of 20 used linear regression analysis, 4 of 20 used multiple linear regression, 7 of 20 used binary logistic regression, 1 of 20 used canonical correspondence analysis, 1 used principal component analysis, and 2 studies did not perform statistical analysis. Strong associations between environmental variables and target microorganisms were rare ($r;R^2;$or $r_s < 0.4$), but did show statistical significance as well as consistency in the direction of the relationships in several studies. For example, Draper et al. (2016), Solaiman et al. (2020), and Topalcengiz et al. (2017) reported significant negative relationships between *E. coli* and pH, however the strengths of these relationships were weak with correlation coefficients of -0.354, -0.300, and -0.198, respectively. Several authors reported relationships between precipitation and fecal microorganism concentrations, but the strongest association was only $R^2 = 0.395$ (Topalcengiz et al., 2017). In the 20 studies in Table 2.1, only a couple reported strong relationships. Solaiman et al. (2020) reported a strong correlation between the concentrations of *Aeromonas* spp. and nitrate in pond waters ($r = 0.800$) and Partyka et al. (2018) found a moderately strong relationships ($R^2 = 0.530$) between *E. coli* concentrations and geographical and chemical variables in reservoir waters using multiple linear regression. The strongest relationships in the 20 summarized studies were those reported in Durham et al. (2016) who used stepwise multiple linear regression for variable selection to predict *E. coli* concentrations in both the summer and winter in 6 different Texas lakes. While the R^2 values for the individual lakes were as low as 0.174, the authors

reported strong (0.806) to very strong (0.924) values with certain combinations of variables for some of the lakes. The strongest association (0.924) utilized geese counts, total solids, precipitation, pH, ammonia, and nitrate concentrations to predict *E. coli* concentrations. The average R^2 value across seasons and lakes in their study was 0.522. While the regressions presented in Durham et al. (2016) for lake waters showed strong results, the majority of the remaining studies did not find strong associations between target microorganisms and environmental parameters.

It is likely that conventional statistical approaches that rely on linear combinations of predictors are not well suited for microbial water quality studies which contain complex and non-linear relationships between inputs and the output. These complex relationships may result from subsets of data in which the directionality and strength and/or reported relationships may be different than the values for other subsets. For example, Bradshaw et al. (2016) reported varying efficacy of *E. coli* as an indicator of *Salmonella* concentrations when dissolved oxygens levels were < 11.3 mg L^{-1}, pH was < 6.65, and temperature was ≥ 13 °C, but outside the specified environmental conditions these parameters were not useful in modeling the relationships between the studied fecal microorganisms. Davies-Colley et al., (2000) reported that *E. coli* die-off in pond water was greatly increased when pH was above 8.5 as compared to more moderate levels of 7 to 8.5. The authors also noted an interdependence of die-off rates on dissolved oxygen, pH, and sunlight levels with separate threshold values above and below which die-off rates would change. Weller et al. (2020) reported polynomial relationships between the detection of shigatoxins (*eaeA-stx*) and air temperatures in New York streams and Arizona

irrigation canals. Whitman et al. (2004) reported exponential decline of *E. coli* concentrations with increased solar insolation within Great Lake waters. In another study, it was reported that *E. coli* concentrations agricultural streams displayed non-linear relationships with pastured land whereby if this land use type exceeded a threshold value of 50% of a studied area, there was a significant increase in mean concentrations (Scott et al., 2017).

Overall, relationships between target microorganisms or their markers and environmental covariates in microbial water quality studies have been shown to be complex and often non-linear whereby threshold values determine the sign of the relationship. Additionally, the dependencies of microbial concentrations on water quality covariates can be different or unique to certain clusters of data within dataset. Machine learning (particularly regression trees and random forest algorithms) allows us to discern these domains which may contain unique dependencies of microbial concentrations on water quality covariates. Elucidating these covariate subspaces would be an arduous task to complete with conventional regression methods.

2.5.3 Machine learning and Microbial Water Quality

Machine learning (ML) is a subsection of artificial intelligence (AI) within the field of computer science which revolves around creating algorithms that can solve complex problems with little to no human intervention. These algorithms are provided training datasets in which to learn the associations between the predictor variables (independent variables/inputs) and the output (or target/dependent variable) without any assumptions about the functional forms of the relationships. Therefore,

classic statistical assumptions including normality and homoscedasticity of data or linearity of relationships are not required (Quetglas et al., 2011). Once models are developed with the algorithms, some portion of the total dataset, which the algorithm was not trained on, is used for testing/validation. The results of predictions or classifications show how well the models generalize on unseen data and therefore provide insight into the future performance of the model.

Model evaluation differs depending on whether the ML algorithm is being used to solve regression or classification problems. For regression, ML algorithms are evaluated based on the absolute or relative errors computed with observed and predicted values. Common metrics include root-mean-squared error (RMSE), mean absolute error (MAE), percentage error (MAPE), and the coefficient of determination (R^2). Classification tasks do not use these metrics, but rather common results include those reporting model selectivity and sensitivity, classification accuracy and precision, and recall, among others. Classification metrics focus on the number of true and false positives and negatives and analysis of confusion/error matrices.

Popular ML algorithms include random forest (RF), support vector machines (SVM), k-nearest neighbor (kNN), gradient boosting machines (GBM), and artificial neural networks (ANN), among others (Kuhn and Johnson, 2013). These algorithms are fundamentally different in how they operate and produce results. For example, the RF algorithm produces hundreds to thousands of regression trees which all produce individual predictions which are then averaged to get a final prediction. The kNN algorithm makes new predictions based on similarities between characteristics of the to-be-predicted value with other values in the dataset (known as neighbors). It is also

quite common for variants of each algorithm to appear within studies as well and comparisons to baseline models are usually performed. These variants (e.g. conditional random forests, weighted kNN, and linear or polynomial SVM kernels, etc.) are created with similar foundations as the original algorithms, but usually include some modifications which may be preferable in certain circumstances (Weller et al., 2021). For example, if a dataset contains a high-degree of linear relationships in the data, then a researcher may discover that the linear SVM is more suitable than one designed to handle non-linearity like the polynomial or radial basis function kernels which are both more computationally intensive as they use functions to transform the original feature space into a higher dimensional space (Burkov, 2019). The selection of the optimal algorithm is largely dataset dependent and researchers must often conduct some exploratory analysis to make the determination of which algorithms to utilize in their studies.

Machine learning has been gaining popularity within the field of microbial water quality as numerous studies have now demonstrated the powerful capabilities of these algorithms. Table 2.3 lists 20 studies published between 2015 and 2022 which have successfully utilized ML to create models which are either used to predict specific concentrations (regression) or classify ranges (e.g. low, medium, high or compliant/noncompliant) of FIB and other foodborne pathogens. The performance of these models is typically very good especially when compared to results gained through classic statistical approaches (Quetglas et al., 2011). In microbial water quality studies which have evaluated conventional statistical models alongside ML models using the same datasets, ML models have been shown to consistently

outperform the former (Mas and Ahlfeld, 2007; Motamarri and Bocelli, 2012; Thoe et al., 2012; Thoe et al., 2014; Stidson et al., 2012; Avila et al., 2018; Panidhapu et al., 2020).

Table 2.3. Results of 20 recent microbial water quality studies which utilized machine learning methodology.

Target Organism	Water Type	Predictors	Models	Best model	Metric	Ideal Value	Performance	Reference
EC	Agricultural streams	AD, GP, HYD, MET	ANN, ANFIS, FCM, SVM, RT, RF	ANFIS	R^2	1	0.962	Abimbola et al., 2020
S. spp., EHEC	Irrigation canals	GEOG, LU, PHYS, MET, HYD	cRF	-	AUC	1	0.84 for S. spp. 0.92 for EHEC	Belias et al., 2021
EC	Beaches	MET, PHYS	14 models	GBM	AUC	1	0.76	Brooks et al., 2016
S. spp.	Irrigation Ponds	PHYS, MET, MIC	RF, SVM, NB, ANN	ANN + RF*	AUC ; ACC	1 ; 100%	0.98 ; 94.9%	Buyrukoglu et al., 2021
TC	River	MET,HYD	ANN	-	R^2	1	0.56 - 0.87	Choi et al., 2018
EC	Estuaries	MET, HYD	ANN	-	R^2	1	0.83	Garcia-Alba et al., 2019
EC,EN	Beaches	MET, HYD	cB, XgB, RF, SVM, ANN	cB	R^2	1	0.71 for EC 0.68 for EN	Grbčić et al. (2021)
EC FC	River	PHYS,MET	BBN, NB, RF	BBN	ACC	100%	76% to 95.5%	Panidhapu et al., 2020
EC,EN	Beaches	MET, HYD	ANN, SVM	ANN	R^2	1%	0.332 - 0.962	Park et al., 2018
S. spp.	Irrigation Ponds	PHYS	kNN, SVM, ANN	kNN	ACC	100%	52 - 77%	Polat et al. (2020)
EN	Beaches	RS	ANN	-	ACC	100%	76%	Laureano- Rosario et al 2019
EC	Beaches	MET, HYD	GLS, ANN, RF	RF	R^2	1	0.95	Searcy and Boehm, (2021)
EC,TC	Drinking Supply	PHYS,MET,HYD	LASSO, RF,	TPOT	R^2	1	0.86	Sokolova et al. (2022)

			TPOT, VAR					
EC	Irrigation Ponds	PHYS	RT	-	R^2	1	0.676 - 0.926	Stocker et al., 2019
EC	Irrigation Canals	PHYS, MET, SED	SVM, BLR, RC	SVM	TPR ; TNR	100%	93 - 100 % 88.3 - 95.7%	Tousi et al., 2021
FIB	River	PHYS, MET, SO	ANN	-	ACC	100%	86%	Vijayashanthar et al. (2018)
EC	Urban Streams	GEOG, LU, MET	RF	-	R^2	1	0.46 - 0.50	Wang et al., 2021a
EC	Beaches	MET, HYD	MLR, PLS, SPLS, RF, BN	RF	ACC	100%	78 to 82.3 %	Wang et al., (2021b)
EC	Agricultural Streams	GEOG, LU, PHYS, MET, HYD	26 ML models	b-kNN	R^2	1	0.47	Weller et al. 2020a
L. spp., S. spp., EC	Agricultural Streams and Canals	GEOG, LU, PHYS, MET, HYD	cRF	-	R^2	1	0.72 for AZ 0.45 for NY	Weller et al. 2020b

MET = meteorological variables, HYD = hydrological variables, LU = land use variables, GEOG = geographic variables, PHYS = physicochemical variables, RS = remotely sensed variables, AD = animal density, SO = proximity to sewage outfall, GP = proximity to grazing pastures, SED = sediment properties, MIC = microbial populations, R^2 = coefficient of determination, ACC = accuracy, TPR = true positive rate, TNR = true negative rate, S. spp. = *Salmonella* spp, L. spp. = *Listeria* spp., EC = generic *E. coli*, FIB = fecal indicator bacteria (multiple) , EHEC = enterohemorrhagic *E. coli*, FC = fecal coliforms, MLR = multiple linear regression, SVM = support vector machine, RF = random forest, BN = Bayesian networks, cRF = conditional random forest, BLR = binary logistic regression, RT = regression trees, ANN = artificial neural networks, NB, naïve Bayes, LASSO = least absolute shrinkage and selection operator, GAMMLS = general additive model for location and shape, VAR = vector auto regression, XgB = 'xtreme' gradient boosting machines, RC = ridge classifier, b-kNN = boosted k-nearest neighbors, TPOT = tree-based pipeline optimization tool, SPLS = sparse partial least squares, BBN = Bayesian belief networks, GLS = generalized least squares, ANFIS = adaptive neuro-fuzzy inference systems, FCM = fuzzy c-means.

Explanations for better performance of the ML algorithms as compared to conventional statistical approaches are numerous, but may be attributed largely to the complexity of aquatic environments. The relationships between predictors and the target variable (e.g. *E. coli* concentration) are often non-linear or unknown, and datasets may contain subsets of data with specific direction and strengths of relationships, which if analyzed as a whole, may show inconsistent relationships or no associations between independent and dependent variables. In addition, aquatic environments can be incredibly complex and dynamic systems with many interconnected biotic and abiotic variables which may change in tandem (Strogatz, 2001; Pascual & Dunne, 2006). Aquatic datasets can be bulky, contain noisy data, redundancy, internal relations, and outliers (Park et al., 2003). Discerning the nature of these systems is therefore difficult for biologists and ecologists to describe with conventional statistics which are constrained by assumptions of the nature of the datasets (Quetglas et al., 2011). ML algorithms are flexible and do not make any assumptions regarding the nature of the dataset and are not restricted in abiding by a specific functional equation form. They are capable of fitting data automatically and in seconds even with very large datasets.

While numerous published studies have utilized ML within the field of microbial water quality, a much smaller set have been applied to irrigation water quality (Table 2.3). Within this subset, the majority of studies have applied ML to solve classification problems while only one study utilized ML to predict concentrations of *E. coli.* The study performed by Weller et al. (2020) predicted *E.*

coli concentrations in NY streams which could be used for irrigation using ML algorithms. The applicability of ML for use within microbial irrigation water quality studies has not been fully elucidated. More work is needed to both identify the best algorithms as well as investigate the accuracy of these algorithms for predicting FIB based on levels of readily sensed environmental parameters.

The relationships between target microorganisms or their markers and environmental covariates in microbial water quality studies have been generally shown to be complex and often non-linear whereby threshold values determine the sign and strength of the relationship. Additionally, the dependencies of microbial concentrations on water quality covariates can be different in separate domains of the covariate space. Machine learning and particularly regression trees and random forest algorithms allows us to discern these domains which may contain unique dependencies of microbial concentrations on water quality covariates. Elucidating these covariate domains would be an arduous task to complete with conventional regression methods.

2.6 Research Needs

A substantial volume of research has been previously conducted to study the prevalence and dynamics of FIB and pathogens in the environment waters. A smaller subset of research has been conducted to evaluate these issues with regard to the microbial quality of irrigation waters. However, several issues that are critical for the design and implementation of microbial monitoring as well as for the development of successful predictive modeling remain unexplored, thus hampering effective and efficient water quality determinations.

Since information regarding the spatial and temporal dynamics of FIB in irrigation water sources is extremely scarce, this literature review provided an up-to-date overview of the limited scientific information currently in print. The extent of the spatial variability in FIB concentrations across waterbodies is largely unknown which may have serious consequences for monitoring and risk assessment. Temporal dynamics are also rarely reported as is the time of day in which sampling occurred in the majority of studies. If concentrations can be expected to change throughout the day in response to environmental stressors then the time of day when sampling occurred may have substantial impacts on the results of water quality assessments. Therefore, it appears highly pertinent to evaluate both the spatial and temporal dynamics of FIB in waters used for irrigation. It is likely that the choice of where and when to sample should not be one made arbitrarily or based on where and when is most convenient. More work is needed to elucidate the spatiotemporal dynamics of FIB in order to improve monitoring design and risk assessment.

The majority of previous studies have drawn conclusions from samples collected only at the water surface which may not be representative of conditions at depths where the largest volumes of water are contained. Three-dimensional (3D) sampling of waterbodies used for irrigation is extremely uncommon. Yet, when irrigation is initiated, waters from deeper depths are what is withdrawn, at least initially before substantial mixing occurs, and therefore may be the most important waters to monitor. Knowledge on the 3D distributions of FIB in irrigation ponds may be especially useful for simulating the withdrawal of water and estimating the concentrations of FIB which will reach fields. Currently, the vast majority of hydrodynamic modeling studies involving ponds have focused on waste- or storm-water stabilization ponds which are designed to utilize temperature, sedimentation, sunlight, predation, and competition as factors to reduce populations of bacteria, viruses, and parasites while also increasing hydraulic retention time to allow sufficient time for these factors to reduce the quantity of microorganisms within the water (Dias et al., 2017). No such study has been conducted with regard to withdrawal of water for irrigation with implications for microbial water quality monitoring and food safety. Therefore, conducting 3D sampling of irrigation ponds for the use of hydrodynamic modeling will improve our knowledge on changes in microbial water quality throughout the duration of an irrigation event and provide insight into the concentration ranges that can be expected in the withdrawn water and by extension the water which reaches crops and soils.

Correlative and regression results in irrigation waters often report poor associations between environmental variables and densities of FIB or pathogens. The

complexity of aquatic habitats which are simultaneously affected by a multitude of factors responding in concert to changes in the system is a likely reason why conventional statistical approaches struggle in discerning relationships between inputs and outputs. Machine learning provides a powerful approach to elucidating the functional forms of these relationships and can also be used to evaluate which predictor variables are the most important for predictions and degree in which specific predictors affect the model error.

Machine learning is becoming commonly applied within microbial water quality studies, but most previous research studies have focused on flowing or coastal waters (Table 2.3). Results from irrigation water quality studies may be expected to differ depending on the water source and if the irrigation water is from a lotic (e.g. stream, river, coastal) or lentic (e.g. pond, lake, reservoir) source. Differences in the most important predictor variables will likely also differ between waterbody types. For example, water velocity would likely not be an important factor in lakes while thermal stratification would not be an important factor for streams or creeks. Therefore, it appears prudent to evaluate the applicability of ML algorithms to make predictions on the microbial quality of water in ponds which are commonly used to provide irrigation water. Additionally, the only ML studies which involved irrigation ponds and canals have been conducted in the South West and South East regions of the United States. No such studies have been conducted in the Mid-Atlantic region which may contain unique relationships between predictors and outputs due to differing environmental conditions such as temperature and solar radiation.

Chapter 3: Persistent Patterns of *E. coli* Concentrations in Two Irrigation Ponds from 3 Years of Monitoring

3.1 Introduction

The microbial quality of water used for irrigation is a critical component of environmental quality that should be monitored to reduce the incidences of food-borne outbreaks associated with fecal contamination. The linkages between microbial water quality and public health incidents have been documented in the US and elsewhere worldwide (Pachepsky et al., 2011; Uyttendaele et al., 2015; Jongman and Korsten, 2016; Allende et al., 2018). Due to the difficulty in the detection and enumeration of pathogenic microorganisms in the environment, the levels of generic Escherichia coli (*E. coli*) are measured and compared to regulatory standards for microbial water quality evaluation. Currently, the Food and Drug Administration sets the allowable levels of *E. coli* in irrigation waters in the U.S. under the Food Safety Modernization Act (FSMA) (FDA, 2016) and *E. coli* is also used as a microbial water quality indicator internationally as well (Uyttendaele et al., 2015; Allende et al., 2018).

The initial guidance provided by the FDA for the evaluation of microbial water quality of irrigation source waters does not provide information regarding the selection of water sampling location(s). The dynamics of *E. coli* concentrations in surface waters have been shown to display substantial spatial and temporal variability (USEPA, 2010; Quilliam et al., 2011; Jeon et al., 2020). An et al. (2002) reported that *E. coli* densities on an inland lake displayed large spatial variation with high

concentrations measured in areas with heavy recreational activity (swimming or motor boating) and where waterfowl tended to congregate. Increased levels of both Salmonella and *E. coli* O157 have been observed with decreased distance from rangeland in wells, ponds, drainage ditches, standing waters, and reservoirs used for irrigation (Benjamin et al., 2013). Other researchers have documented high indicator organism concentrations in areas where streams or rivers enter lakes and lower concentrations when moving outward from these locations (Brookes et al., 2004; Davis et al., 2005). Therefore, the decision on where and when to sample may have significant effects on the results of microbial water quality monitoring assessments.

Datasets resulting from microbial water quality assessments of irrigation water sources are somewhat limited and have usually focused on the temporal component over the spatial one (Draper et al., 2016; Havelaar et al., 2017; Lee et al., 2018; Topalcengiz et al., 2017). In these studies, samples are collected from as close to the irrigation intake as possible which is considered the "golden standard" (Lee et al., 2018), or from the sprinkler systems where the water is distributed (Marine et al., 2015; Pagadala et al., 2015). Many studies exist that document *E. coli* concentrations in irrigation sources such as ponds, lakes, and reservoirs, but rarely are samples taken across the water body due to limited available resources, difficulties in accessing potential sampling locations and/or due to the intended goals of the research.

Denser spatial sampling yields better information on the spatial distribution of concentrations that may be present. Lee et al. (2018) documented large spatial differences of *E. coli* and Salmonella concentrations among 3 locations around the perimeter of an irrigation pond measured from March to September. The authors

related the spatial differences to potential microhabitats existing within the pond along with the physicochemical parameter preferences of the microorganisms. They also found the samples taken near the intake (pond interior) contained lower concentrations of *E. coli* than those sampled at the bank. Conversely, Harris et al. (2018) reported higher *E. coli* concentrations in interior sampling locations as compared to bank locations in two GA ponds. Pagadala et al. (2015) measured total aerobic bacteria, *E. coli*, and total coliforms concentrations in irrigation wells, pond water samples, and at the end of drip lines, but found no significant differences among the concentrations of all three groups. In another study, tenfold or greater differences were observed between pond water at the source and what is released at the end of drip lines or pivot sprinkler heads (Antaki et al., 2016). Another group of authors sampled 5 locations along the flow path of 6 different California reservoirs and found that the microbial water quality in downstream irrigation supplies was not statistically associated with conditions within the reservoirs, rather there was a greater similarity to water quality just upstream of the reservoirs (Partyka et al., 2018).

The need for monitoring environmental variables across spatial units with substantial internal heterogeneity is quite common across environmental disciplines. One feature helpful in monitoring design is the existence of persistent spatial patterns in spatiotemporal variations. Some areas exhibit environmental variable values that tend to be lower than the average across the study area. On the contrary, other areas have values of the same variable that tend to be higher than the average. Such phenomenon is well known in hydrology and atmospheric sciences and is referred to

as temporal stability (Cosh et al., 2008; Vanderlinden et al., 2012; Vereecken et al., 2016).

One approach to quantify the temporal stability is to use the mean-relative-difference (MRD) analysis which compares the relative differences between the point measurements and averages across the study area. The analysis returns information on the consistency, or stability, of a measurement taken in each location relative to the averaged measurements in all locations within the sampling extent. In this way, locations with values consistently lower, approximately equivalent, and higher than the spatiotemporal average can be determined. The MRD analysis has been successfully applied in the analysis of across- the-field soil moisture contents (Vachaud et al., 1985), soil moisture, and electrical conductivity (Cosh et al., 2008; Pedrera-Parrilla et al., 2017) as well as crop yields (Huang et al., 2018). Jeon et al. (2020) used the MRD analysis to discern temporal stability of *E. coli* concentrations in a PA creek which showed dependency on the sampling season and land use. Pachepsky et al. (2018) documented temporal stability in *E. coli* concentrations over a single growing season in two irrigation ponds in Maryland.

Another powerful approach to elucidating spatial patterns is the Empirical Orthogonal Function (EOF) analysis. This method was successfully applied to spatiotemporal variations found for various environmental variables, e.g., rainfall (Saravan et al., 2020), soil moisture (Perry and Niemann, 2007; Korres et al., 2010), snow cover thickness (Pan et al.,2017), crop yield (Kim et al., 2020), phytoplankton community structure (Liu et al., 2021), etc. Pattern analysis with EOFs typically concludes with identifying natural factors responsible for each pattern (Lee et al.,

2021). The EOF analysis was applied to study *E. coli* concentrations in a large creek in Pennsylvania (Jeon et al., 2019) and the patterns found were related to land-use change and tributary effects. The benefit to performing the EOF analysis is that not only is the spatial variation identified but the temporal component is as well in the form of expansion coefficients (ECs) and EOFs and ECs can be analyzed separately. Additionally, while the MRD method provides a single spatiotemporal pattern, the EOF analysis can identify multiple patterns within datasets which each contribute to some percentage of the overall variability in the system.

Although irrigation ponds and reservoirs provide a substantial fraction of irrigation water in the world, neither the MRD or EOF analysis has been applied to study *E. coli* concentrations in these surface water sources. Knowledge of spatial and temporal patterns in *E. coli* concentrations in irrigation sources can improve monitoring design by characterizing the variability across a waterbody which will help identify representative sampling locations.

The objective of this work was to 1) apply and compare two distinct pattern recognition analyses to characterize and compare the spatiotemporal variability of *E. coli* concentrations in irrigation ponds using single- and multi-year data, and 2) to test the hypothesis that the patterns of *E. coli* correlate and may be predicted using the patterns of more easily measurable water quality variables.

3.2. Materials and Methods

3.2.1 Data Collection

The field sites and the data acquisition are described earlier by Stocker et al. (2019). In brief, two working irrigation ponds (Fig. 1) were sampled during the growing seasons of 2016 − 2018 at the locations indicated in the site maps. Irrigation water is withdrawn in locations 6 and 4 from ponds P1 and P2, respectively. No animal manure is applied to the surrounding lands of either property. Ponds are mostly rain- and runoff-fed. P1 receives some water from an ephemeral creek at location 23 and from another pond at location 6 when the water level is low. P2 receives water via the culvert at location 12 during rainfall events and has level-dependent outflows between locations 24 and 25. Animal inter- actions with the ponds are controlled by the farm managers, but P1 has been observed to occasion-ally shelter small flocks of geese. Location 9 at P1 is used as an entry point by the landowners for recreational activities. There is a permanent residence at the South-Western part of P2 which is equipped with a septic system and domesticated dogs are free to wander the area near the pond. The two ponds are separated by a distance of roughly 130 km.

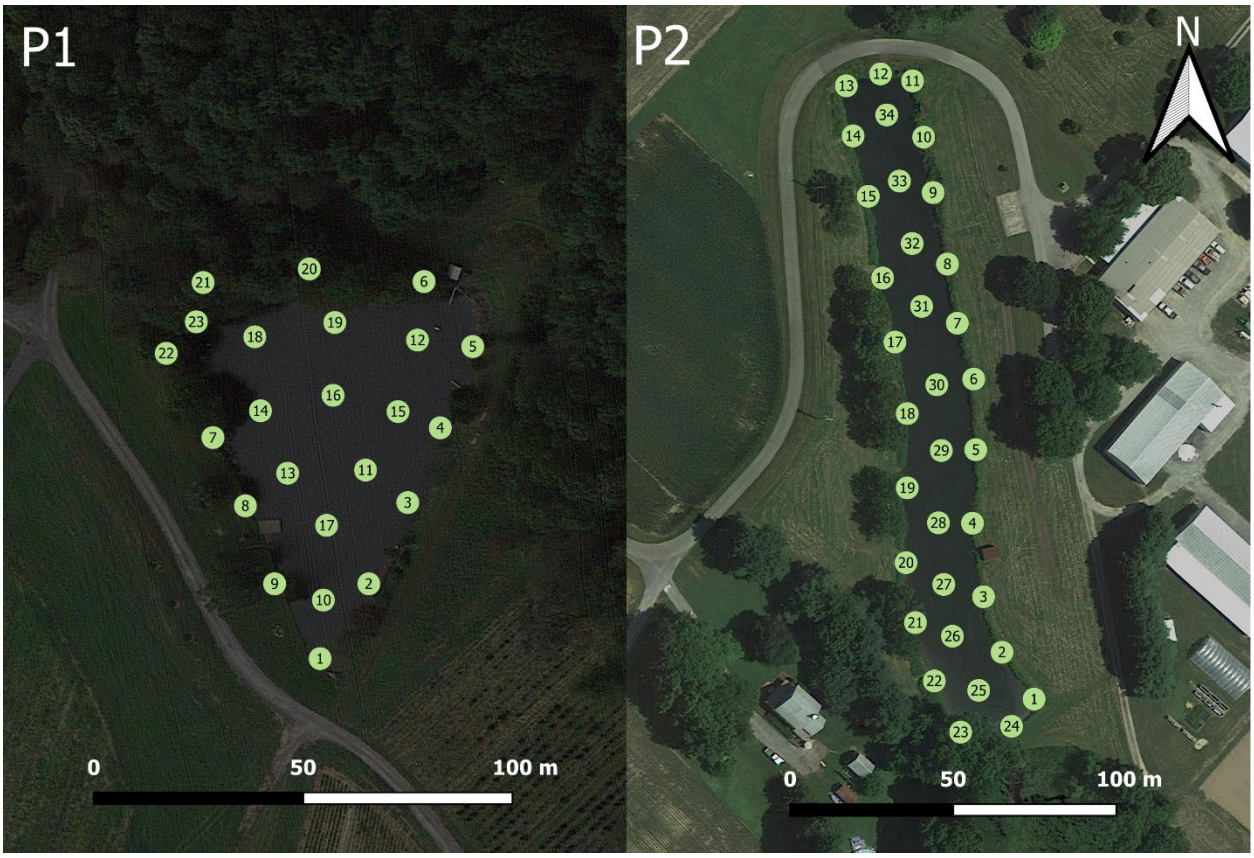


Figure 3.1. Sampling locations used for the 2016 – 2018 observation period. P1 and P2 have 23 and 34 locations, respectively. Locations 20-23 in P1 were not sampled in 2016. In 2016, sites 16-19 in P1 were added after the 2nd sampling date. Pond P2 had locations 31-34 added after the first sampling date in 2016.

Sampling was performed approximately biweekly primarily in June, July, and August. Care was taken to avoid sampling less than 2 days after a rainfall event. In 2018, the increased occurrence of rainfall led to larger than biweekly gaps between some sampling events. Nevertheless, in the two instances indicated below, sampling occurred less than 8 h after heavy rainfall events. Weather data, including air temperature and precipitation, were collected from on-site weather stations. Sampling on all dates started at 9:00 am and ended around 10:30 am. Samples were collected from the 0 to 20 cm depth interval by hand using a multipurpose sampling platform (Kim et al., 2020), a small boat, or in the case of nearshore samples, using a 500-mL

sampling rod. All samples collected were immediately placed in ice-packed insulated bags and then in a dark cooler on ice for transport. The coolers were brought to the laboratory within 2 h after sampling concluded.

Concurrently with sampling, a YSI 556 MPS (Xylem, Yellow Springs, Ohio) meter was used to measure temperature (C;°C), specific conductivity (SPC; µS cm^{-1}), dissolved oxygen (DO; mg L^{-1}), and pH in 2016. Turbidity (NTU;NTU) was measured back in the laboratory with a Lamotte 1799 turbidity meter (Lamotte, Chestertown, Maryland). In 2017 and 2018, a YSI EXO-2 sonde meter was used to measure the same set of parameters within the 0–20 cm sampling depth range but was also equipped with the sensors to measure turbidit, fluorescent dissolved organic matter (FDOM; µg L^{-1}), phycocyanin (PC; relative fluorescent units (RFU)), and chlorophyll-a (CHL; RFU). *E. coli* was enumerated based on EPA method 1603 (USEPA, 2014) which utilizes membrane filtration. The filtered volume was 50 to 150 mL of pond water. Each sample was filtered and plated in duplicate and the CFU counts were then averaged before reporting.

3.2.2 Mean Relative Difference (MRD) Analysis

The MRD analysis is a method used to characterize temporal stability of target measurements across space and time and utilizes relative differences in concentrations as opposed to absolute concentrations. The relative difference RD_{ij} between x_{ij} which is the observation of a variable x at location i at time j, and the spatial average of x at the same time $\langle x \rangle_j$ is defined as:

$$RD_{ij} = \frac{x_{ij} - \langle x \rangle_j}{\langle x \rangle_j} \quad (1)$$

The MRD for location *i* becomes:

$$MRD_i = \frac{1}{N_t}\sum_{j=1}^{j=N_t} RD_{ij} \quad (2)$$

where N_t is the number of observation times and i = 1,2, ….N_i, where N_i is the total number of locations.

The standard error $SERD_i$ of the set $RD_{i,1}$, $RD_{i,2}$….$RD_{i,\,Nt}$ of the relative differences at location *i* over the observation period is computed along with the MRD_i as follows:

$$SERD_i = \frac{\sqrt{\frac{1}{N_t - 1}\sum_{j=1}^{N_t}(RD_{ij} - MRD_i)^2}}{N_i{}^{0.5}}$$

The SERD serves as a metric to evaluate the degree of temporal stability present at a specific location. If $SERD_i$ is large, this value indicates that the MRD_i in location *i* is not stable. Conversely, if the value is low, this result indicates a strong temporal stability of the *E. coli* concentrations.

3.2.3 Empirical Orthogonal Function (EOF) Analysis

The *E. coli* concentration datasets were analyzed using Empirical Orthogonal Functions (EOFs). EOF analysis has proven to be a powerful tool to discern spatial patterns in spatiotemporall datasets obtained by measuring a single scalar variable in space and time (Navarra and Simomcini, 2010).

The EOF analysis describes patterns not in variations of the target variable per se but in variations of its “anomalies” defined as deviations of the target variable in measurement locations from the average across all locations at the same time. If measurements of the target variable *Y* were done in *m* locations *n* times, then the anomaly is

$$\Delta(m,n) = Y(m,n) - \bar{Y}(n) \quad (1)$$

Where m=1,2,3,…,M, and n=1,2,3,…, N, are locations and times of measurements, respectively, and $\bar{Y}(n)$ is the average value of $Y(m,n)$ across all locations m=1,2,3,…, M at the time n.

The EOF analysis presents the anomalies as sums:

$$\Delta(m,n) = \sum_{k=1} \mathrm{EOF}_k(m)\mathrm{EC}_k(n) \quad (2)$$

Where $\mathrm{EOF}_k\ (m)$ is the kth spatial pattern defined for each of locations m and $\mathrm{EC}_k(n)$ is the expansion coefficient defined for each sampling time n for the kth spatial pattern.

Other names for EOFs are scores and loadings while amplitude functions or principal components are other names for ECs. These names reflect the fact that the EOF analysis is a particular case of the principal component analysis. Each of the EOFs explains some part of the variability in the anomaly values. Algorithms for EOFs computation generate EOFs in descending order by the value of explained variance. The proportion of the total variance in the data explained by the EOF_1 is the largest; EOF_2 explains less variance than EOF_1 but more than any of others, etc. In this work, the singular value decomposition was performed to obtain the EOFs and expansion coefficients (ECs) (Bjornsson and Venegas, 1997).

3.2.4 Data Analysis

All values of *E. coli* were $\log_{10}$-transformed prior to statistical analysis. In cases where *E. coli* was not detected in a water sample (i.e., a concentration of 0 CFU

100 mL^{-1}), the concentration was first assigned a value of one-half the detection limit (i.e. 0.5 CFU 100 mL^{-1}) and was then log-transformed.

The Kolmogorov-Smirnov test was used to compare distributions of *E. coli* concentrations between the pond interior and bank locations. The non-parametric Spearman correlation coefficient was calculated to determine similarities in the MRD or EOF values of locations between individual years as well as with the 3-year average. A value of 1 and -1 indicate extremes whereby there is a strong positive and negative monotonic relationship, respectively. PAST (Paleontological Statistics Software v3.2 (Hammer et al., 2001) was used to perform the abovementioned analyses. Site maps were created using QGIS v3.18. Figures were generated using Sigmaplot v13 (Systat Software, California). Sampling dates which were associated with rainfall events (6/22/2016 and 8/8/2016 at P2) were analyzed separately for all statistical testing.

EOFs analysis was performed in R using the sinkr package (Taylor, 2017; R Core Team. 2013). In this work, we used North's rule of thumb which is described in Overland and Preisendorfer (1982) and examines the uniqueness of singular values whereby overlapping error limits between neighboring singular values indicate a possible mixture of signals. Those that do not show overlap can be retained for interpretation. In this work, only EOF_1 in the 3-year datasets for both ponds was found to be statistically significant and therefore this pattern will primarily be discussed.

3.3 Results

3.3.1 Overview of the dataset

Daily average air temperature, daily precipitation, and average *E. coli* concentrations on each date are shown in Figure 2. For P1, there was a negative Spearman correlation between the days since previous rainfall and the *E. coli* concentration when data from all years was pooled (r_s = -0.544; p = 0.029). For P2, this relationship was also negative (r_s = -0.380; p = 0.146), indicating that *E. coli* concentrations in the surface water samples generally decreased with increasing separation from rainfall events.

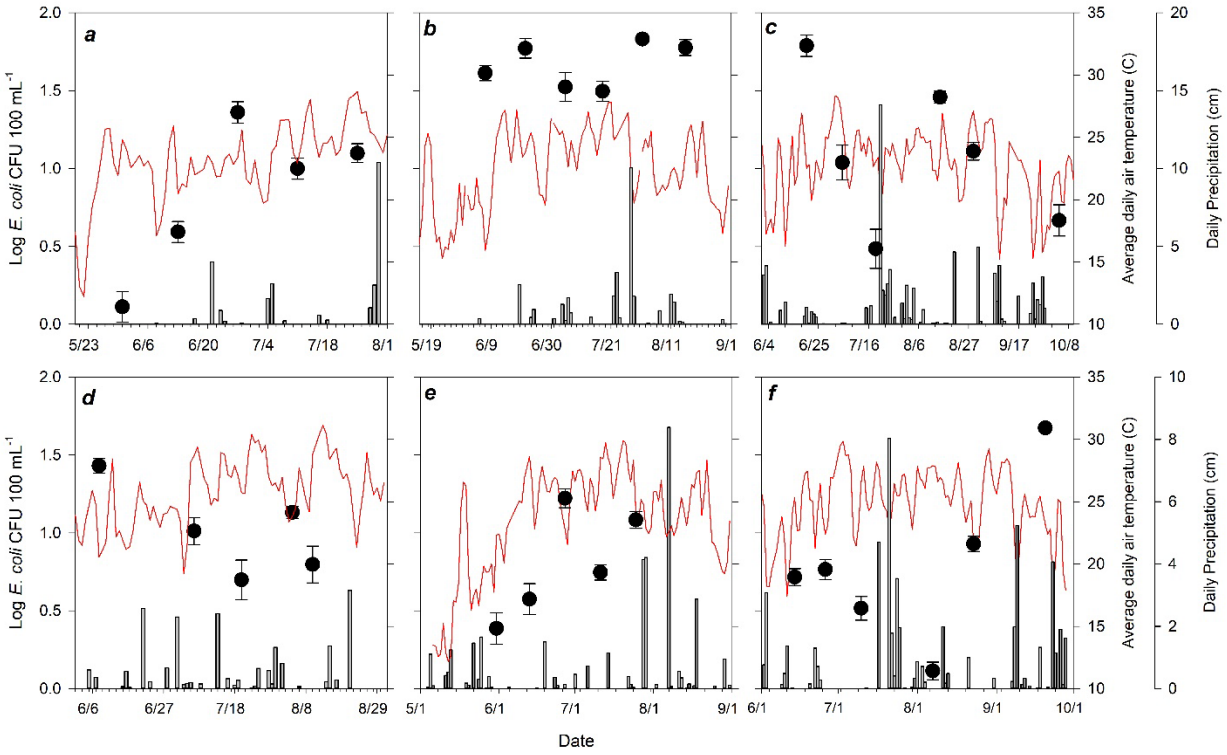


Figure 3.2. Daily precipitation (cm), average daily temperature (°C), and average across the pond *E. coli* value at P1 (a, b, c) and P2 (d, e, f) in 2016, 2017, and 2018, respectively. The mean *E. coli* value is displayed for specific sampling dates and is separated by the standard error of the mean by the ± symbol.

Average water temperature values in P1 showed considerable differences with a max and min value of 22.7 ºC and 29.6 °C, respectively, over the three-year observation period (Supplemental Tables S1 and S2, Appendix B). The average water temperature during all observation dates was 26.5 ± 1.9 °C (average ± S.E.). Annual averages of pH in P1 were very similar in 2016 and 2017 with values of 8.64 ± 0.33 and 8.81 ± 0.30, respectively, but in 2018 this value was noticeably lower at 7.64 ± 0.24. Average daily DO concentrations in P1 across years ranged from 7.89 to 13.1 mg L^{-1}, with a mean of 10.14 ± 1.35 mg L^{-1}. Average daily values of SPC ranged from 130.00 µS/cm to 175.82 µs/cm and had an inter-annual average of 147.88 ± 14.15 uS/cm. SPC showed small standard error across P1 on most sampling dates; however, on the 8/29/2018 the error was relatively large due to locations 20 and 21 have a combined average of 31.4 µS/cm whereas all other locations had an average of 142.5 µS/cm. CHL and PC levels in P1 were relatively low in both 2017 and 2018 with values generally < 5 RFU (100 RFU is the sensor maximum value) indicating minor populations of planktonic photosynthetic organisms in the water body. Average FDOM concentrations in P1 in the same years showed considerable variability across sampling dates which ranged from 2.11 ppb to 36.65 ppb in 2017 and from 1.92 ppb to 38.39 ppb in 2018.

Average daily water temperatures in P2 displayed a greater range than P1 with min and max values of 21.5 °C and 31.9 °C and an experimental average of 27.5 ±

2.6 °C. Annual average pH values were 7.6 ± 1.2, 8.5 ± 0.7, and 8.2 ± 0.8 in 2016, 2017, and 2018, respectively. D.O. values were typically higher in P2 than in P1 over the study period, and there were a greater proportion of sampling dates with DO levels > 10 mg L^{-1} in P2 than P1. The average DO over observation dates in P2 was 12.6 ± 3.7 mg L^{-1}, and the CV was 29.21 %. Daily average values of SPC ranged from 125 μS/cm to 179 μS/cm with a mean of 154 μS/cm, which were all very similar to the same metrics in P1 (129 μS/cm, 176 μS/cm, and 153 μS/cm, respectively). P2 contained levels of CHL and PC which were consistently several times higher than P1 and were similar across years. FDOM levels in P2 were also consistently higher and showed much less variability across sampling dates as compared to those measured in P1. Average annual FDOM concentrations in P2 were 27.22 ppb and 34.91 in 2017 and 2018, respectively.

3.3.2 Measurements taken on rainfall dates

On two dates at P2 within the 3-year observation period, samples were collected during or shortly after (< 8 hours) heavy rainfall events. These dates were 6/22/2016 and 8/8/2017 and interestingly, both dates had very similar *E. coli* concentrations with very low standard error at 3.07 ± 0.02 and 3.06 ± 0.17, respectively, indicating a high degree of mixing in the pond. Indeed, values of water quality parameters also demonstrated very low errors relative to the measured values (Tables S1.1 and S1.2, Appendix B). On the 6/22/2016 rainfall event, inflow water was sampled from the culvert at location 12 and was found to have an average *E. coli* concentration of 3.92 log CFU 100 mL^{-1} which was about 7 times higher than the

average value in the pond based on untransformed concentration data. DO and pH levels were lower than average on both rainfall dates while other parameters showed inconsistent trends. CHL and PC pigment levels were notably low on 8/8/2017 while average FDOM levels were the higher than any other date across the 2-years of observations that included this measurement.

3.3.3 Inter-annual variability of *E. coli* concentrations

In 2016, *E. coli* concentrations steadily increased throughout the sampling period from late May to late July in P1 (Figure 3). The 2017 sampling season showed consistently higher concentrations of *E. coli* than observed for P1 during the 2016 or 2018 sampling seasons. The 2018 sampling season in P1 showed the greatest variability in concentrations and did not show any consistent trend across the season. Mean logarithms of *E. coli* were 0.88 ± 0.51, 1.65 ± 0.31, and 1.09 ± 0.61 for 2016, 2017, and 2018, respectively (± separates mean and standard deviation). A Kruskal-Wallis ANOVA on ranks showed that *E. coli* concentrations, when pooled together by year, were significantly different ($P < 0.001$). Pairwise testing of any pair of years also showed significant differences ($p < 0.05$). The coefficient of variation for the log-transformed average *E. coli* concentrations at P1 was 59.1 %, 18.8 %, and 55.9% for 2016, 2017, and 2018, respectively.

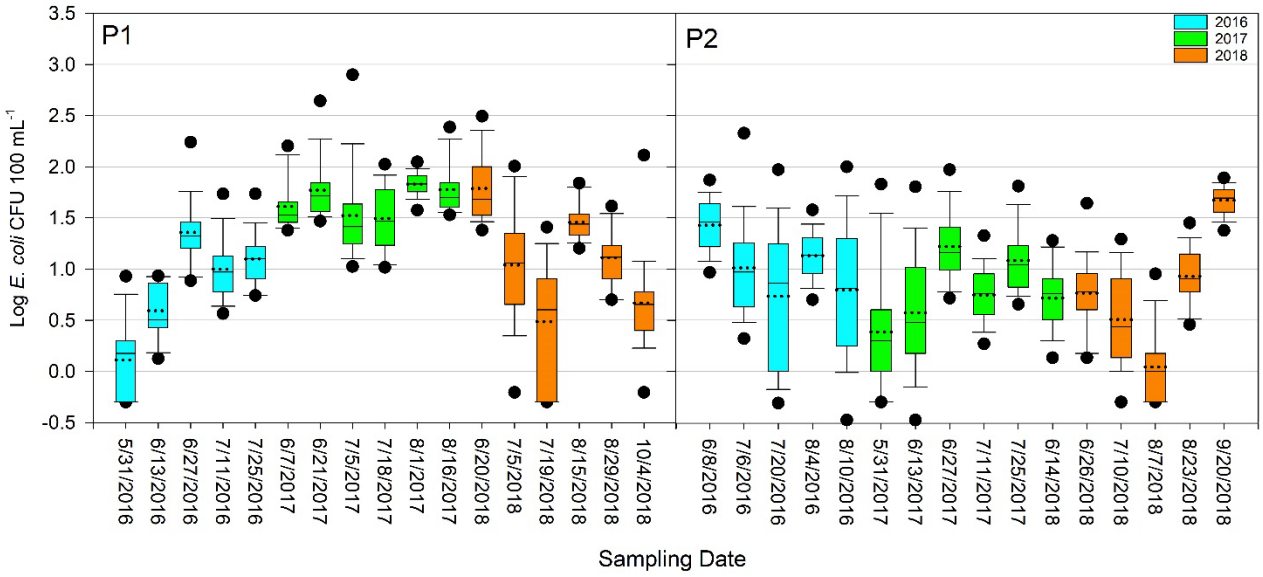


Figure 3.3. Box and whisker plots showing the spread of the observed data. Upper and lower portions of boxes show the 25th to 75th percentiles, respectively, (interquartile range) while whiskers show the 10th and 90th percentiles of the data, respectively. The solid and dotted lines represent the median and mean, respectively. Dots show the 5th and 95th percentiles.

Concentrations in P2 were similar to those values measured in P1 during the 2016 and 2018 years and displayed no consistent trend across years (Figure 3). In P2, the across-season averages of *E. coli* were 1.01 ± 0.57, 0.80 ± 0.53, and 0.77 ± 0.59 log CFU 100 mL^{-1} for 2016, 2017, and 2018, respectively. A Kruskal-Wallis ANOVA on ranks showed that *E. coli* concentrations significantly differed between years in P2 ($P = 0.001$). Mann-Whitney pairwise testing showed that the concentrations measured in 2016 significantly differed ($P < 0.05$) from 2017 and 2018, but concentrations in 2017 did not significantly differ from those observed in 2018 ($P = 0.533$). In general, P2 did not show a consistent within-season trend of increase or decrease in concentrations with the progression of the sampling season. For P2, the CVs of logarithms of *E. coli* were 56.4 %, 67.3 %, and 77.2 % for 2016,

2017, and 2018, respectively. The experiment-wide CVs for each pond when the data was pooled across years was 47.1 % and 67.8 % for P1 and P2, respectively.

3.3.4 Site-specific variability of *E. coli* concentrations

Concentrations of *E. coli* varied substantially across the two ponds during every year of the observations (Figure 4). In accordance with Fig. 2, Figure 3 shows that there were typically lower *E. coli* concentrations observed in P1 in 2016 and higher concentrations observed in 2017 with 2018 appearing somewhere intermediate. In general, the samples from the pond interiors contained lower concentrations than those collected near the banks. On average, the concentrations of *E. coli* measured in bank and interior samples was 1.38 ± 0.16 and 1.12 ± 0.09 log CFU 100 mL^{-1}, respectively. A Kolmogorov-Smirnov test for sameness of cumulative probability distributions showed that concentrations observed in the pond interior were significantly lower than those measured at the bank on multiple sampling dates, but the number of cases of such significance differed by year. In 2016, 3 of the 5 sampling dates (2^{nd} , 4^{th} , 5^{th} dates) were found to contain significantly higher ($P < 0.05$) *E. coli* concentration distributions in the bank samples than the pond interior. In 2017, this significant difference occurred for 4 of 6 dates (2^{nd} , 3^{rd} , 5^{th} , 6^{th}). Conversely, in 2018, only 1 of the 6 sampling dates contained significantly different distributions (3^{rd} date) while the other dates were found to contain more similar concentrations between the two location groups albeit always higher on the banks.

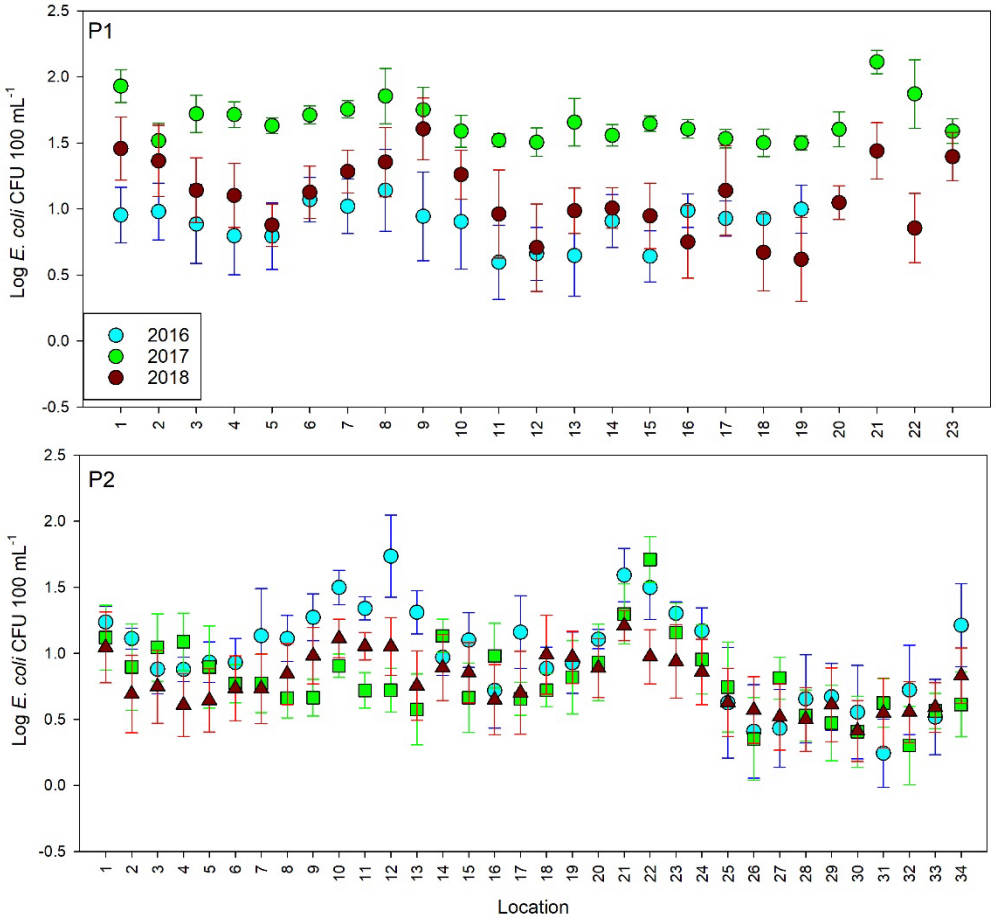


Figure 3.4. Observed concentrations of *E. coli* at each location by year during the 2016 to 2018 monitoring period. Error bars show the standard error of the mean.

In P2, the inter-annual differences in *E. coli* concentrations across the sampling locations were less pronounced than in P1. This level of difference meant that despite large differences in the spatial averages observed throughout the experiment, the trends of high and low concentrations by locations were preserved. Locations 8 through 13 show divergent concentrations in 2016 and 2017 whereas 2017 and 2018 were largely similar. The interior locations 25 through 33 demonstrated consistently lower concentrations than the bank samples collected across the rest of the pond. The 3-year average for the bank and interior locations was

0.97 ± 0.17 and 0.56 ± 0.11 log CFU 100 mL^{-1}, respectively. Interestingly, locations 25 and 34, which are adjacent on three sides to the bank showed concentrations more similar to bank samples than interior samples. The Kolmogorov-Smirnov test showed that in 2016 the 2nd , 3rd , and 5th sampling dates contained significantly different distributions of concentrations between the bank and interior. In 2017 only the 1st and 2nd sampling dates were found to have significantly different distributions ($P < 0.05$) and in 2018 all sampling dates, but the 3rd date ($p = 0.068$) were significantly different.

3.3.5 Temporal Stability Assessment of *E. coli* concentrations

The mean-relative-difference analysis pooled data from all 3 years for each pond and showed that some locations demonstrated temporal stability in *E. coli* concentrations with locations that were consistently lower, about, and above the spatial average on each sampling date (Figure 5). The consistently low and consistently high locations or groups of locations may be thought of as "cold" and "hot" spots or zones of *E. coli* concentrations in the pond, respectively. The degree of temporal stability of any location presented in Figure 4 is indicated largely by the size of the error bars present. Examination of the size of the error bars reveals locations that could potentially belong to another group of values (e.g. location 22 in P1 and location 12 in P2). However, the largest number of points is contained within the intermediate range and, thus, these locations can be informative for determining the relative concentrations across all other locations.

The MRD graphs showed similar dynamics by location between the years although there was usually some shuffling of locations within a given group (low,

medium, and high) (Supplemental Figure S1, Appendix C). For instance, in P1, the hot spots in 2017 consist of locations 3, 6, 9, 7, 8, 22, and 21 (presented in order by MRD value) and in 2018 this group consisted of locations 17, 9, and 21. Locations 9 and 21 did not change group, but location 17 went from being low to being high in 2017 and 2018, respectively. Another example may be seen in the ranked order of locations in P2 within the cold and hot spots in 2017 and 2018 (Figure S1.2, Appendix C).

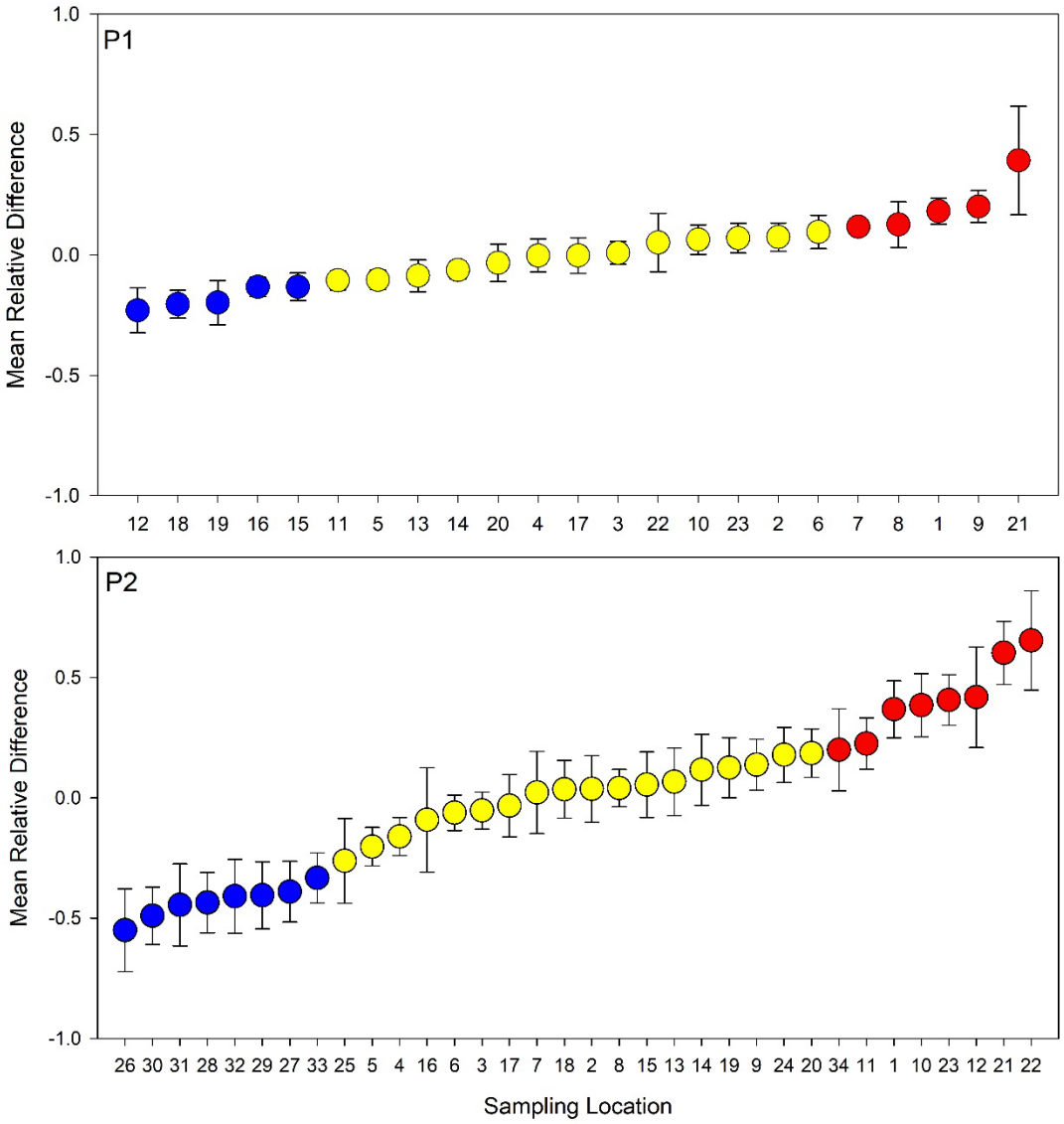


Figure 3.5. Mean relative difference (MRD) values for Pond P1 (P1) and Pond P2 (P2). Symbol colors were selected based on those that are below the 25th quartile (blue), within the 25th and 75th quartiles (yellow), and above the 75th quartile (red).

Overall, there was a moderate to high level of consistently in the ranked order of locations between years. In P1, the Spearman correlations between MRD values across years was $r_s = 0.524$ $p < 0.021$ for 2016 and 2017, $r_s = 0.842$ $p < 0.001$ for 2016 and 2018, and $r_s = 0.487$ $p = 0.018$ for 2017 and 2018. The correlation coefficients for each year's MRD and the 3-year average MRD was $r_s = 0.843$, 0.611, and 0.869 for 2016, 2017, and 2018 comparisons, respectively ($p < 0.001$ for each). P2 also showed a high degree of consistency in the ranks of locations from the MRD analysis. The Spearman correlation coefficients were 0.467 $p = 0.005$, $r_s = 0.789$ $p < 0.001$, and 0.452 $p = 0.007$ for comparisons between 2016 and 2017, 2016 and 2018, and 2017 and 2018, respectively. The strength of the correlations were generally strong between the 3-year average MRD values and each year: $r_s = 0.933$, $r_s = 0.645$, and $r_s = 0.896$ (all $p < 0.001$) for 2016, 2017, and 2018, respectively.

3.3.6 Correlation of MRDs of *E. coli* and Water Quality Variables

Average values of water quality variables are shown in Supplemental Tables S1 and S2 for P1 and P2, respectively (Appendix B). An MRD analysis was also performed on values of all measured variables and a correlation analysis was then performed to relate MRD values from *E. coli* concentrations with those values from the measured variables (Table 3.1). In P1 in 2016, the MRDs between *E. coli* and all measured water quality variables displayed significant negative relationships with similar strengths. In 2017, the direction of the correlations was similar during 2016

for that set of reported variables, however, the MRDs for pH and DO did not show significant relationships with *E. coli* MRDs. Correlations with C, SPC, turbidity, phycocyanin, and chlorophyll-a were found to be significant. FDOM showed a positive relationship with *E. coli* MRDs in both 2017 and 2018, but only in the latter year was the relationship significant. In 2018 in P1 pH, C, and DO MRDs all showed significant negative correlation with *E. coli* MRDs while SPC, NTU, PC, CHL all showed positive relationships. Correlations of MRDs from the pooled datasets among all the years show several significant relationships. Temperature, pH, and DO MRDs demonstrated significant negative relationships while the MRDs for NTU, PC, CHL, and FDOM showed moderate positive relationships with *E. coli* MRD values. SPC demonstrated a weak negative coefficient value when all years were pooled.

Table 3.1. Spearman correlation coefficients (r_s) between *E. coli* MRDs and MRDs of the listed water quality variables.

	P1				P2			
	2016	2017	2018	All Years	2016	2017	2018	All Years
pH	**-0.681**	-0.136	**-0.489**	**-0.485**	**-0.426**	**-0.642**	**-0.490**	**-0.394**
C	**-0.646**	**-0.541**	**-0.609**	**-0.635**	**-0.419**	-0.311	**-0.351**	-0.303
DO	**-0.509**	-0.051	**-0.810**	**-0.437**	-0.312	**-0.523**	**-0.556**	**-0.450**
SPC	**-0.617**	**-0.434**	0.042	-0.299	-0.098	0.015	-0.044	-0.288
NTU	N.M.	**0.530**	0.305	0.359	N.M.	0.263	**0.436**	**0.607**
PC	N.M.	**0.602**	**0.52**	**0.591**	N.M.	0.311	**0.417**	**0.582**
CHL	N.M.	**0.508**	**0.452**	**0.602**	N.M.	**0.439**	0.176	**0.524**
FDOM	N.M.	0.242	**0.673**	**0.669**	N.M.	0.149	0.209	0.206

Values in bold indicate statistically significant ($P < 0.05$) correlations. Parameters are abbreviated as C (temperature), DO (dissolved oxygen), SPC (specific conductivity), NTU (turbidity), PC (phycocyanin), CHL (chlorophyll-a), and FDOM (fluorescent dissolved organic matter). N.M. stands for not measured.

Like P1 in P2 during 2016, all correlations between *E. coli* MRDs and those values for water quality variables were negatively correlated, however, only the comparisons with pH and C were significant. In 2017, *E. coli* MRDs were negatively correlated with those values for pH, DO, and C with the former two showing significant relationships. SPC, NTU, PC, CHL, and FDOM all showed positive correlation with *E. coli* MRDs, but only the CHL comparison was significant. In P2 during 2018, pH, C, and DO showed low to moderate negative significant relationships with *E. coli* MRDs while SPC showed a very weak negative correlation. NTU, PC, CHL, and FDOM MRDs were all positively correlated with *E. coli* MRDs with the former two being significantly related. When all years of data for P2 were pooled, pH, C, DO, and SPC showed low to moderate negative relationships with *E. coli* MRDs while NTU, PC, and CHL showed moderate and significant positive relationships. MRD values of FDOM showed a weak positive relationship with *E. coli* MRDs which was not significant. The MRD correlations between water quality variables and *E. coli* showed the same directional relationship between ponds albeit to different strengths and significances.

3.3.7 Empirical Orthogonal Function Analysis

The empirical orthogonal function analysis was performed on logarithms of *E. coli* counts per 100 mL^{-1} measured at each pond, date, and location. The total explained variance in the anomalies differed for each pond and by each of the observation years (Table 3.2). EOF_1 for P1 explained 34%, 48%, and 54% percent of the total variance within the datasets for 2016, 2017, and 2018, respectively. The second and third EOFs in P1 also tended to explain a non-trivial fraction of the total

variance. In 2016, the sum of EOF_2 and EOF_3 explained a larger amount of variation than EOF_1 (34.17%). In 2017 and 2018 the variance explained by EOF_1 and the sum of EOF_2 and EOF_3 were similar at approximately 50% and 40%, respectively. Altogether, the sum of the first three eigenvalues s at P1 explained a similar fraction of the total variance at 87, 94, and 90 % for each respective year. Pooling data from all years produced an explained variance of 37% for the first P1 EOF. The first EOF in the 3-year P1 dataset was the only one to show a significant pattern whereas EOFs from individual years were not found to be significant.

Table 3.2. Percentage of total variance explained by the Empirical Orthogonal Functions Analysis.

	Pond 1				Pond 2			
EOF	2016	2017	2018	All Years	2016	2017	2018	All Years
1	34.17	48.04	53.75	**37.17**	60.23	50.81	44.65	**40.53**
2	29.59	26.24	22.50	17.55	19.70	25.52	21.10	13.74
3	23.22	19.76	13.73	10.96	12.98	13.25	14.58	8.90
4	8.03	4.21	5.14	7.23	5.34	7.44	9.93	8.58
5	4.99	1.75	3.88	6.17	1.75	2.98	7.13	6.11
6	-	-	1.00	5.92	-	-	2.62	5.01

Bold values indicate statistically significant percentage of the total variance explained.

Total variance explained by the first EOF at P2 was generally higher than P1 at 60, 51, and 45 % for 2016, 2017, and 2018, respectively. In 2016 the EOF_2 explained much less variation than EOF_1, and the sums of the trailing EOF_s were also low compared to EOF_1, indicating a robust spatial pattern identified by the first EOF. In 2017 and 2018, variances explained by the first and second EOFs were similar. Summing the first three EOFs for P2 also produced high total explained variance at

93, 90, and 80 for three observation years, respectively. EOF_1 for the pooled dataset for P2 explained 41% of the explained variance. Like P1, the EOF_1 from the 3-year dataset was the only pattern to demonstrate statistical significance.

EOF_1 values from P1 between pairs of years showed weak positive correlations for comparisons between 2016 and 2017 ($r_s = 0.263$; $p = 0.276$) and 2017 and 2018 ($r_s = 0.332$; $p = 0.121$). However, the comparison between 2016 and 2018 showed a stronger and significant relationship between EOF_1 values ($r_s = 0.533$; $p = 0.018$). For P2, EOF_1 values between locations and across years demonstrated significant positive relationships with correlation coefficients of 0.429, 0.804, and 0.489 for the same comparisons as the above annual pairs ($p = 0.011$, $p < 0.001$, and $p = 0.003$, respectively).

Figure 6 shows the spatial distribution of EOF_1 values from the 3-year datasets of *E. coli* concentrations from the two ponds. For both ponds, lower EOF_1 values were observed in the pond interiors. The Kolmogorov-Smirnov test demonstrated that there was a significant difference in the distribution of EOF_1 scores between the interior and banks samples in both P1 ($p = 0.007$) and P2 ($p < 0.001$).

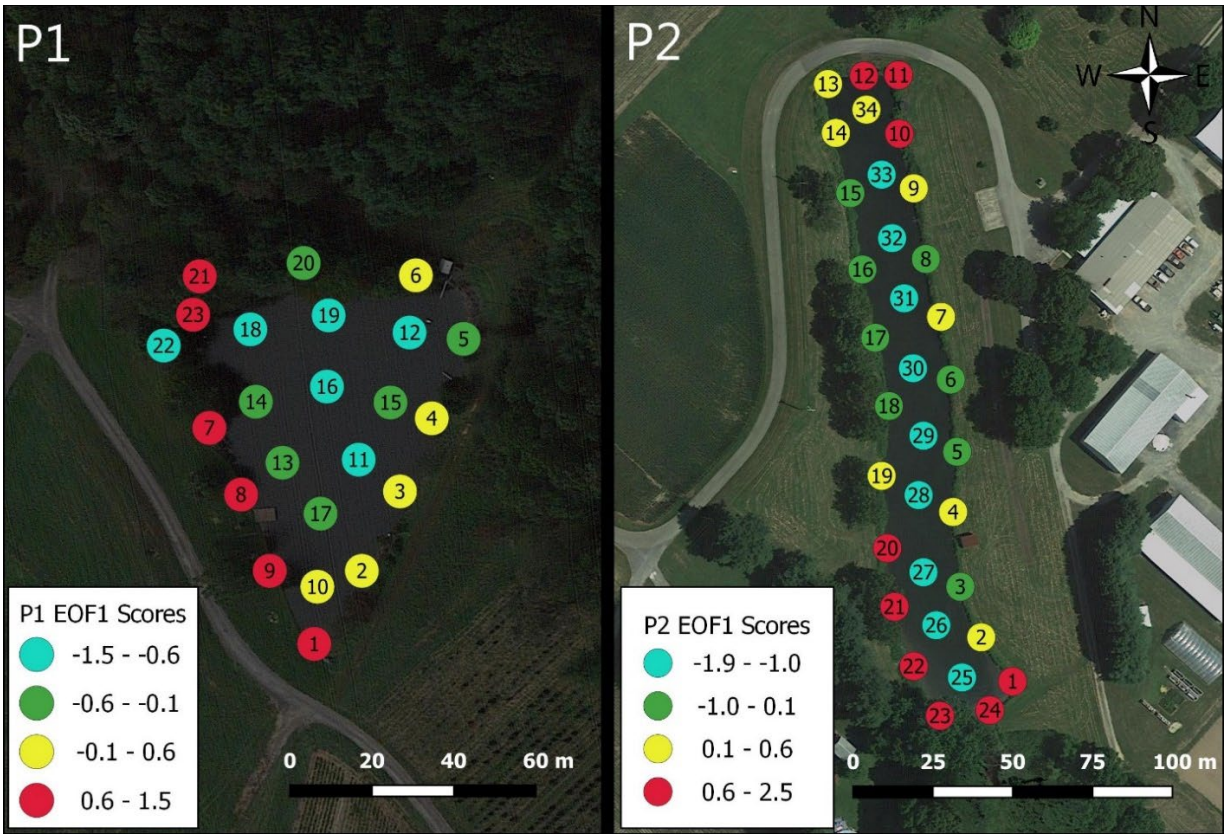


Figure 3.6. Spatial distribution of EOF_1 values of *E. coli* across the two observation ponds between 2016 and 2018.

Several *E. coli* hot spots and cold spots were identified by EOF_1 in both ponds. These designations were determined by sampling locations with the highest and lowest EOF_1 values, respectively, relative to the other locations. For P1, this included the bank adjacent to the road up to the Western corner of the pond and the pond apex. Locations closest to the ephemeral inflow were also identified as hotspots; however, spot 22 stands as an outlier and an identified "cold spot". The EOF analysis also revealed hot and cold spots in P2. Hot spots were in the Northern and Southern ends of the pond adjacent to the ephemeral inflow and the outflow, respectively. Sampling locations along the banks were primarily comprised of EOF_1 scores in between extremes, while the interior contained mostly cold spots.

A comparison between EOF_1 score distributions among years can be seen in Figure 7. In most cases, the sign and amplitude of the EOF_1 scores are in agreement among all three years of observations. However, there are some notable exceptions. For instance, in P1, bank locations 6 and 22 have large EOF_1 values for 2016 which are not present in any other year, indicating a strong deviation from the spatial average was observed at this location during this year. There are notable differences in P2, particularly in the bank sampling locations from 2 to 9 and 16 through 19. Locations 1, 10 through 12, 20 through 24, and 25 through 33 all have similar deviations between years.

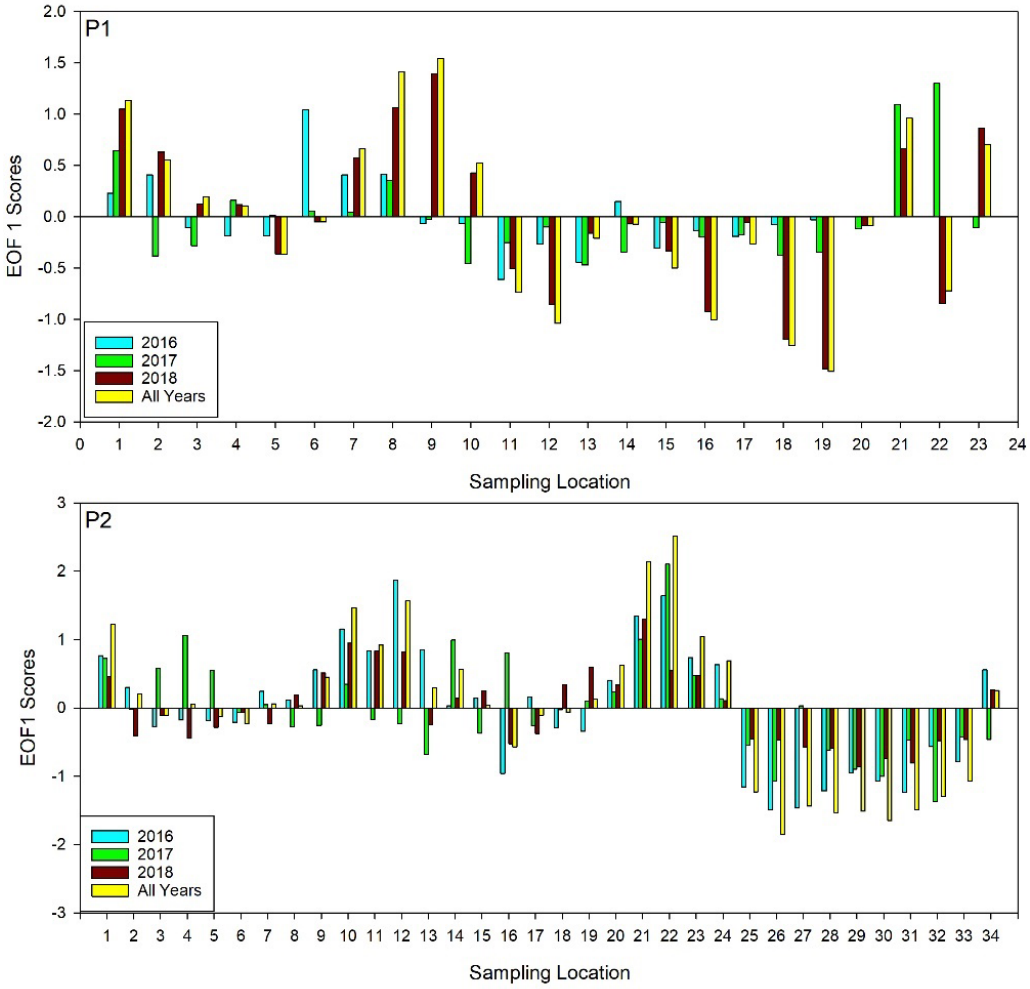


Figure 3.7. EOF_1 values for specific locations by year alongwith the 3-year average.

Graphs of the expansion coefficient 1 (EC_1) associated with the predominant empirical orthogonal function (EOF_1) are shown in Figure 8. EC_1 shows the weight with which the EOF_1 contributed to the anomaly values on each sampling date (Eq. (2). The EC_1 values were almost entirely positive for both ponds across all dates. The one exception is the 6/8/2016 sampling date at P2, where the EC_1 is slightly negative.

Values of EC_1 associated with specific sampling dates were correlated with the observed daily variances of the *E. coli* dataset. P1 showed a positive, yet non significant, relationship ($r_s = 0.380$ $p = 0.132$) and P2 displayed a strong positive and significant relationship with the daily variances observed in the dataset ($r_s = 0.932$, $p < 0.001$). The EOF_1 and EC_1 was able to effectively capture both the spatial and temporal variability of the datasets.

A correlation analysis was performed between the EOF_1 scores and the original datasets on *E. coli* concentrations by date and location to assess on which dates the predominant spatial pattern in EOF_1 was present and to what degree. Results of the correlations are shown below each graph of EC_1 values in Figure 6. On a small number of sampling days, the dominant spatial pattern was not strongly observed, and the resulting correlation coefficient was not significant ($p > 0.05$). For P1 these dates were 5/31/2016, 7/5/2017, 7/18/2017, and 10/4/2018. In P2, the only the 6/8/2016 date that did not show a significant correlation between *E. coli* concentrations and the EOF_1

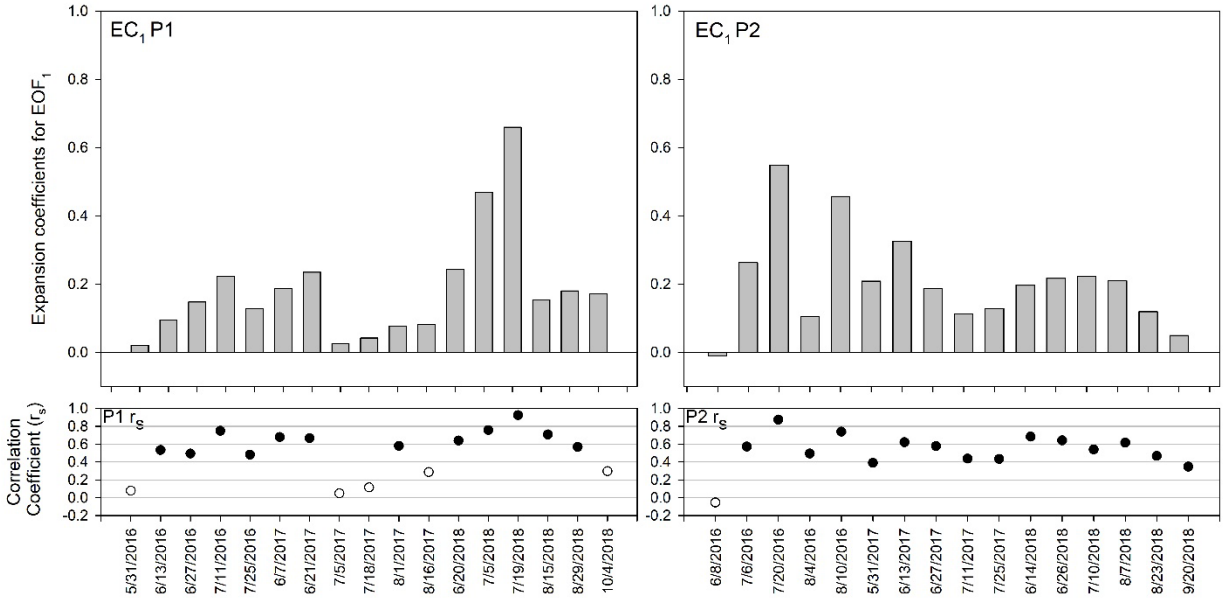


Figure 3.8. Values of expansion coefficients (E.C.s) of the leading empirical orthogonal function pattern (EOF_1) (top panel). Values of Spearman coefficients for comparisons of EOF_1 and individual sampling dates (bottom panel).

3.3.8 Correlation of *E. coli* and Water Quality Variable EOF_1 values

EOFs were also computed for select water quality parameters where paired measurements were consistently taken alongside *E. coli* measurements between 2016 and 2018. These included temperature (°C), specific conductivity (SPC), pH, and dissolved oxygen (DO). NTU EOFs were not computed due to large gaps in the 2016 dataset. In P1, the Spearman correlation coefficient (r_s) between EOF_1 for *E. coli* and temperature demonstrated a significant negative relationship (Table 3.3). Correlations between EOF_1 for *E. coli* and both pH and DO in P1 demonstrated significant negative relationships. EOF_1 patterns between *E. coli* concentrations and specific conductance showed no correlation. In P1, DO and pH were positively correlated as were DO and temperature. Specific conductivity scores were negatively correlated with pH, DO, and temperature.

Table 3.3. Spearman correlation coefficients (r_s) for *E. coli* and water quality variables. Correlations for P1 and P2 are shown below and above the diagonal line, respectively.

	EC	DO	pH	C	SPC
EC		**-0.403**	**-0.460**	-0.321	-0.059
DO	**-0.491**		**0.955**	**0.838**	**-0.527**
pH	**-0.591**	**0.811**		**0.776**	**-0.509**
C	**-0.707**	**0.748**	**0.797**		**-0.626**
SPC	-0.001	0.166	0.087	0.091	

Values marked in **bold** indicate statistical significance ($p < 0.05$)

In P2, spatial patterns from EOF_1 of *E. coli* and temperature demonstrated negative relationship. Patterns between *E. coli* and both pH and DO showed significant negative correlation. Like in P1, EC and SPC EOF_1 values were not correlated. DO and pH EOF_1 values were strongly correlated. Values of EOF_1 for temperature were also strongly correlated with both DO and pH EOF_1 values in P2.

EOF_1 values for water quality parameters across both ponds are shown in Supplemental Figures S3 – S7 (Appendix C). Examination of Figure 2 and Supplemental Figures S3 – S7 shows that there were similar trends of high and low EOF_1 values between bank and interior locations for *E. coli* and those of water quality parameters. For example, the EOF_1 values for *E. coli* and temperature show opposite symbol colors in the banks versus interior which agrees with the reported negative correlation (Figure 2 and Figure S4).

3.3.9 Relationship between MRD and EOF values for *E. coli* and water quality variables

Since both the MRD and EOF analyses rank measurement locations by their typical deviations from the average across the pond, it was of interest to assess the similarity of these rankings. To evaluate the agreement between the 3-year MRD and EOF_1 values for each location, linear regression and Spearman correlation analyses were performed. Figures 9 and 10 show the linear regression results between MRD and EOF values for Ponds 1 and 2, respectively. Values for these analyses for P1 showed a strong significant and positive relationships ($R^2 = 0.735$; $p < 0.001$). A stronger monotonic relationship on the agreement of the rankings of locations between the two methods was observed ($r_s = 0.919$; $p < 0.001$). In Figure 9 it can also be seen that both analyses captured earlier observations of interior locations containing smaller MRD and EOF values. Most locations were clustered close to the 1:1 regression line with only a few exceptions. To the right of the line, the greatest differences were between locations 21 and 22 with residuals of -0.955 and -0.928, respectively, which were about twice as large as the next location which was 19 with a residual of -0.526. The greatest error on the left side of the line was for locations 8 and 9 with residuals of 0.583 and 0.819, respectively.

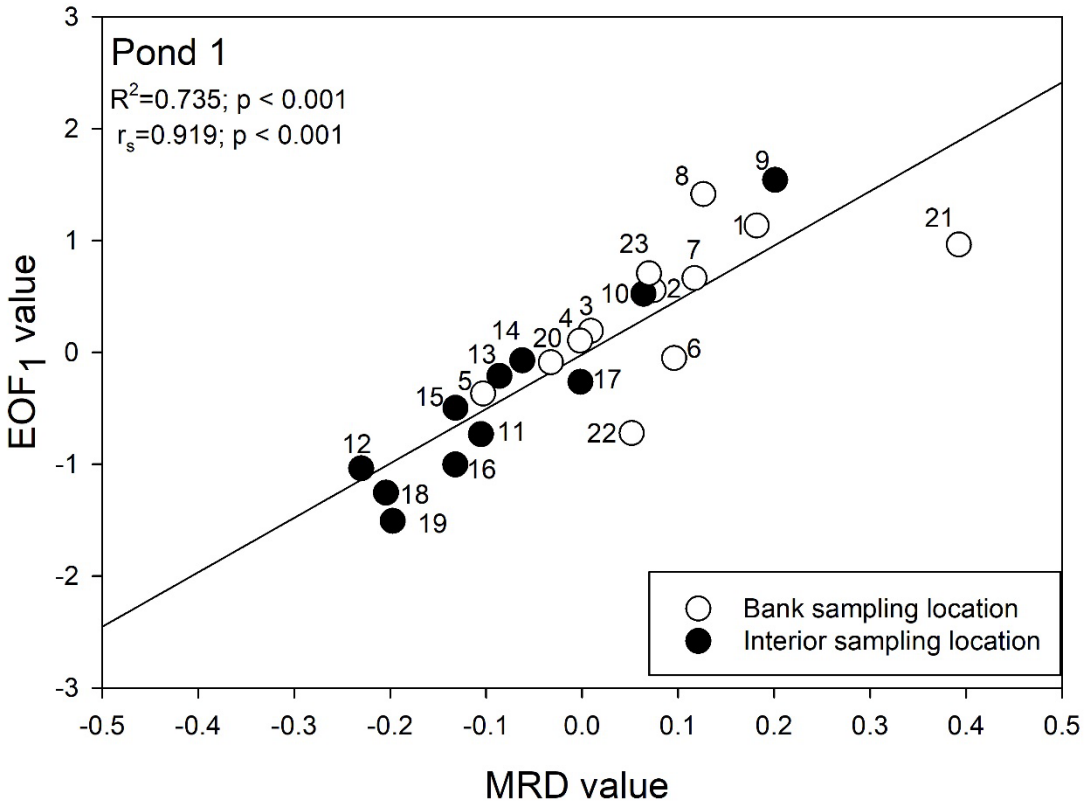


Figure 3.9. Linear regression between sampling site values calculated from the MRD and EOF analyses for Pond 1.

For P2, the agreement between methods as determined by the linear regression analysis was much stronger than P1 ($R^2 = 0.961$; $p < 0.001$). The monotonic Spearman correlation showed a slightly stronger relationship ($r_s = 0.973$; $p < 0.001$). Like P1, the linear regression in Figure 10 shows the clear groupings of interior and bank sampling locations whereby interior locations contain lower values from both analyses with location 34 being the only exception. Residuals from the regression were lower than in P1 with the greatest positive values being 0.600 (location 4) and 0.557 (location 5) and the lowest values being -0.435 (location 34) and -0351 (location 23).

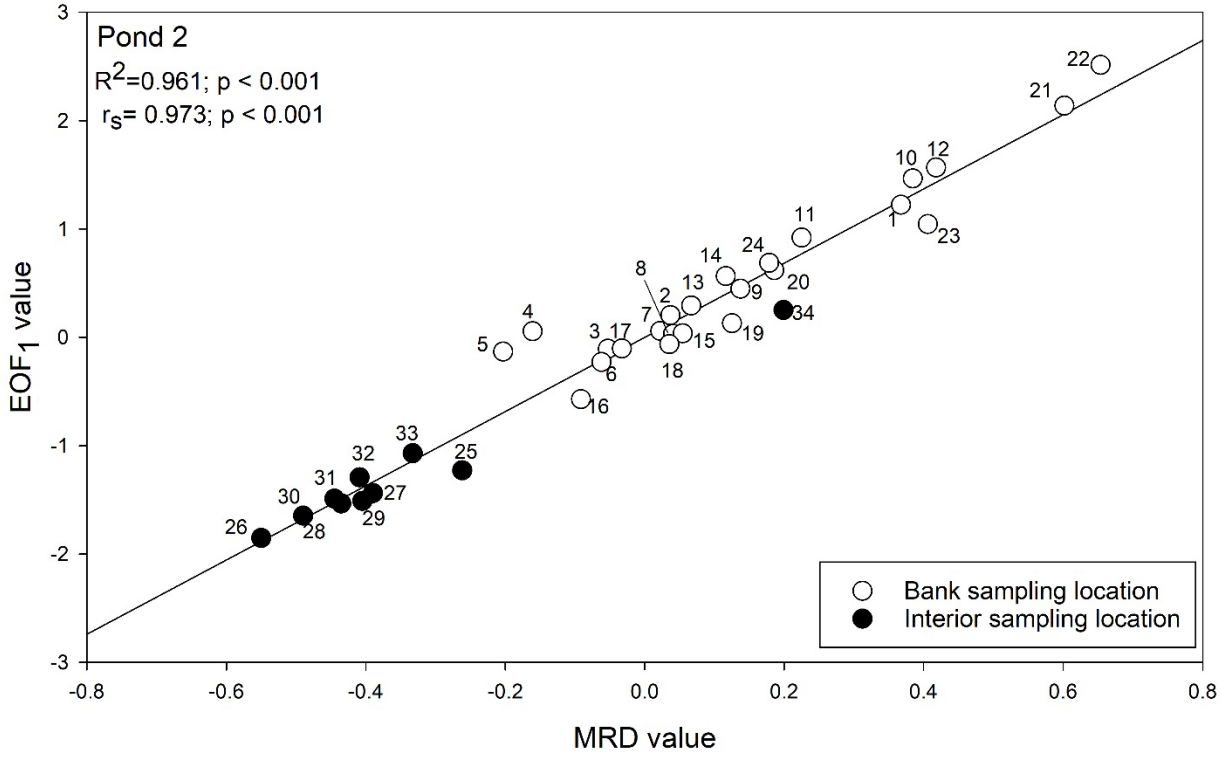


Figure 3.10. Linear regression between sampling site values calculated from the MRD and EOF analyses for Pond 2.

Examination of Tables 3.2 and 3.3 show that there is a perfect agreement between the sign of the relationships between EOF or MRD values for *E. coli* concentrations and those calculated for water quality variables. Interestingly, in most cases the strength of the relationship is similar as well. For example, the Spearman correlation value of -0.403 and -0.491 were recorded between the 3-year EOFs of *E. coli* and dissolved oxygen and for the MRD analysis in P1 these values were -0.437 and -0.450, for P1 and P2, respectively, with all cases showing statistical significance. Likewise, correlation coefficients for pH were -0.460 and -591 for P1 and P2 EOFs, respectively, while these values for MRD were -0.485 and -0.394 with all showing significance. SPC demonstrated consistent negative correlations in both analyses

however the MRD analysis showed a stronger relationship. Some differences between analyses can be seen for temperature which showed significant negative association with *E. coli* for P1 EOF and P2 MRD but non-significant association with *E. coli* in P1 MRD and P2 EOF although the sign of the relationship was maintained across analyses.

3.4. Discussion

The two ponds in the present study demonstrated satisfactory microbial water quality over the 3-year period. On average, no sampling dates exceeded the water quality criteria limit which is a geometric mean value of 126 CFU 100 mL^{-1} (log 2.1 CFU 100 mL^{-1}) as established by the US Food and Drug Administration (Tables S1 and S2, Appendix B) (US FDA, 2019). The STV values also were below the threshold value of 410 CFU 100 mL^{-1} (data not shown).

The concentration ranges of *E. coli* in the two studied ponds differed by year (Fig. 3). The intra-annual variability of indicator organisms measured in the same season during consecutive years is not commonly reported, but the available literature has shown that concentrations may vary considerably. For example, Topalcengiz et al. (2017) measured *E. coli* concentrations for 3 years at six different agricultural ponds in Central Florida between 2012 and 2014. When analyzing their summer data (May and June) for 2013 and 2014, it was seen that in 5 of 6 ponds the concentrations in 2014 ranged from 3 to 13 times greater than those observed in 2013 and in 1 pond

the concentration average was 118 times greater in 2014. Similarly, Jokinen et al. (2019) documented significant inter-annual differences in *E. coli* and *campylobacter* in an irrigation reservoir in Alberta, Canada. In year 1 of their study, microbial water quality exceedances were attributed to runoff from surrounding agricultural lands which occurred earlier in the season, whereas, in year 2 most of the exceedances occurred later in the season when there was little precipitation and many instances of exceedance were attributed to wild birds and cattle. Conversely, Durham et al. (2016) reported similarity in concentrations of *E. coli* in the summer months for 6 lakes in Lubbock, Texas between 2011 and 2013. The largest differences in concentrations between summer seasons across their study were attributed to heavy precipitation events that are known to efficiently transport fecal material and organisms to surface waters with runoff as well as to disturb bottom sediments which seem to be a large reservoir for *E. coli* in the environment (Jenkins et al., 2015).

The present study was conducted with emphasis on avoiding measurements taken within close proximity to rainfall events to ascertain the expected concentrations during baseflow and when farmers are most likely to irrigate. For this reason, only two sampling events were performed at P2 shortly after heavy rainfall. During these sampling dates the concentrations of *E. coli* in the pond were around 1000 CFU 100 mL^{-1} on both dates and the inflow water had concentrations roughly ten times greater. Indeed, the USEPA has listed rainfall events as the greatest cause for temporal differences in indicator densities in surface waters (USEPA, 2010). Jenkins et al. (2015) reported a two orders of magnitude increase in fecal indicator bacteria for three ponds during stormflow events observed in their study. The authors

attributed the elevated levels to runoff and sediment disturbances. Harris et al. (2018) measured the levels of *E. coli* and *Salmonella* before and after rainfall events in two irrigation ponds in GA and reported significantly higher concentrations of both organisms after precipitation events which agrees with the results of several other studies that involved measurements conducted during precipitation events (Steele et al., 2005; Chen and Chang, 2014; Kleinheinz et al., 2010). Farmers would likely need not irrigate crops for some time after heavy rainfall events, but more data is needed to determine the optimal waiting period before resuming irrigation and what factors may affect the wait time. One example comes from Jenkins et al. (2015) who reported that it took about 5 days for levels of *E. coli* and enterococci to return to pre-storm levels. Interestingly, the levels of Salmonella and *E. coli* O157 remained significantly higher 5 days after the event in their work.

E. coli concentrations in both ponds showed a substantial spatial variability in measurements taken across the ponds (Fig. 4). It was observed that on numerous dates the concentrations of *E. coli* were significantly higher in bank samples than in interior samples. Possible explanations for this observation may include decreased shading of interior locations relative to bank locations as solar insolation has been shown to increase indicator organism inactivation rates in pond water (Davies-Colley et al., 2000). Pond interior locations also had higher pH and DO levels than the banks (data not shown). Davies-Colley et al. (1999) reported that when pH values exceeded 8.5 the inactivation of *E. coli* in pond water increased, but the inactivation may also be partially related to elevated DO levels that are linked with photo-oxidative damage to fecal-microorganisms in water. Bank samples are typically located along shallower

parts of ponds. The wind-driven movement of water against bank soils and sediments may cause the release of microorganisms and the settling process would be less effective near the banks as compared to the interior due to wave action. Finally, bank locations may be expected to have higher concentrations of *E. coli* because this interface is where runoff enters which may be rich in fecal bacteria and/or nutrients dependent on the conditions and surrounding land. It should be noted that the hot spots identified in P1 are at locations 9 which coincides with small beach where recreational swimmers and waterfowl sometimes enter the pond and location 21 where an ephemeral creek is located. In P2, location 12 is considered a hot spot as the location is adjacent to the inlet. Locations 21 and 22 may also be consider hot spots since both locations are adjacent to the residential property containing domesticated animals and a septic system.

Correlation coefficients were calculated between MRD and EOF values for *E. coli* and those of water quality variables (Tables 3.2 and 3.3). In performing this analysis, we compared spatiotemporal patterns of variables rather than simply correlating the raw values which is the typical procedure. There was a considerable amount of consistency in the interrelationships among MRD and EOF values of *E. coli* and the values of the water quality variables. Correlation coefficients between pH and *E. coli* demonstrated a consistently negative relationship. A similar finding was reported by Draper et al. (2016) who observed significant, albeit weak, relationships between pH and aerobic plate counts, total coliforms, fecal coliforms, and enterococci in irrigation waters on 39 farms in Pennsylvania over a 2-year period. Topalcengiz et al. (2017) also reported significant negative relationships between pH and *E. coli*,

enterococci, and total coliforms in 6 ponds in Central Florida in a 3-year study. As previously noted, elevated pH levels (> 8.5) have been linked with increased inactivation of indicator organisms in pond water (Davies-Colley et al. 1999). DO MRDs also showed consistent negative relationships with *E. coli* MRDs. The increase of pH and DO oftentimes occur simultaneously due to photosynthetic activity by planktonic organisms in water as dissolved CO_2 is converted into organic components that may serve as nutrients, and release DO as a by-product. This process also spurs the dissociation of HCO_3^- which further increases the pH (Zang et al., 2011). The positive relationship between *E. coli* MRDs and those of turbidity and dissolved organic matter may be related to the shielding effects that suspended materials provide for fecal microorganisms in water (Whitman et al., 2003). Similarly, the positive relationships with both PC and CHL may also be linked with decreased light penetration as well as the potential for cyanobacteria and algae to release dissolved organic carbon and other nutrients back into the water column which may benefit *E. coli* survival (Davis et al., 2005).

The MRD analysis was used to identify sampling locations which were consistently lower, about, or higher in *E. coli* concentrations than the average across the pond. Differences in hot and cold spots could then be related to activities on the shore or physiochemical properties of the water and the effects on fecal microorganisms in terms of changing the habitat conditions. Additional usefulness in computing MRD values for sampling locations is introduced by gaining the ability to use locations with low error, or high temporal stability, to estimate the concentrations across the rest of the pond. For example, if one were to develop an MRD distribution

of *E. coli* in a pond, a location characterized by the consistent display of small errors might be used to help predict the concentrations in different locations based on the pre-established relationships between the locations in question. Similarly, the locations between the 25th and 75th percentiles of the MRD curve (indicated in yellow in Fig. 4) might be useful to help approximate the average concentrations across the rest of the pond and would reduce the need for intensive sampling regimes. A similar type of procedure was performed by Pedrera-Parrilla et al. (2017) who were able to calculate an MRD curve for soil water contents and electrical conductivity from multiple sampling locations across a field. The authors defined a representative location from a MRD distribution as one with low error which could be used as a constant bias to acquire the average soil water content and electrical conductivity over the observation area. One may expect a much weaker temporal stability in aquatic systems than in soil systems due to mixing and movement of water. However, the results of the current study indicate that even with the occurrence of mixing, the temporal stability of *E. coli* concentrations were still observed. The uncertainty of the MRD curves was greater for P2 than P1 which may reflect differences in the relative degree of mixing, the fecal microbe inputs, and microbial survival conditions between the two ponds.

The EOF analysis informed not only about the temporally stable anomalies in specific locations but also about contributions of individual locations to the overall variability. The observed correlations between expansion coefficients and variance of logarithm concentrations across the pond on the observation dates illustrate the role of the first EOF as the most informative pattern. If one truncates (Eq. 2) and leaves only

the first term of the sum, then for each observation time, the variance of the anomalies will be equal to the variance of the EOF_1 multiplied by the squared EC_1 for this observation time. Because the first EOF does not explain the total variance, the above proportionality can only be observed as approximate estimation.

In general, the EOFs explained more variance over one year than over the combined 3-year period (Table 3.3). However, the emergence of interpretable signals (patterns) from the noise according to North's test was not noted for any EOF found in a 1-year observation period. Three years of data were necessary to delineate the significant spatial patterns in the data. It remains to be seen whether the sheer amount of sampling dates or the temporal extent of observations control the relationship between EOF and noise in this case. Additionally, the authors note that the first several EOFs in various years explain a non-trivial fraction of the total variance explained however they were not found to be significant. Researchers may elect to choose different significance tests or analyze non-significant EOFs and ECs which may provide additional insight into their data.

The two spatiotemporal analyses utilized in this study provided results which were largely in agreement which held true for the concentrations of *E. coli* as well as water quality parameters. The MRD analysis uses relative anomalies, whereas the EOF analysis works with absolute anomalies. This presents some advantages to the MRD technique when there are large temporal variations of the studied variable. On the other hand, the MRD analysis generates a single pattern. The EOF analysis reveals several different spatial (EOFs) and temporal (ECs) patterns that may also represent large fractions of the variability and can be attributed to separate

mechanisms affecting the observed spatiotemporal variability. Additionally, there are several methods to determine the significance of EOFs (Babamoradi et al., 2013; Bjornsson and Venegas, 1997; Hammer et al., 2001), but no such techniques are known for the MRD analysis. Overall, each of the spatial pattern recognition methods has useful features and the methods can be complementary (Vereecken et al., 2016) which was evident in this work.

It must be acknowledged that the present study only studied two irrigation ponds in the Mid-Atlantic region over the course of 3-years. Indeed, many thousands of ponds exist in this area with a smaller number actively used for crop irrigation. The objective of this work was not to document *E. coli* dynamics in a large number of ponds, rather it was to test the hypothesis that there may exist spatial- and temporally stable hot- and cold-spots of *E. coli* concentrations which has implications for water quality sampling and the design of future monitoring programs. Pachepsky et al., (2018) showed that temporally stable patterns of *E. coli* can be detected within a single year of observations. This work aimed to assess the inter-annual variability of those patterns and answer the question of whether they are stable outside of the span of a single year. The findings presented show that within these two ponds, the temporal stability of *E. coli* concentrations can indeed be persistent. Future work should aim to determine the applicability of the MRD and EOF methods on additional study sites within and outside the region examined in this work.

The existence of stable spatial patterns of *E. coli* concentrations has consequences for both for microbial quality monitoring and management of irrigation ponds. We note that the discovery of stable spatial patterns in this work required high

spatial density of sampling that has not been implemented in previously published research. The spatial detail of the monitoring should be increased for both detecting and accounting for temporally stable patterns. Results of this work show that sensing water quality variables can be instrumental for establishing *E. coli* MRD or EOF patterns because of the relationships observed with water quality parameters and *E. coli.* Additionally, the use of drone-based imagery may be an appropriate source of data on spatial variation of major microbial habitat-affecting parameters such as chlorophyll-a, suspended solids, and dissolved organic matter concentrations (Pyo et al., 2017; Kutser et al., 2020). Future work should evaluate additional methods to increase the detection and knowledge of spatial patterns in microbial water quality studies such as geographically weighted regression and cluster analysis which were both recently shown to provide promising results within this field (Wang et al., 2021; Giao et al., 2021).

3.5 Conclusions

Large inter-annual spatial and temporal variability of *E. coli* concentrations was documented in two irrigation ponds from biweekly grid sampling for three years. Bank and interior sampling locations typically contained significantly different concentrations. In each of the ponds, the MRD analysis identified groups of sampling locations that had *E. coli* concentrations consistently below or consistently above the average concentrations across each pond. The EOF analysis provided similar results as the MRD analysis in terms of the ranking of locations based on the observed concentrations in space and time which was confirmed by linear regression and correlation analysis. Such stable spatial patterns were related to differences in habitat

conditions for *E. coli* across the ponds and pathways in which the microorganism may enter. For each pond, the MRDs demonstrated significant positive relationships across years which exemplified the multi-year persistence of the observed patterns. The MRD patterns in dissolved oxygen, pH, and temperature showed consistently negative relationships with *E. coli* patterns in both ponds for the 3 years of the study, while the turbidity, phycocyanin, chlorophyll-a, and dissolved organic matter showed positive relationships in all cases. Correlations between EOF_1 values for *E. coli* and water quality parameters contained the same sign of the relationships and often similar strengths as the correlations performed in the MRD analysis. Both the MRD and EOF analyses show promise to be beneficial for the improvement of monitoring and management of the microbial quality of irrigation water from ponds.

Chapter 4: Accounting for the Three-Dimensional Distribution of *Escherichia coli* Concentrations in Pond Water in Simulations of the Microbial Quality of Water Withdrawn for Irrigation

4.1 Introduction

The microbial quality of irrigation water is a matter of public safety and concern. The World Health Organization reported that an estimated 600 million people in the world suffer from diseases after eating contaminated food every year (WHO, 2015). The consumption of fresh produce has been increasingly viewed as a serious matter influencing food safety and health (Kearney, 2010). The microbial quality of irrigation water is recognized as a substantial factor affecting the contamination of produce, including fresh fruits and vegetables (Pachepsky et al., 2014; Oliveira et al., 2012; Akinde et al., 2016).

Generic *Escherichia coli* (*E. coli*) and enterococci are commonly used as fecal indicator microorganisms for the determination of the microbial quality of irrigation water (Steele et al., 2005; Boehm et al., 2014; Chandrasekara et al., 2015). In the United States, the US Food and Drug Administration (FDA) implements the Food Safety Modernization Act (FSMA) to prevent the risk of foodborne disease (FDA, 2018). As part of the FSMA, the Produce Safety Rule sets standards on the microbiological quality of irrigation water. This rule proposes specific standards of *E. coli* concentrations to evaluate the microbial water quality based on two calculated values: the geometric mean (GM) and the statistical threshold value (STV). The STV is calculated as an approximation of the 90th percentile of the

water quality distribution and no more than 10% of samples used to calculate the GM should exceed this value. Both the GM and STV are calculated using 20 grab samples collected over a 2 to 4-year period. Ponds, impoundments, and agricultural water recovery basins are common sources of irrigation water (Dieter et al., 2015).

The microbial quality of irrigation waters in the USA was essentially unknown until recently (Pachepsky et al., 2014). Since then, survey studies have been conducted in various regions of the United States, including studies on farm ponds and agricultural reservoirs in New York State (Jones et al., 2014), Georgia (Antaki et al., 2016; Gu et al., 2012; Lee et al., 2018), Florida (Havelaar et al., 2017), Virginia (Truitt et al., 2018), Mississippi (Partyka et al., 2016), and California (Partyka et al., 2016). The majority of work was done by taking a single sample from a single depth when sampling irrigation ponds or reservoirs. In many cases, the depth was not reported. Where the sampling depth was recorded, it was either 20 cm (McEgam et al., 2013) or 50 cm (Gu et al., 2015; Lee et al., 2018; Truitt et al., 2018). A depth of 1 m was also used to characterize the microbial water quality near the irrigation intake (Antaki et al., 2016).

Taking more than one sample during a sampling event has been shown to result in a substantial spatial variability of *E. coli* concentrations measured at the same depth across irrigation ponds. Pahl et al. (2013) took three samples across a pond on a biweekly and later weekly basis in the summer of 2009 and reported standard deviations of logarithms of *E. coli* concentrations between 0.25 and 0.5. Pachepsky et al. (2018) took surface samples in 34 locations across an irrigation pond on a biweekly basis in the summer of 2016 and reported standard deviations of

logarithms of *E. coli* concentrations between 0.1 and 0.75. The latter authors also demonstrated that there was a spatial pattern in *E. coli* concentrations during the summer months of June, July, and August, meaning that some areas in the pond mostly had higher than average concentrations of *E. coli* and other areas mostly had lower than average concentrations of *E. coli*.

The spatial variation of fecal indicator concentrations across a waterbody is commonly studied in research on wastewater stabilization ponds (Dias et al., 2017; 2018, Li et al., 2018). It has been shown that considering variations of microbial concentrations in three dimensions is beneficial for understanding the pond efficiency (Maiga et al., 2009). Three-dimensional hydrodynamic modeling has been used to explain functions of wastewater stabilization ponds, such as aerobic and anaerobic carbon and nutrient removal, sedimentation and mixing, and algae growth, among others (Sah et al., 2011; 2012, Khan et al., 2013). To the best of our knowledge, no 3D information on the microbial water quality in irrigation ponds has been reported so far.

We hypothesized that information on the 3D distribution of *E. coli* concentrations in irrigation ponds may be useful for explaining and predicting the temporal variation of the microbial quality ofwater going to the fields. If irrigation water is taken at different times from different parts of a pond, and the *E. coli* concentrations in these parts are different, then one can anticipate differences in the irrigation water quality at different times during an irrigation event.

The objective of this work was to research the 3D distributions of *E. coli* concentrations in an irrigation pond in Maryland and to estimate the dynamics of *E.*

coli concentrations at the water intakeduring irrigation events using hydrodynamic modeling.

4.2 Materials and Methods

4.2.1 Study Area

The study was carried out at an excavated pond located at the University of Maryland Wye Research and Education Center (38°54'58.43"N– 76° 8'28.95"W). The pond is about 22 m wide and 200 m long, with an average depth of 1.5 m (Figure 1). The pond is surrounded by dense shrubs andgrasses, with a few small trees. There are agricultural supply storage facilities and a parking lot on the eastern side of the pond. The crops around the pond receive chemical fertilizers in March and animal manures are not applied. The water level in the pond between irrigation events is controlledby precipitation, as well as by an ephemeral creek that enters through a culvert at the northern-end inflow. This creek directs overland flow from the surrounding corn fields to the pond. The water levelis restricted by a water level-dependent orifice outflow drain that flows to a ponded marsh-like areathat drains into a small creek. The irrigation pipe is located in the pond interior and is indicated in Figure 1. The irrigation water was drawn at a rate of 0.01 m^3/s for durations ranging from 1 to 8 h.

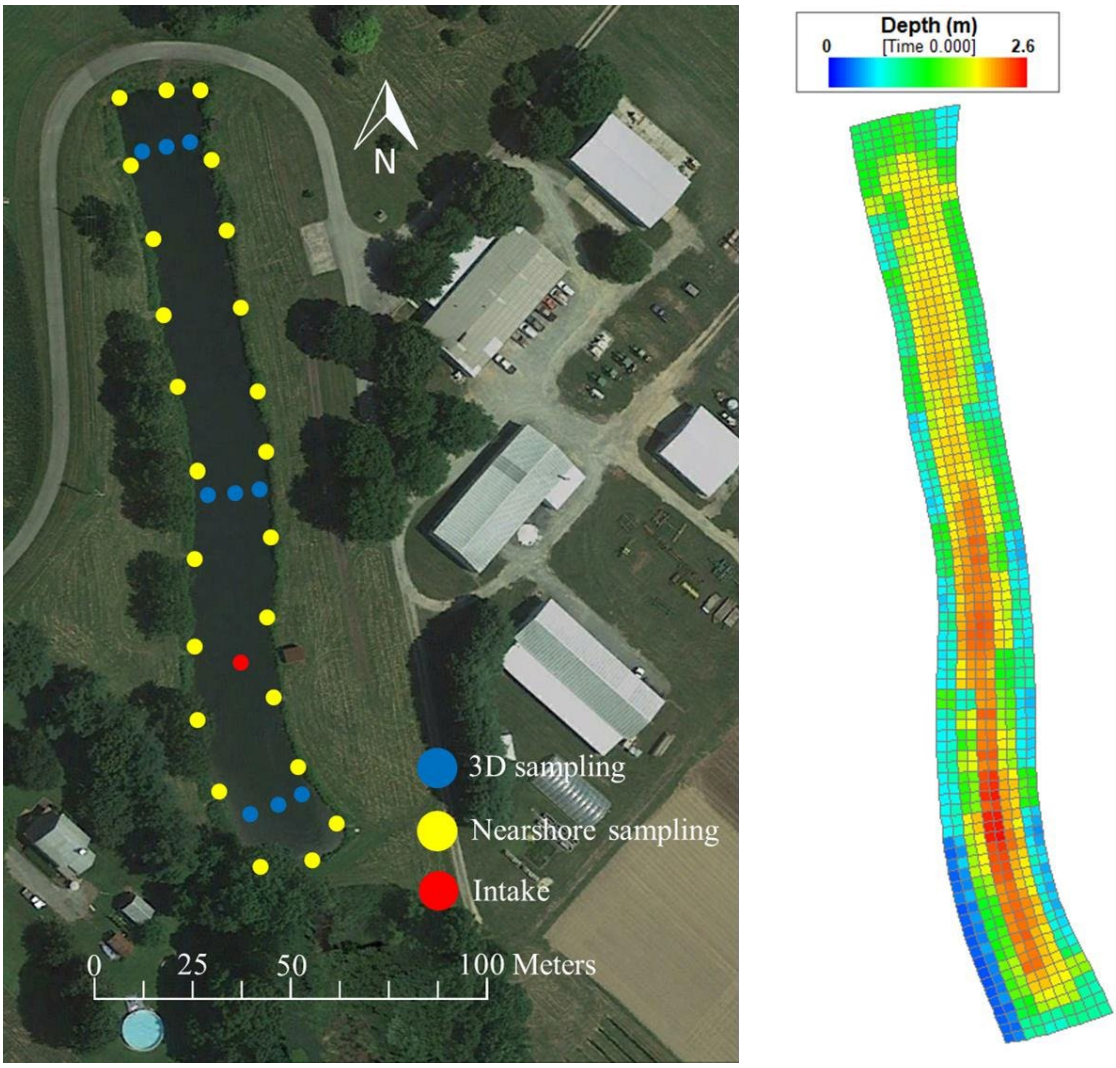


Figure 4.1. Map of the study area showing the locations of 3D sampling for the pond interior and nearshore sampling (left-hand side). The top view of the computational mesh is also shown (right-hand side).

4.2.2 Data Acquisition

Water sampling was conducted between 12:15 and 13:30 on the 24 August 2017. All sampling locations were geotagged with the handheld global positioning system (GPS) device (BE-2300; Bad Elf, Tariffville, CT). Interior water samples were collected using the Mobile Sensing and Sampling Hyper-Spectral Imaging Platform (MSSHIP, Figure S8, Appendix C). Samples were collected along three

transects of the pond, which encompassed the bottom, middle, and top portions (Figure 1). At each sampling position, samples were collected at depth increments of 0.5 m (i.e., 0, 0.5, 1, and 1.5 m). Care was taken to monitor the water depth at each sampling location so that interaction between the sampling pump and the bottom sediments was avoided. For this reason, some locations only provided samples from as deep as 0.5 m, while others provided samples collected to a depth of 1.5 m, which was the maximum sampling depth employed in this experiment. Between each location and sampling depth, the water pump was thoroughly flushed with pond water from the target depth of interest prior to collecting the sample so as to minimize cross-contamination. Bank samples were collected every 20 m along the perimeter of the pond with a 500-mL grab sampler. A total of 26 and 24 interior and bank samples were collected, respectively. The bathymetry data were collected using a handheld digital depth sounder (Vexilar LPS1, Minneapolis, MN). Approximately 200 depth sample points were collected in the pond. Weather data were recorded at the Wye Research and Education Center, which is less than 1 km from the study pond.

The precipitation and maximum/minimum temperature were recorded on an hourly basis, and the wind speed and direction were recorded every 15 min. No rainfall occurred on the sampling date and the closest event consisted of a 57 mm rainfall one week before the monitoring day. The average temperature, wind speed and direction during the sampling period were 22 °C, 1.9 m s^{-1}, and northward (354.1°), respectively.

Membrane filtration was used to enumerate *E. coli* (EPA method 1603, US EPA, 2002). Approximately 100 mL of sampled water was vacuum filtered through

0.45 um filters (Millipore Corp., Bedford, MA, USA), which were placed on modified mTEC agar plates (Difco, Sparks, MD, USA). The plates were placed in a 37 ◦C incubator for 2 h and were then transferred to a 44.5 ◦C incubator for 22 h. After the incubation period, purple colonies were counted as *E. coli.*

4.2.3 Model and Simulation Setup

The environmental fluid dynamic code (EFDC), which was developed at the Virginia Institute of Marine Science, is a comprehensive three-dimensional numerical model (Hamrick, 1992). This model has been successfully applied to a wide range of environmental studies simulating the variation of hydrodynamic and water quality in lakes (Zhao et al., 2012), rivers (Seo et al., 2012), estuaries (Wool et al., 2003), reservoirs (He et al., 2011), wastewater stabilization ponds (Li et al., 2018, Maiga et al., 2009; Sah et al., 2011; 2012; Khan et al., 2013; Luo et al., 2018), and wetlands (Yang et al., 2016). Details of the EFDC have been documented by Hamrick (28). In the EFDC model, fecal coliform bacteria have no interaction with other state variables and only have a die-off term. The kinetic equation is written as:

$$\frac{\partial(FCB)}{\partial t} = -KFCB \times TFCB^{(T-20)} \times FCB + \frac{WFCB}{V} \quad (1)$$

where FCB is the bacterial concentration (CFU 100 mL^{-1}), *t* is the time (day), *KFCB* is the first-order die-off rate at 20 °C, *TFCB* is the effect of temperature on the decay of bacteria ($°C^{-1}$), *WFCB* represents the external loads of fecal coliform

bacteria (CFU 100 mL^{-1} m^3 day^{-1}), and V is the cell volume (m^3). The right-hand side of Figure 1 shows the boundary and corresponding topography of the irrigation pond. The water quality model was set by allocating 960 active cells in the horizontal plane and four layers in the vertical direction. The average cell size was 4.5 m^2.

The kinetic equation was approximated using the finite difference, with a five second time step applied in each computational cell. The irrigation event was simulated by adding the negative component of the water mass balance in the computational cell where the intake was located. This component was equal to the intake rate multiplied by the time step. The interpolation was performed with the EFDC explorer software interpolation routine that uses a Laplace equation filling technique (EEMS, 2020). A five second computational time step was used. Water extraction was simulated by setting a constant sink value corresponding to the actual water extraction rate (i.e., 9.8 L s^{-1}).

Four different *E. coli* concentration distribution scenarios were designed to examine the effectiveness and necessity of 3D information in microbial water quality assessments in the irrigation pond (Figure 2). The first scenario included data from the bank water samples and 3D interior water samples. The second scenario was built using bank water samples and interior water samples taken at the minimal depth (surface samples): the data obtained at the depth 0–15 cm were extrapolated down to the bottom of the pond. The third scenario included the interpolation of nearshore data across the pond and extrapolation of the results as constant values down to the bottom. The fourth scenario differed fromthe third one

by using interior samples instead of nearshore samples. Internal EFDC routines were used for interpolation.

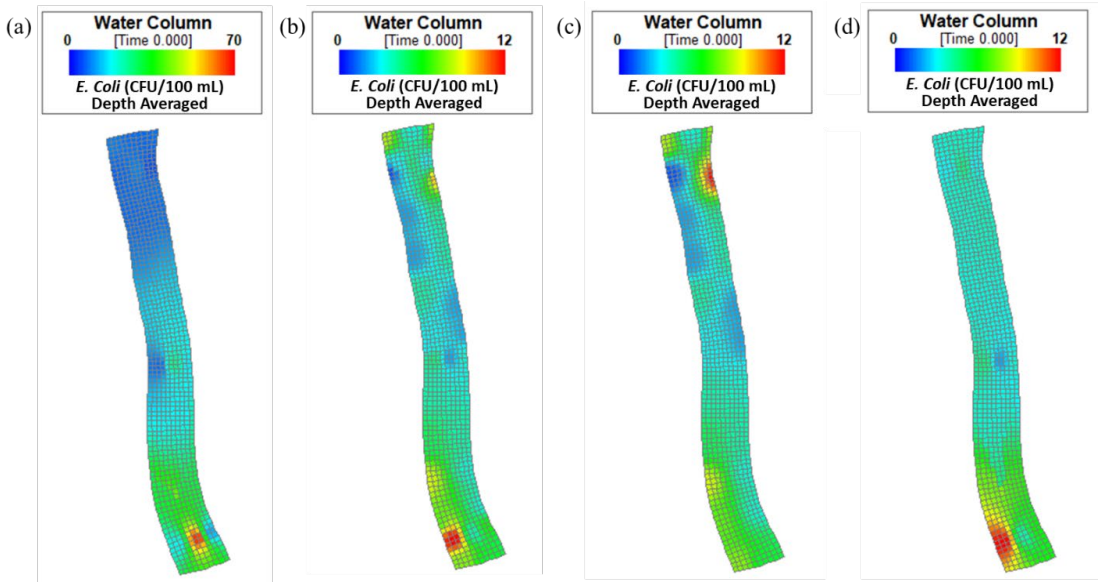


Figure 4.2. Scenarios with the initial Escherichia coli concentration from different sources: (a) 3D sampling; (b) surface layer data (both nearshore and interior) interpolated in 2D and extended over the whole water column; (c) nearshore data interpolated in 2D over the surface layer and extended over the whole water column; and (d) interior surface layer data interpolated in 2D and extended over the whole water column.

A set of scenarios was set for the simulations to examine the temporal variation of the microbialwater quality at the intake, depending on the location of the intake pipe around the pond (Figure 3). The scenarios of different intake locations were designed by moving the intake location at an interval of about 16–25 m from the actual intake position shown in Figure 3. The hypothetical locations corresponded to other sampling sites along the perimeter of the pond.

To assess the zone of influence of the intake pipe, five hypothetical spatial distributions of initial concentrations of an inert tracer were assumed. Each of these

distributions assumed a zero concentration in the section of the pond that included the intake and a concentration equal to one in the section that did not include the intake. The boundaries between the sections were set at locations 2,5, 6, 7, 8, and 9 for the six scenarios, respectively.

E. coli concentrations were log-transformed prior to statistical analysis. Statistics were computedusing the Paleontological Statistics Software Program for education and data analysis (PAST) (Hammer et al., 2001).

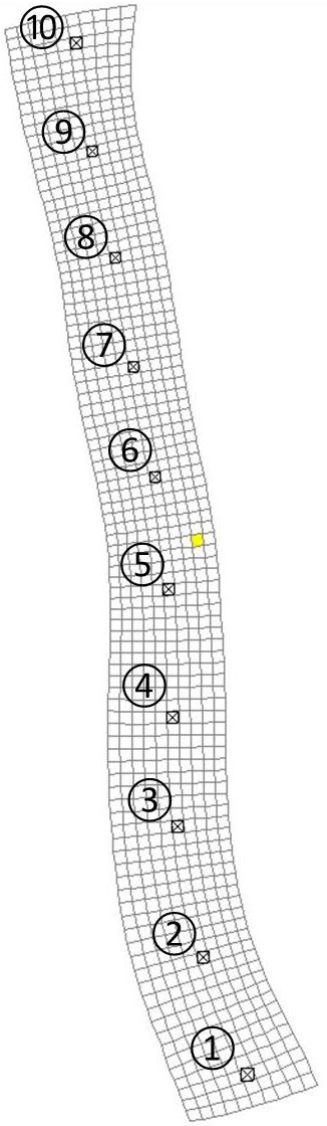


Figure 4.3. Hypothetical irrigation water intake locations. Locations 1–10 correspond with the tenhypothetical irrigation intake positions described in the text. Location 3 is the actual intake location.

4.3. Results

4.3.1 Spatial Variability of *E. coli* Concentrations

Figure 4 shows the lateral distributions of *E. coli* concentrations according to depth. A *t*-test demonstrated that there was no significant difference between the average values of *E. coli* concentrationsnear the shore and in the interior of the pond ($p = 0.491$). Lateral distributions of *E. coli* concentrations at all depths showed higher concentrations in the southern part of the pond than the middle or northern part, except for at the 0.5 m depth. Concentrations of *E. coli* increased with an increasing depth. The average concentrations of *E. coli* in the 0.5 and 1.0 m water depths were about 2 and 4 times higher, respectively, than the average concentration at the surface (0 m). The average concentration at the 1.5 m water depth (bottom layer) was 101.7 CFU 100 mL−1, which was 21 times higher than that of the surface water. An analysis of variance (ANOVA) on the *E. coli* values showed that the concentrations significantly differed by water depth range ($p < 0.001$). Tukey's pairwise testing showed that the 1.5 m depth contained significantly higher concentrations than the 0 m ($p < 0.001$), 0.5 m ($p < 0.001$), and 1.0 m ($p = 0.043$) depths. The 0 and 1.0 m depth ranges also significantly differed ($p = 0.044$), but comparisons of other depth ranges were not statistically significant.

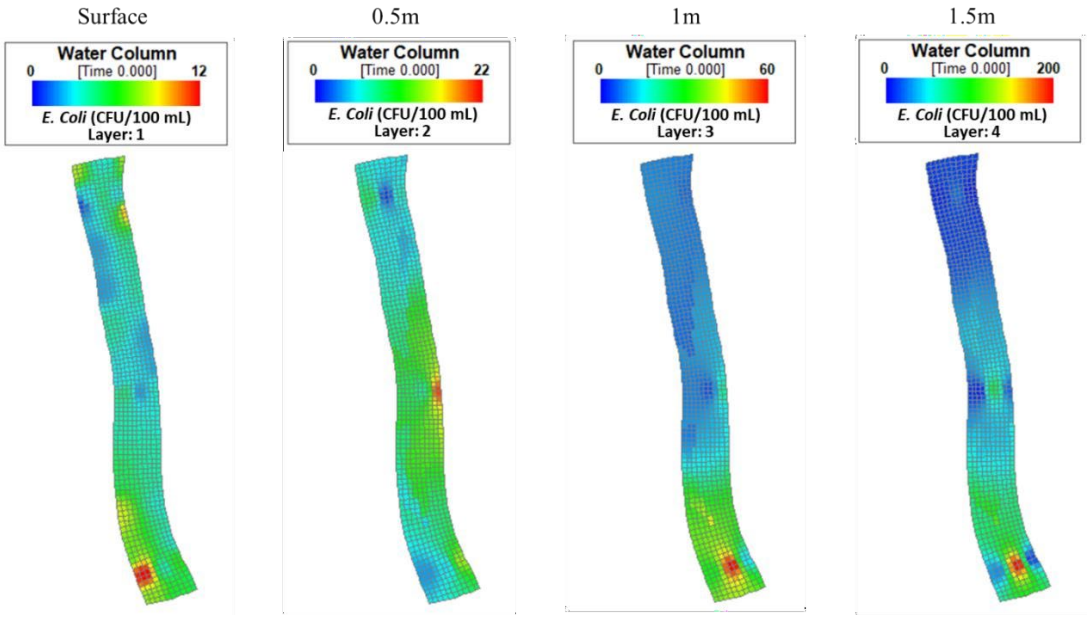


Figure 4.4. Initial *E. coli* concentration at different depths interpolated from the 3D sampling data. Concentrations at the surface (0 m), 0.5 m, 1 m, and 1.5 m depths are shown from left to right.

Figure 5 shows that the predictions of the temporal variation of the *E. coli* concentration in the withdrawn water dependent on the different initial concentration distributions. The simulation results based on the initial distribution using 3D information were substantially different from the simulation results based on initial distributions using only surface water information. In the simulation involving the actual location of the intake pipe (Figure 5. Panel 3), using the 3D dataset yielded an initial concentration of pumped water of 27 CFU 100 mL^{-1} which slowly declined over the 8-h simulation. Using the other three interpolation scenarios in this case all resulted in similar but lower results, which consisted of roughly 5 CFU 100 mL^{-1} in the pumped water over the course of the simulation.

4.3.2 Simulations with Different Intake Locations

Figure 5 shows the temporal variation of *E. coli* in the withdrawn water for several irrigation intake locations. Figure 5 panel 3 presents the change in *E. coli* concentrations of irrigation water taken at the actual pumping location. The other panels represent the change in *E. coli* concentrations when the location of the irrigation pipe is moved upwards along the pond according to the hypothetical locations presented in Figure 3. Simulations involving the 3D data distributions resulted in consistently higher *E. coli* concentrations than the 1or 2D information distributions in all proposed locations. This was particularly true for locations 1–7; however, locations 8–10 contained the smallest differences between scenarios of initial concentration interpolation. Placing the irrigation pipe at location 10 resulted in the most similar result across scenarios.

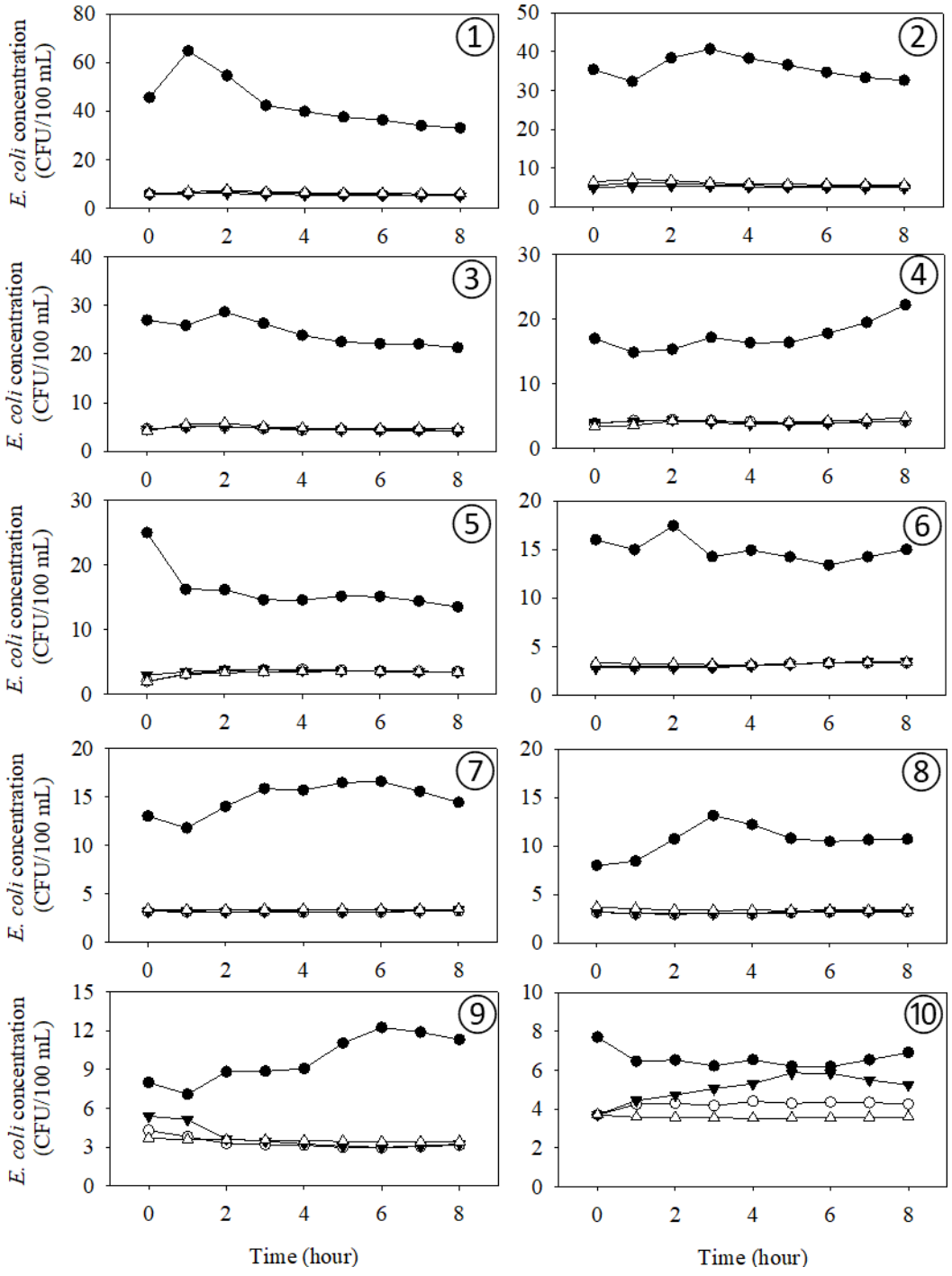


Figure 4.5. *E. coli* concentrations in irrigation water in different irrigation pipe locations under four initial *E. coli* concentration distribution scenarios. ● 3D sampling data; ○ 2D data from surface samples extended to the bottom; ▼2D data from nearshore samples extended to the bottom; and △ data from interior samples extended to the bottom. Panels 1-10 correspond to the hypothetical irrigation intake locations displayed in Figure 3.

4.3.3 Assessment of the Zone of Influence for the Intake

Figure 6a shows the results of setting the boundaries between sections with the tracer being present and absent at different locations within the pond. Figure 6b shows the simulated dynamics of the tracer at the actual intake location, depending on the different microbial contaminated area boundaries under an actual wind direction condition. Scenarios with the boundary between sections at locations 5 to 8 indicated an increase in the *E. coli* concentration during the irrigation event. On theother hand, scenarios with boundaries between sections at locations 2 and 9 maintained a stable concentration during the entire irrigation event time of 8 h. Figure 6c shows the variations of the tracer concentration in the withdrawn water with the wind blowing in the opposite direction to the observed direction. The scenarios with the boundary between sections at locations 2, 5, and 6 demonstrated variation of the *E. coli* concentration during the irrigation event. The other scenarios maintained a zero tracer concentration during the entire 8-h event.

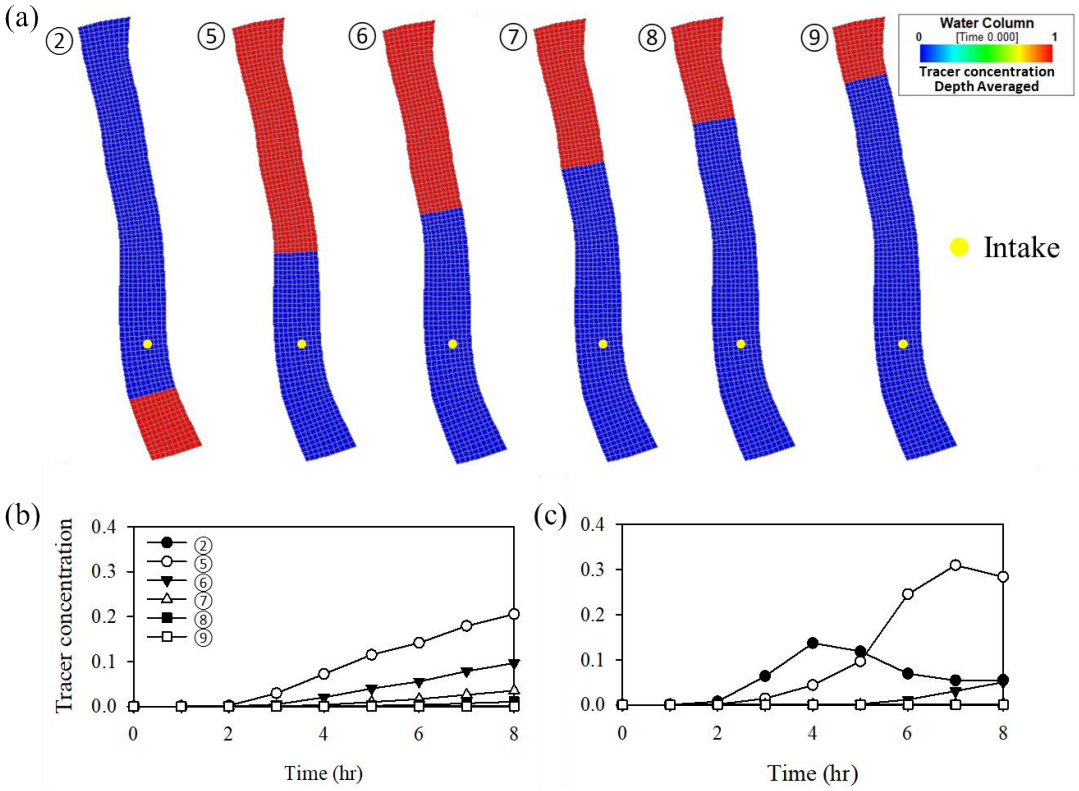


Figure 4.6**.** Tracer distributions and simulation results in scenarios designed to evaluate the zone of influence of the intake. Panel (**a**) shows the boundaries between pond sections with the tracer present (red) and absent (blue). Panel (**b**) shows the simulated concentrations of the tracer withdrawn from the pond at different intake locations over the 8-hour irrigation event. Panel (**c**) displays the same scenario as in (**b**) except with the opposite wind direction.

4.4. Discussion

Both vertical and lateral changes in *E. coli* concentrations were found across the pond (Figure 4). The lateral variations at the 0 to 15 cm depth range were demonstrated for this pond in earlier reports (Pachepsky et al., 2018; Stocker et al., 2019). In these works, the southern part of the pond had concentrations that were persistently higher than other parts of the pond over the summers of 2016 and 2017. The hypothesis was that this hotspot was formed due to the proximity to a residential

property equipped with a septic system, as well as the presence of domesticated pets (Stocker et al., 2019). Vertical gradients of *E. coli* concentrations were well-pronounced in some parts of the pond. A possible reason for gradient formation is the effect of UV radiation. Maraccini et al. (Maraccini et al., 2016) observed that a stronger UVB intensity significantly increased the inactivation rate constants of *E. coli* in an unmixed natural freshwater marsh, a brackish water lagoon, and a marine site. The absorbance of UVB and inactivation rates of *E. coli* likely decreased with depth in this work, thus creating conditions for the formation of vertical gradients of concentrations. The elevated levels of *E. coli* in the southern part of the pond may also be the result of the increased shading that this region receives relative to the middle and northern parts throughout the day. The amount of particulate matter present in the pond would have also likely influenced both the vertical and lateral concentrations of *E. coli*, since it has been shown that particulate and dissolved organic matter may attenuate UV penetration and shield bacteria in water (Smith et al., 2004). Future work may want to investigate the results of simulations when both organic matter and UV radiation measurements are collected and used as model inputs.

The temporal variation of *E. coli* concentrations in the withdrawn water based on 3D information had significantly different patterns compared with results based on three different 2D information scenarios (Figure 2). Under the conditions interpolated using the 3D dataset, *E. coli* concentrations of the intake water were high because of vertical gradients that were detected during sampling. All three scenarios assuming a constant concentration across the depth range gave an estimated *E. coli* withdrawal

concentration that was five times lower than that of the scenario using the 3D information. Therefore, it may not be sufficient to characterize the microbial quality of irrigation water based on *E. coli* concentrations measured in samples taken close to the pond water surface. Additionally, such characterization may mask the possibility that the microbial quality of irrigation water exceeds permitted thresholds when vertical concentration gradients are formed between rainfall periods.

Changing the intake location substantially changed the simulated *E. coli* concentrations and dynamics in the extracted water (Figure 5). The simulated *E. coli* concentration in the extracted water based on 3D information demonstrated higher concentrations than the results of 2D information during irrigation events up to location 9 due to the vertical gradient of *E. coli* concentrations present in the pond. With water uptake at location 10, no difference was found in the results obtained between 3D information and 2D information. Therefore, if the location of the irrigation pipe is moved to location 10 in the studied pond, then the microbial quality of the extracted water may be improved both initially and throughout the course of a multi-hour irrigation event.

The 3D sampling information helps to more realistically predict the temporal variation of the *E. coli* concentration in the extracted water compared to using 2D survey information. However, 3D surveys have disadvantages in terms of time, costs, and labor compared with 1D survey tasks. Therefore, reducing the area in which a 3D survey is performed would alleviate these disadvantages while maintaining a high prediction accuracy of the microbial quality in the extracted water. We have confirmed that the northern-most area from location 9 and the southern-most area

from location 2 do not affect the temporal variation of *E. coli* concentrations in the extracted water taken from location 3 under actual wind direction conditions (Figure 3). The 3D survey information from location 9 to location 2 is sufficient to accurately predict the microbial quality in the extracted water. Furthermore, the decreased survey area will contribute to reducing the time, costs, and labor for subsequent 3D surveys. The results on the selection of the survey area are obviously pond shape-specific. Simulations can be helpful in selecting the parts of a pond that should be researched in a 3D survey.

To examine the impact of the wind direction on the spread of *E. coli* in the studied pond, we compared the intake influence zones which were estimated under the actual wind direction, as well as the opposite wind direction (Figure 6). The survey area was considerably reduced as the wind direction changed from a northerly wind direction to a southerly wind direction (data not shown). The northerly wind encourages more movement of the microbially contaminated water from the upper area to lower areas. Otherwise, the southern wind disturbs the movement from the northern area to the southern area. The wind direction and speed have dramatically influenced water circulation and the spread of microbial contamination in various bodies of surface water (Smith et al., 1999; Zhang et al., 2012; Hatvani et al., 2018). Sokolova et al. (2013) concluded that the microbial contamination from different sources was transported faster or slower to the water intake, depending on the wind speed and direction. Therefore, these simulations demonstrated that the wind direction should be considered when designing a 3D survey area.

We realize that our simulations with a constant inactivation rate only present a first approximation of the fate and transport of *E. coli* in the pond during irrigation. A decrease in the inactivation rate with depth is expected (Maraccini et al., 2016). However, the differences between the inactivation rates at different depths will support and may increase the vertical gradients and possibly lateral variation of concentrations. Quantifying the effect of *E. coli* survival variation in space and time on the variation of the microbial quality of the extracted water presents an interesting research avenue. Future work should also investigate the applicability of the EFDC model for irrigation water sources of different shapes, sizes, and types (such as streams, rivers, or canals). Validating the model by sampling the withdrawn water and pond water over time would also be an interesting venture, but would require intensive 3D sampling, which would be time-consuming and costly. Additionally, the simulation scenarios presented in this work are just a few of many possible ways to configure the model and simulate the irrigation event. Therefore, we encourage future researchers to experiment with differing sample sizes, sampling depths, interpolation methods, wind patterns, and locations of irrigation withdrawal. Finally, it would be interesting to incorporate other measured water quality variables into the model to assess their dynamics during an irrigation event and examine whether levels of other water quality parameters such as dissolved oxygen or pH influence the concentrations of *E. coli* in the withdrawn water.

The results of this work indicate that sampling irrigation water in the field should not be done once. The spatial variability of the microbial concentrations in the water source translates to the temporal variation of concentrations in the irrigation

water reaching the field. The variation of the microbial water quality in the field over time will indicate that there is a spatial variability of the indicator concentrations in the water source. Understanding the reasons for this variability may lead to management actions that will improve the quality of the irrigation water from the source.

4.5. Conclusions

Through simulations using the EFDC model, we were able to estimate the temporal dynamics of *E. coli* concentrations based on hydrodynamic information, such as inflow and outflow rates, and mixing due to wind. *E. coli* concentrations were measured at depth increments of 0.5 m in the interior area and at the water surface along the bank of an irrigation pond. Albeit small, the studied pond is not a homogeneous body of water. Both vertical and lateral changes in *E. coli* concentrations were observed across the pond. Modeling confirmed that the initial 3D distribution of *E. coli* concentrations significantly influenced the dynamics of the *E. coli* concentration in extracted water going to the field. The simulation study confirmed that changing the intake location can change the microbial quality of the extracted water. Moreover, simulations can be helpful in defining the pond sections for which to collect 3D data. The delineation of this section must be done for different realistic wind conditions that may have significant effects on the *E. coli* transport in the pond. The 3D spatial distributions of *E. coli* in irrigation ponds appear to be an underestimated factor which may influence the microbial quality of

water by creating substantial temporal variations. Hydrodynamic modeling can be a useful tool for improving microbial water quality monitoring strategies by considering site-specific environmental characteristics of irrigation water sources.

Chapter 5: Diurnal Variation of *E. coli* Concentrations in Agricultural Irrigation Ponds

5.1 Introduction

Irrigation accounts for 42 % of the total freshwater withdrawals in the United States with most of that coming from surface water sources (Dieter et al., 2015). The microbial quality of these surface waters is immensely important. Pathogens in surface water sources can be transferred via irrigation water to crops and their soils which has led irrigation water quality to be implicated in several multi-state foodborne outbreaks (Greene et al., 2007; Carstens et al., 2019). In recent years, the U.S. EPA listed the leading cause of the impairment of surface waters as being attributed to elevated levels of microbial pathogens (U.S. EPA, 2009). Because pathogens are often difficult to detect and enumerate, fecal indicator bacteria, such as generic *E. coli* which are easier to measure, are used for determining the microbial quality of irrigation sources.

The Produce Safety Rule (PSR) is a part of the implementation of the Food and Drug Administration's Food Safety Modernization Act (FSMA). It specifies the allowable concentrations of *E. coli* in surface water used for irrigation. Under the PSR, growers must sample their irrigation water a minimum of 20 times over the course of a 2- to 4-year period and develop a microbial water quality profile. Failure to meet microbial water quality standards results in the inability to use a water source or the need to undertake corrective measures such as delaying harvest to allow for microbial die-off in the field (Havelaar et al., 2017).

The PSR does not specify where or when to sample an irrigation water source and simply states that samples must be "collected as close as is practicable to harvest" (FDA, 2015). Recent investigations into *E. coli* concentrations in irrigation ponds has revealed that a high degree of spatial variability can exist which may affect the conclusions of microbial water quality assessments (Pachepsky et al., 2017; Lee et al., 2018; Stocker et al., 2019). Temporal variation of *E. coli* concentrations has also been described in the literature. Seasonal variability is routinely reported whereby concentrations are typically higher in the summer months and lower in colder periods (Traister and Anisfeld, 2006; Jeon et al., 2020). *E. coli* concentrations can dramatically increase in response to rainfall/runoff events (Jenkins et al., 2015; US EPA, 2010). However, within-day variability in indicator or pathogen concentrations in irrigation waters is rarely acknowledged or reported in the literature.

Concentrations of *E. coli* in water may change in response to diurnal fluctuations of environmental factors. Diurnal variation in indicator or pathogen concentrations in freshwater bodies typically results in the highest concentrations observed in the morning hours with steady reductions throughout the day and a rebound overnight (US EPA, 2010). These dynamics are most often attributed to increasing solar radiation and temperatures over the course of a day. These factors have been shown to effectively inactivate indicator and pathogen species in microcosm experiments (Solic and Krstulovic, 1992; Sinton et al., 2002; Jenkins et al., 2011; Maraccini et al., 2016) and has been implicated in field monitoring studies (Hughes, 2003; Desai and Rifai, 2013; Maraccini et al., 2016; Lusic et al., 2017; Lothrop et al., 2018). Diurnal increases in dissolved oxygen and pH attributed to

photosynthetic activity of algae and cyanobacteria during the day have also been shown to negatively affect *E. coli* concentrations in water (Liu et al., 2016; Nguyen et al., 2015; Cho et al., 2016; Dias et al., 2017). Concurrent action of environmental stressors can lead to die-off but also can induce a "viable but not culturable" physiological state in *E. coli* and pathogens in which cells lose their capacity to multiply while maintaining metabolic activity in a so-called "survival mode" which may also lead to lower counts of bacteria measured in the afternoon (Han et al., 2020). Concentrations of fecal bacteria in surface waters have been reported to rebound during the nighttime hours possibly due to cellular repair/recovery and reduction in stressors which during the day may result in a negative net population change (Boehm et al., 2009; Desai and Rifai et al., 2013; Dias et al., 2017).

Diurnal or intra-daily variation in *E. coli* concentrations has been reported to occur in numerous recreational waters including streams and creeks (Maey et al., 2006; Traister and Anisfeld, 2006; Stocker et al., 2016), Great Lakes (Whitman et al., 2004), and coastal environments (Boehm et al., 2009; Mudd et al., 2012; Lusic et al., 2017; Wyer et al., 2018). Daily dynamics have also been documented in retention and treatment ponds, and wetlands (Fallowfield et al., 1996; Davies-Colley et al., 2000; Nguyen et al., 2015; Dias et al., 2017). However, we are aware of only one publication that has reported diurnal microbial variations within agricultural irrigation sources. Lothrop et al. (2018) measured the concentrations of *E. coli* and total coliforms in irrigation canals in Southern California and Arizona at four different time intervals including before 9:00, between 9:00 and 12:00, 12:00 and 13:00, and after 13:00. The authors observed decrease in concentrations throughout the day with

the last time interval samples containing about half the *E. coli* concentrations collected before 9:00.

It was demonstrated that differences in water quality may substantially affect daily *E. coli* dynamics. Such differences can be expected in terms of fluctuating dissolved oxygen, pH, and temperature, as well as the accumulation of cyanobacteria and algae, suspended solids, and the degree in which light penetrates the water. The canal water studied in Lothrop et al. (2018) was shallow and non-turbid which would be conducive to die-off of *E. coli* throughout the day. Poor water clarity has been hypothesized to subdue diurnal oscillations of *E. coli* in lake waters and therefore it would be interesting to know if deeper lentic waters can be expected to exhibit diurnal concentration oscillations (Ismail et al., 2015). To date, no data are available on the effect of water quality parameters on *E. coli* concentrations in irrigation ponds

The objectives of this work were to 1) assess the intra-daily variations of *E. coli* concentrations in irrigation ponds 2) summarize the range of inactivation rates, and 3) determine the most influential water quality variables governing the changes in *E. coli* concentrations throughout the day.

5.2 Materials and Methods

5.2.1 Pond Monitoring

Sampling was conducted at three irrigation ponds in Maryland, USA. These irrigation ponds are referred to as P1, P2, and P3 and were chosen because they were either actively or previously constructed and used for irrigation purposes. Sampling of these ponds was conducted between June and September which was considered the

growing season and when irrigation withdrawal would primarily occur. The ponds were sampled at different frequencies with P1 and P2 being sampled 5 and 8 times, respectively, over three years (2019 to 2021) and P3 being sampled 4 times over 2 years (2020 -2021). All ponds were located within a 1-hour drive from the laboratory so that samples could be processed the same day as collection.

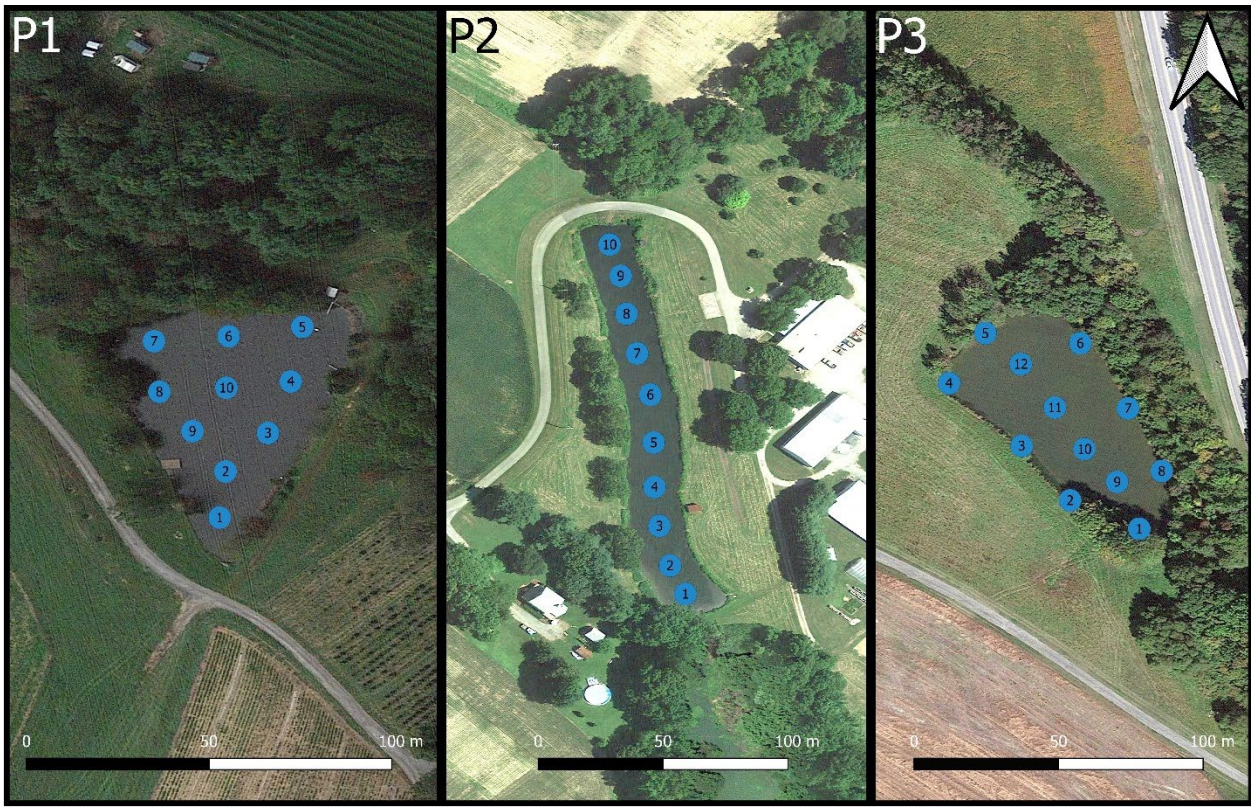


Figure 5.1. Maps of sampling locations used in this study.

Pond 1 is a 4036 m^2 man-made embankment pond with an average depth of 2.7 m (Fig 1). There is vegetation consisting of deciduous trees and shrubs along the Northern and Eastern banks. The remaining banks have grass cover. To the West and Southwest of the pond are crop fields which receive inorganic fertilizers in late April or early May. When the water level in this pond gets low, the farm operators will

occasionally pump water from another pond which is creek-fed. The inflow pipe is located near location 4 and the irrigation outflow is at location 5. Through correspondence with the farm operators, irrigation was not conducted at any pond in the study on days in which sampling occurred.

Pond 2 is located at the University of Maryland Wye Research Center in Wye Mills, MD, USA. This 4197 m^2 pond is an excavated pond with an average depth of 2.7 m and most of the bank is covered with grass and dense shrubs. Large trees are also present along the perimeter but are not close to the water. This pond is surrounded by crop fields, farm buildings, and one residential property. In March, the surrounding fields receive inorganic fertilizer and, no animal manures are used. This pond is primarily fed through rainfall which enters mainly through an ephemeral creek that leads into a culvert near sampling location 10. This culvert tends to have a substantial inflow only when precipitation has recently occurred. On the south end of the pond there is a water-level dependent outflow drain located near location 1. The irrigation pump intake is located near point 3.

Pond 3 is a retired irrigation pond located at the USDA's Beltsville Agricultural Research Center (BARC) in Beltsville, Maryland, USA. This man-made pond is 4,383 m^2 and used to provide irrigation water to the adjacent fields which are primarily corn fields. The Western perimeter of the pond is boarded by a 10-m grass buffer strip whereas the remainder is adjacent to a forested area. To the West and Southwest of the pond are separate fields planted with corn and orchard grass. The surrounding fields receive no nutrient amendments. The pond is runoff-and groundwater-fed with an inflow pipe extending into and hovering above the pond

water surface between locations 1 and 2. The pipe drains the surrounding fields which are at a higher elevation and has a consistent flow of approximately 3 L min^{-1}. A level-dependent outflow is located at sampling site 6 which drains into a riparian creek. The pond depth ranges from 30 cm to 130 cm with an average depth of 90 cm.

Sampling locations in P3 were the same for all observation dates. In 2019, the P1 and P2 sampling sites are what is depicted in Figure 1. In 2020 and 2021 sampling schemes changed and the number of samples was changed from 10 to 5 whereby odd-numbered locations continued to be sampled. Sampling always began within ± 10 minutes of 9:00, 12:00, and 15:00 on each event and the sampling sites are presented in the order of which they were sampled. Sampling was conducted with emphasis to collect all samples as close to the specified times as possible while ensuring proper collection procedures. Sampling always began within ± 10 minutes of 9:00, 12:00, and 15:00 on each event and the sampling sites are numbered in the order of which they were sampled. There was on average a ± 30-minute time span within which sampling and sensing were done across the ponds. For example, some samples at the 9:00 sampling event could have been taken at 8:55 while others were taken at 9:15 or 9:25.

Water samples were always taken at depths of 0 - 20 cm which was considered the water surface. Samples were taken from a small boat and GPS was used to provide consistency of sampling locations between different sampling dates.

After collection samples were immediately placed into a cooler. Samples were then transported to the lab for analysis. Water quality variables were measured in-situ with a YSI EXO-2 sonde (YSI Inc., Yellow Springs, OH) at the same time and

locations as the water samples were collected. The YSI sonde measured temperature (C;°C) (abbreviation; unit), specific conductance (SPC; µS cm^{-1}) dissolved oxygen (DO; mg L^{-1}), pH, fluorescent dissolved organic matter (FDOM) (relative fluorescent units, RFU), chlorophyll-a content (CHL; RFU), phycocyanin (PC; RFU), and turbidity (NTU;NTU). Surface and subsurface photosynthetic active radiation (PAR and sPAR, respectively) measurements were taken in the field with submersible sensors (Apogee Instruments, Logan Utah). The PAR meters measured the photosynthetic photo flux density (µmol m^2 s^{-1}) at the wavelength range from 400 to 700 nm and applied an immersion effect correction factor for underwater readings. Concentrations of ammonia (NH_3; mg L^{-1}), nitrite (NO_2; mg L^{-1}), nitrate (NO_3; mg L^{-1}), and orthophosphate (OP; mg L^{-1}) were analyzed with a SEAL AQ300 discrete nutrient analyzer (SEAL, Minnesota, USA). Total carbon (TC; mg L^{-1}) (inorganic and organic; (TIC; mg L^{-1}) and (TOC; mg L^{-1}), respectively) and total nitrogen (TN; mg L^{-1}) were analyzed on a Vario TOC cube (Elementar Hanau, Germany) using high temperature combustion and tandem TN_b and TC detectors. All nutrient samples were processed according to prescribed methodology provided by the respective companies.

Samples in the lab were kept on ice throughout processing. Laboratory processing involved membrane filtration for *E. coli* enumeration based on EPA method 1603 (U.S. EPA, 2014). Briefly, 50 to 150 mL of pond water was filtered, depending on turbidity, through a 0.45 µm filter (EMD Millipore) which was then aseptically transferred onto petri dishes with modified mTEC media. Agar plates with filters were then incubated at 22 °C for 2 hours followed by incubation for 22 hours at

45 °C. After incubation, purple colonies were counted as *E. coli*. All samples were filtered in duplicate and counts were then averaged.

5.2.2 Data Analysis

E. coli concentrations were decimal log-transformed prior to applying statistical tests. All samples with non-detectable concentrations of *E. coli* (e.g. < 1 CFU 100 mL^{-1}) were assigned a value of 0.5 CFU 100 mL^{-1} prior to log transformation (USEPA, 2009). If water disturbances occurred during sampling (human or wildlife activity) then samples near those locations were not used in the analyses. The Kruskal-Wallis one-way analysis of variance (ANOVA) was used to test if time of day was a significant factor on *E. coli* concentrations. The same test was used to statistically compare medians of the *E. coli* concentrations across ponds. If significance was detected then a pairwise Mann-Whitney post-hoc test was performed to evaluate if median concentrations significantly differed between sample sets collected at 9:00, 12:00, and 15:00. The $\alpha = 0.05$ significance level was used for all statistical comparisons. Intra-daily die-off rates were computed for concentration changes between 9:00 and both 12:00 and 15:00 as well as for 12:00 and 15:00. Rates were calculated by dividing the difference in natural logarithms of the pond-wide average *E. coli* concentration at two time points by the elapsed time in hours. The Kolmogorov-Smirnov test was used to evaluate the equality of continuous probability distributions of die-off rates between ponds. The Spearman correlation coefficient (r_s) was calculated between *E. coli* and water quality parameters to assess the strength and direction of statistical relationships. Correlations coefficients were

calculated using paired point measurements of *E. coli* and water quality parameters. Statistical significance was assessed with $\alpha = .05$.

Regression tree analysis was used to determine the most influential environmental covariates of the *E. coli* concentrations (Stocker et al., 2019). The 'partykit' package (Hothorn et al., 2015) was used in the R statistical software (R Development Core Team, 2015). This package implements a type of regression trees called conditional inference trees (cTREE). Conditional inference trees use hypothesis testing to determine covariates responsible for splitting the dataset into homogenous groups with respect to the output variable (*E. coli* concentrations). This approach resolves complications of overfitting and covariate selection bias associated with traditional regression trees and uses a pre-determined significance level ($\alpha = .05$) as a split-stopping criterion (Guan et al., 2021). The only constraint used in creating the cTREEs was that at least 10 samples must be present in the terminal nodes. The cTREEs for each pond were created using the pooled datasets containing all point observations collected at that study site over the monitoring period.

5.3 Results

5.3.1 Physicochemical water quality parameter changes and correlations with *E. coli* concentrations

The concentrations of *E. coli* in the ponds as well as other water quality variables by time of day are shown in Table S3 for P1 and P2 and Table S4 for P3 and the P3 inflow water, respectively (Appendix B). For all ponds, the levels of DO, pH, and °C increased throughout the day. Other parameters showed inconsistent intra-

daily trends. Changes in DO and pH throughout the day were most pronounced at P2 while the other ponds only experienced small increases. For example, in P2 from 9:00 to 15:00, the average increase of DO and pH throughout the observation period were about 7 mg L^{-1} and 1.6 units, respectively, whereas in P1 this amounted to about a 1 mg L^{-1} and 0.3 unit increase, respectively in P1. For P3 these increases were about 1.85 mg L^{-1} and 0.9 units, respectively, between 9:00 and 15:00. PAR and sPAR levels increased from 9:00 to 12:00 but decreased by the 15:00 sampling time in all ponds. NTU concentrations in the ponds were similar from 9:00 to 15:00 but P2 was found to be in general about 2 times as turbid as P1 and P2 which contained similar concentrations. FDOM concentrations were similar between Ponds 2 and 3 with both having in the range of 2 to 3 times greater concentration than P1. In all three ponds the concentrations of FDOM slowly decreased from 9:00 to 15:00 as did the levels of photosynthetic pigments PC and CHL. Both PC and CHL were several times higher in P2 relative to P1 and P3. Concentrations of nutrients measured in the ponds were low and were found to typically be below the limit of detection. The exceptions were NO_3 and OP in P1 and P2, respectively, which were consistently above detection limits. Nutrient analysis was not consistently performed for P3 but when it was performed all nutrients but NH_3 were below the detection limits.

Of the 18 environmental variables measured relatively few showed significant correlations with concentrations of *E. coli*. In P1, the DO and pH levels showed significant negative correlation with *E. coli* (r_s = -0.714 p = .003 and r_s = -0.646 p = .009, respectively) while concentrations of nutrients such as total dissolved nitrogen, ammonium, and nitrate all showed significant positive correlations (r_s = 0.682 p =

.005, 0.575 p = .025, and 0.761 p = .001, respectively). For P2, DO and pH were also negatively correlated with *E. coli* concentrations (r_s = -0.518 p = .006 and r_s = -0.630 p < .001) while dissolved total organic and inorganic carbon concentrations showed significant positive correlations (r_s = 0.590 p = .005 and r_s = 0.551 p = .010, respectively). In P3, the only covariate significantly correlated with *E. coli* concentrations was specific conductivity (r_s = -0.629 P = 0.028). Temperature was negatively correlated with *E. coli* in all ponds (r_s = -0.207, -0.133, and -0.238 in P1, P2, P3, respectively, all p > .05). The levels of PAR and sPAR were weakly negatively correlated with *E. coli* concentrations in all ponds (r_s < 0.200) and non-significant (P > 0.05). The only exception was with sPAR in P3 which had a moderate negative correlation with *E. coli* (r_s = -0.650 p = .0670).

5.3.2 Intra-daily variation in *E. coli* concentrations

The intra-daily dynamics of *E. coli* at the studied ponds are shown in Figure 5.2 and summarized in Table 5.1. In all samplings at the P1, the highest average concentrations were observed during the 9:00 sampling time. In all but the 7/2/2019 sampling date, concentration magnitudes were in the order of 9:00 > 12:00 > 15:00. The average concentrations in P1 across the study were 0.716 ± 0.336, 0.656 ± 0.265, and 0.636 ± 0.251 (average ± S.E.) log CFU 100 mL^{-1} for the 9:00, 12:00, and 15:00 sampling times, respectively. Time of day was found to be a significant factor in all but the 6/6/2019 and 9/6/2019 sampling dates (p < .05) (Table S5).

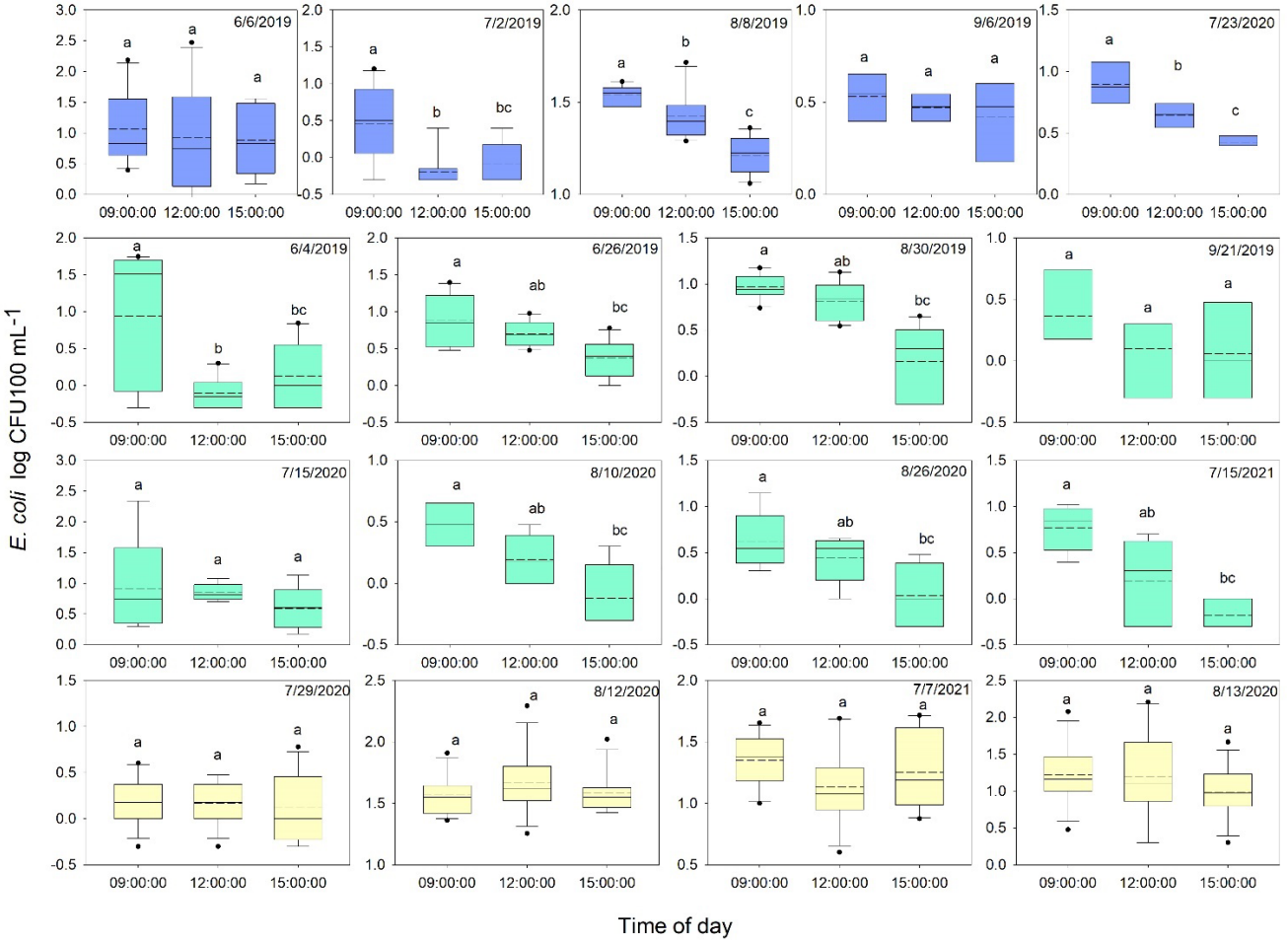


Figure 5.2. Intra-daily changes in *E. coli* concentrations at P1 (top; blue), P2 (middle; green), and P3 (bottom; yellow). Means and medians are displayed with dashed and solid lines, respectively. The upper and lower sides of the boxes represent the 3rd (Q3) and 1st (Q1) quartiles of the *E. coli* distribution with the distance between showing the interquartile range (IQR). The whiskers show maximum and minimum concentrations values not considered outliers. Upper and lower dots show outliers which are calculated as values greater than Q3+1.5*IQR or less than Q1-1.5*IQR, respectively. Means not sharing any letter are significantly different by the Tukey pairwise test at the 5% level of significance.

Similar to P1, P2 showed a trend of highest concentrations in the 9:00 sampling events with a gradual decrease throughout the day with the sampling on 6/4/2019 being the only exception. In 6 of 8 sampling dates, concentrations in P2 were shown to significantly differ by time of day (Table 5.1). Where one-way ANOVA results were significant, Tukey pairwise comparison testing showed that *E. coli* concentrations in P2 during the 9:00 and 12:00 sampling times were typically not significantly different nor were the 12:00 and 15:00 but the 9:00 and 15:00 sampling times were always significantly different. Study-wide average concentrations in P2 were 0.74 ± 0.08, 0.40 ± 0.13, and 0.15 ± 0.08 log CFU 100 mL^{-1} for the 9:00, 12:00, and 15:00 sampling times, respectively.

Table 5.1. *E. coli* dynamics in the studied ponds at-a-glance

Attribute	P1	P2	P3
ANOVA significance of time as a factor ($P < 0.05$) (Table S3)	3 of 5 dates	6 of 8 dates	0 of 4 dates
Time of the highest average concentrations	9:00	9:00	9:00[a]
Time sequence of average concentration decrease	9 > 12 > 15[a]	9 > 12 > 15[a]	9 > 12 > 15[a]
Average reduction of (*E. coli* CFU 100 mL^{-1}) from 9:00 to 12:00	40%	50%	4%
Average reduction of (*E. coli* CFU 100 mL^{-1}) from 12:00 to 15:00	21%	317%	12%
Average reduction of (*E. coli* CFU 100 mL^{-1}) from 9:00 to 15:00	167%	568%	17%
Inactivation rates (k hr^{-1}) from 9:00 to 12:00 (Table S4)	0.187 ± 0.082	0.266 ± 0.087	0.027 ± 0.051
Inactivation rates (k hr^{-1}) from 12:00 to 15:00 (Table S4)	0.065 ± 0.047	0.205 ± 0.071	0.042 ± 0.054
Inactivation rates (k hr^{-1}) from 9 to 15:00 (Table S4)	0.126 ± 0.032	0.235 ± 0.031	0.034 ± 0.020

[a]On all but one observation date. Supplemental Tables S5 and S6 show p-values for ANOVA testing and die-off rates (k) for each date, respectively.

Concentrations of *E. coli* in P3 showed a much weaker intra-daily trend than P1 and P2 (Figure 2). However, concentrations were highest in the 9:00 sampling time than later in the day in all but the 8/12/2020 sampling date where an increase was observed at the 12:00 sampling time. Concentrations were not found to significantly differ throughout the day on any observation date in P3 ($p > .05$). Average concentrations across all dates in P3 were 0.93 ± 0.30, 0.85 ± 0.35, and 0.82 ± 0.31 for the 9:00, 12:00, and 15:00 sampling times, respectively. The average decrease in *E. coli* concentrations from 9:00 to 12:00 in P3 was only 3.6 % with a 16.5 % and 12.3 % decrease for 9:00 to 15:00 and 12:00 to 15:00, respectively.

Measurements of the inflow at P3 were taken on several occasions. When measured, the SPC of the inflow water source was typically between 2 to 2.5 times that of the pond average (Table S4, Appendix B). Levels of DO, pH, C, NTU, CHL, PC were all lower than what was measured in the pond whereas SPC and FDOM were 2 to 3 times higher than in the pond. Concentrations of *E. coli* in the inflow water were often observed to be several times larger than the pond average for the same day. In 2020, *E. coli* concentrations in the inflow water were on average 5 times greater than what was measured in the pond (data not shown). In 2021, inflow water measurements were taken on the same day as intra-daily measurements. Concentrations of *E. coli* in the inflow were 7 times greater on 7/7/2021 and 3.5 greater on 8/13/2021.

When data from specific sampling times (e.g. 9:00, 12:00 or 15:00) were pooled across the study for each pond it was found that median *E. coli* concentrations between ponds did not significantly differ in measurements made at 9:00 ($p = 0.379$) but did for those made at 12:00 and 15:00 (both $p<0.001$). Mann-Whitney pairwise testing showed that for the pooled 12:00 sample sets, P3 contained significantly higher median concentrations than P2 ($p < 0.001$) but for P1 this difference did not meet the level of statistical significance ($p = 0.060$). The difference between P1 and P2 was not significant ($p = 0.122$). Pairwise testing also showed that P3 contained significantly higher concentrations at 15:00 across dates than both P1 ($p = 0.015$) and P2 ($p < 0.001$).

Cumulative probability distributions of *E. coli* die-off rates were largely non-overlapping among study sites (Figure 3). Average values of k in P2 were about 2 to

2.5 larger than in P1 and about 7 times larger than P3 (Figure 5.4 and Supplemental Table S6, Appendix B). Median die-off rates at all ponds between different times of day were not found to significantly differ (P = 0.987, 0.564, and 0.945 for P1, P2, and P3, respectively). The Kolmogorov-Smirnov test showed that distributions of *k* for all ponds were significantly different (P = 0.006, 0.027, and < 0.001 for the comparisons of P1 and P2, P1 and P3, and P2 and P3, respectively). The ordering of *k* values computed between specific times of day (e.g. 9:00 and 12:00 or 12:00 and 15:00) did not show any discernable trend of higher or lower values for any of ponds.

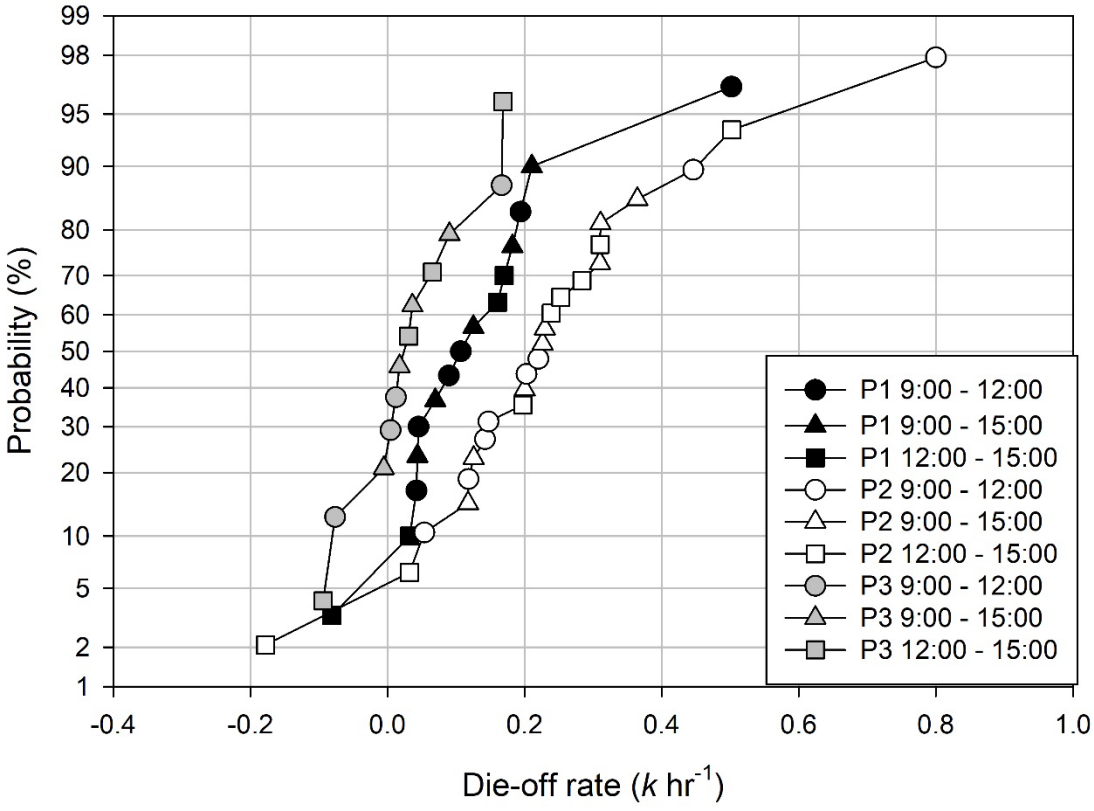


Figure 5.3. Cumulative probability distributions of *E. coli* die-off rates (h^{-1}).

5.3.3 Leading Environmental Covariates

The cTree analysis for the full datasets revealed environmental covariates of the *E. coli* concentrations which split the datasets into the most homogenous subgroups using significance testing of mean values in the resulting split groups. For P1, the sole splitter of the dataset was the level of dissolved oxygen with a threshold value of 10.39 mg L^{-1} (Supplemental Figure S8, Appendix C). Above this threshold, the mean concentration of *E. coli* was close to zero at 0.174 log CFU 100 mL^{-1} whereas below this value the mean concentration was 1.23 log CFU 100 mL^{-1}.

For P2, pH was the leading splitter of the dataset with a threshold value of 8 (Supplemental Figure S9, Appendix C). Below this, NTU created the next split in the dataset with higher levels of NTU associated with higher concentrations of *E. coli.* Above a pH of 8, the concentration of orthophosphate was responsible for creating the next most homogenous subgroups of *E. coli* concentrations with a threshold value of 0.068 mg L^{-1}. With higher orthophosphate concentrations, there was higher observed *E. coli* concentrations.

For P3, the level of SPC created the first split with a threshold value of 90.7 µS cm (Supplemental Figure S10, Appendix C). Below this value NTU was responsible for the next split which resulted in higher *E. coli* concentrations with higher observed turbidity values. Above an SPC of 90.7, the DO concentration was responsible for the next split with a value of 9.07 mg L^{-1}. Below this, temperature made the final split of the P3 dataset with a threshold value of 31.52 °C and low *E. coli* concentrations with warmer temperatures.

5.4. Discussion

Intra-daily variation of *E. coli* concentrations was observed in all three of the studied ponds but to different degrees (Figure 5.2; Table 5.2). P2 showed the strongest intra-daily variation and had the highest average die-off rates while P3 showed the weakest trend. Factors controlling *E. coli* populations appeared to be different in each of the studied ponds.

Solar radiation metrics including PAR and sPAR were not chosen as important in the cTREE of any pond (Supplemental Figures S8 to S10, Appendix C). Levels of PAR and sPAR between all ponds were similar and correlations of these variables with *E. coli* concentrations were generally weak albeit consistently negative. PAR (400 – 700 nm) as well as UV-A (320 – 400 nm) and UV-B (290 – 320 nm) are commonly reported as disinfecting agents in waste stabilization ponds (Dias et al., 2017). The latter two wavelength ranges were not measured in this study but would be pertinent to include in future monitoring efforts. Another possible reason why PAR/sPAR did not appear as important could be the minimal depth in which these, as well as UV-A/UV-B penetrate the water. This penetration is dependent on water properties such as turbidity with higher levels of NTU resulting in lower penetration. Kohn and Nelson (2007) reported a 99% reduction in UV-B and UV-A at depths of only 2.5 and 8.0 cm in waste stabilization pond water whereas Bolton et al. (2011) found that UV-B, UV-A, and PAR were limited to depths of 8, 15, and 43 cm in a facultative pond. In a shallow maturation pond, Dias and von Sperling (2016) found UV-A/UV-B penetration was limited to the top 10 cm whereas PAR was detected at 30 cm. Overall, solar radiation was likely not the single driver of concentration

reduction throughout the day. Like radiation, temperature ranges for the ponds were similar and P3 which had the highest average temperatures showed the lowest declines in *E. coli* concentrations throughout the day.

A possible explanation for the large reductions in P2 relative to the weaker changes in the other ponds is the large increases in both pH and DO throughout the day due photosynthetic activity of phytoplankton species. Levels of CHL and PC were on average 4 to 5 times greater in P2 than the other ponds (Supplemental Tables S3 and S4, Appendix B). The P1 tree showed that DO concentrations were the single most important covariate affecting *E. coli* concentrations while in Pond 2 the most important covariate was pH (Supplementary Figures S8 and S9, respectively, Appendix C). One expects coupled changes in DO and pH in pond water because photosynthesis adds O_2 to the water as well as removes CO_2 which consequently makes the water more alkaline (Dias et al., 2017). Elevated levels of DO have been shown to negatively affect concentrations of fecal bacteria in water particularly in the presence of sunlight which promotes the formation of radical oxygen species (Curtis et al., 1992). These oxygen radicals can cause photo-oxidative damage to bacteria by way of single strand breaks in cellular DNA (Davies-Colley et al., 2000). The survival of gram-negative bacteria in water has been reported to exhibit dependence on the pH level. For example, pH levels above 8.0 or 8.5 (Davies-Colley et al., 2000; Liu et al., 2015), 9.0 (Smallman, 1986; Pearson et al., 1987), 10 (Mendonca et al., 1994; Mayo and Noike, 1996) have all been linked with elevated die-off rates of *E. coli* as well as other indicator organisms and pathogens as compared to ranges closer to neutral.

The levels of FDOM as well as TC, TIC, and TOC were all greater in P2 than in P1. However, in P3 the FDOM levels were on average higher than in P2 which may indicate other variables have greater control over *E. coli* concentrations than suspended materials. Indirect exogenous photooxidative damage has been described as a mechanism for reducing fecal bacteria populations in water. This occurs when dissolved organic or inorganic materials in water absorb and transfer light energy which can cause physical or chemical alteration of other materials and the creation of radical oxygen species (Dias et al., 2017). Mechanisms for the within-day die-off are likely site-specific and identifying the relative role of specific factors is made difficult due to variables changing in concert throughout the day. Levels of FDOM may also be responsible for reducing light penetration and scattering in the studied ponds as well as levels of algae and cyanobacteria which reduce water clarity (Sukias et al., 2001).

Turbidity (NTU) was responsible for the left split of the cTree of the P2 dataset with greater NTU associated with higher *E. coli* concentrations (Fig. S2). Turbidity (NTU) has been well documented to positively correlate with *E. coli* concentrations in surface waters with shielding being the most likely factor for the direction of the relationship as well as the co-transport of suspended particulates and *E. coli* in sediments and runoff (Stocker et al., 2016; Davies-Colley et al., 2018; Weller et al., 2020). The P2 right split was determined by concentrations of the nutrient orthophosphate. Microcosm studies focusing on the response of *E. coli* and fecal coliforms to nutrient additions have shown that orthophosphate and other forms of phosphorous can dramatically stimulate growth rates in water (Chudoba et al.,

2013; Shelton et al., 2014) which could mean that die-off was reduced due to concentrations of orthophosphate that led to growth in *E. coli* populations.

The P3 dataset was first split by levels of SPC. This observation may be explained by the constant inflow water which contained higher concentrations of both *E. coli* and SPC relative to that in the pond. This constant influx of *E. coli* into the pond may also explain the weak intra-daily trends that were observed at this study site. It is possible that the die-off rates in the pond were slightly larger than the replenishment of *E. coli* into the pond from the inflow water source. Identification of inflows to agricultural ponds appears important as these can act as a constant source of microorganisms and may also result in localized hot spots of fecal bacteria (Stocker et al., 2021). We cannot rule out the effect of undiscovered inflows in P1 or P2 as a source of *E. coli* to the ponds which may have affected the trends observed in the current study.

Temperature (C) was also responsible for splitting the dataset into terminal nodes of the regression tree with elevated temperatures equating to lower *E. coli* concentrations. Meta-analyses focusing on the dependency of fecal indicator and pathogen survival on temperature in a variety of surface water types (e.g. agricultural, pristine, river etc.) have shown overwhelmingly negative associations between fecal bacteria and temperature (Blaustein et al., 2013; Pachepsky et al., 2014).

Probability distributions of intra-daily die-off rates for *E. coli* concentrations were found to significantly differ across ponds (Figure 5.3; Supplemental Table S6, Appendix B). Micro- or mesocosm and or field data which reports diurnal *E. coli* die-off rates is scarce in the literature as most small-scale studies are run with constant

conditions and do not incorporate temperature or solar oscillations. In this work, average *k* values for P1 and P2 were similar to those reported by Struck et al. (2008) for *E. coli* in outdoor pond mesocosms with reported values of 0.148 hr^{-1} and 0.196 hr^{-1} for June and July, respectively. Similar rates (0.15 – 0.29 *k* hr^{-1}) were reported in the work of Desai and Rifai (2013) for a watershed stream which was sampled throughout the day. Another study which measured *E. coli* concentrations in 16 Colorado streams and rivers at 2 hour intervals from 8:00 to 16:00 reported die-off rates between 0.125 and 0.25 hr^{-1} for most study sites (Burke, 2018). P3 demonstrated very low die-off rates compared to the other ponds which was likely due to an influx of *E. coli* from the inflow water entering the pond. The rates for P3 were very similar to the 0.0324 hr^{-1} rate which was reported for lake waters (Ge et al., 2012) and lake and reservoir waters (0.010 hr^{-1}) (Blaustein et al., 2013). It is possible that the larger variation in both DO and pH in P1 and P2 are responsible for the higher die-off rates. It should be noted that comparison of rates across different scales of studies (e.g. lab, small- or field-scale) must be done cautiously especially since lab and small scale studies repeatedly sample the exact same water whereas in rivers, ponds, or lakes there is constant movement of water.

E. coli die-off rates have been calculated in this study using the pond-wide averages. Alternatively, measurements at each sampling location at different times could be used to generate information on local die-off rates. The common practice of die-off rate computations presumes measurements are made to the same volume of water. In the field, water movement and mixing excludes the possibility of having the same volume of water present at different times in the same location. Therefore, the

"local" die-off rates would include the effect of both local *E. coli* survival factors and water flow. The effect of local survival factors can be estimated using the methodology of Jenkins et al. (2011) who deployed floating anchored microcosms for the study of fecal microorganism die-off in agricultural pond water. The die-off rates were found for the volumes of water contained in microcosms rather than for the pond-scale as it was done in this work. It should be noted that this study only documented changes in *E. coli* concentrations over a part of the daytime hours. It would be interesting to expand the sampling window by several hours in either direction and/or to collect samples at intervals which span the course of 24 hours or more. Such information would provide insight into the night-time concentration rebounds.

Limitations of this work are that only 3 irrigation ponds were studied which were all located in the same region of the United States. Future work should investigate the intra-daily or diurnal variability in *E. coli* or other fecal bacteria concentrations in other regions where temperature, solar radiation regime, and microbial community structures may be substantially different. Additionally, all sampling in this study was performed by taking surface (0 to 20 cm) grab samples. The intra-daily dynamics of *E. coli* at different depths may be expected to vary less than near the surface due to reduced influence of effects of temperature and solar radiation. After all, most of the water used for irrigation from ponds is not pulled from the surface, rather it comes from the greater volume of water which is typically in the center where depths are greater and therefore understanding concentration dynamics at depth would be worthwhile to elucidate. The concentrations of *E. coli*

encountered in this study were low. It would be interesting to locate and study irrigation ponds which have higher concentrations ranges than those reported here to determine if there is a dependency of die-off on the initial concentrations.

Overall, the results in this work show that significant variation in *E. coli* concentrations can occur in agricultural irrigation ponds throughout a day and to varying degrees this is dependent on a wide range of environmental factors. Intra-daily variation in *E. coli* concentrations can affect the conclusions of microbial water quality testing with regard to the regulatory standards. The authors of this paper therefore agree with the recommendations put forward by Liu et al., (2006) and Whitman et al. (2004) for Great Lake monitoring and that put forth by Lothrop et al., (2018) for Southwestern irrigation canals that sampling should occur in the morning hours when concentrations of *E. coli* are typically the highest. Such an approach would provide the most conservative estimate on the microbial quality of a waterbody whereas sampling in the afternoon hours may give an overly optimistic result which may not correctly characterize the risk of using an irrigation source.

5.5. Conclusions

Intra-daily variation of *E. coli* concentrations was encountered in all studied ponds. The most consistent observed trend was the concentration change in the order of 9:00 > 12:00 > 15:00. Die-off rates were appreciably different for each pond which may be related to the relative levels of physicochemical water quality parameters present in each pond and the degree in which they changed between sampling times. The cTREE analysis showed that levels of dissolved oxygen, pH, turbidity, specific

conductance were the best splitters of the datasets with respect to *E. coli* concentrations. Dissolved oxygen showed a strong negative association with *E. coli* in when DO was greater than 10.39 mg^{-1} in P1 but in P3 there was a weaker relationship with a lower threshold value (9.07 mg L^{-1}). The concentrations in P2 were largely affected by the pH level with differences occurring above and below a value of 8.0. Turbidity was positively associated with *E. coli* concentrations in P2 and P3 but the threshold values which created the splits was different for each pond (14.41 NTU and 5.28 NTU for P2 and P3, respectively) and contained different average concentrations in the resulting nodes. Design of future microbial water quality monitoring studies should account for the importance of sampling time within a day and document the most influential controls over *E. coli* concentrations in irrigation ponds which appear to be site-specific.

Chapter 6: Prediction of *E. coli* Concentrations in Agricultural Pond Waters: Application and comparison of Machine Learning Algorithms

6.1 Introduction

Food safety is a fundamental public health concern which is threatened when waters with poor microbial quality are used for the irrigation of fresh produce. In the U.S. and around the world, regulatory or advisory thresholds on the microbial quality of irrigation waters are based on the concentrations of *Escherichia coli* (*E. coli*) measured in the water source (US Food and Drug Administration., 2015; Allende et al., 2018; Wen et al., 2020). Irrigation with water of substandard microbial quality has been implicated with foodborne outbreaks associated with the consumption of contaminated produce (Nygård et al., 2008; Kozak et al., 2013; Gelting et al., 2015). Additionally, it is known that pathogenic microorganisms transferred with irrigation water can internalize into crop tissues which extends their persistence and reduces the efficacy of post-harvest washing (Solomon et al., 2002; Martinez et al., 2015).

Ensuring that water used for irrigation meets the recommended criteria is vital for protecting public health and reducing incidences of foodborne outbreaks. Concentrations of fecal indicator organisms, primarily *E. coli*, are commonly used to characterize microbial water quality. Researchers have investigated the relationships between fecal microorganisms and water quality parameters such as dissolved oxygen, pH, turbidity, and nutrient levels with a goal of improving timeliness and predictability of microorganism concentrations in a water source (Francy et al., 2013;

McEgan et al., 2013; Stocker et al., 2019). Such dependencies have varied considerably across studies. This may at least partially be explained by the complexity and non-linearity of relationships of fecal microorganisms with multiple water quality parameters, which in turn exhibit complex relationships.

Machine learning (ML) algorithms have been extensively shown to outperform traditional multivariate analyses in numerous aquatic ecology studies where the two analyses have been compared (Quetglas et al., 2011). The advantage of machine-learning methods is in their ability to mimic linear relationships between dependent and independent variables. Machine learning models are also able to assess associations and quantify predictability in the absence of the knowledge needed for developing process-based models (Thomas et al., 2018). Within the field of microbial water quality, several researchers have used ML algorithms to develop models which could be used to make rapid water quality determinations in rivers, streams, Great Lakes beaches, groundwater, drinking water wells, and water distribution systems (Brooks et al., 2016; Mohammed et al., 2018, 2021; Panidhapu et al., 2020; Tousi et al., 2021; Weller et al., 2021; White et al., 2021).

A large number of ML algorithms have been proposed and implemented in a variety of different research disciplines (Kuhn and Johnson, 2013). Common research goals are (a) to obtain an accurate predictive relationship between the predictors and target variables, and (b) to determine and rank the most influential predictors. By achieving these goals, researchers may be able to eliminate unimportant predictors from measurement programs which can potentially save a great deal of time and resources. The application of ML regressions in the field of microbial water quality is

relatively new (Park et al., 2015; García-Alba et al., 2019; Stocker et al., 2019; Abimbola et al., 2020; Ballesté et al., 2020; Li et al., 2020; Belias et al., 2021; Wang et al., 2021). For agricultural waters, research into predicting *E. coli* concentrations using ML was done for streams (Weller et al., 2021), but so far no studies have utilized ML regressions to predict *E. coli* concentrations in agricultural irrigation ponds which serve as an important source of irrigation water across the United States and abroad.

The objectives of this work were (i) evaluate and compare the capabilities of several popular ML algorithms for predicting concentrations of *E. coli* from water quality parameters in irrigation ponds, and (ii) to determine the most influential predictors for the estimation of *E. coli* concentrations using a multiyear dataset.

6.2. Materials and Methods

6.2.1 Field Sites and Data Collection

Two working irrigation ponds in Maryland were sampled during the 2016–2018 growing seasons. Sampling typically occurred on a biweekly schedule between May and August. Specific details of the sampling procedures can be found in Pachepsky et al. (2018) and Stocker et al. (2021). Briefly, each pond was sampled in a grid-like pattern the maps of which are shown in Supplementary Figure S11 (Appendix C). Pond P1 is located in central Maryland and provided irrigation water to the surrounding fruit fields. The northern part of the pond is surrounded by a forested area whereas the other two sides are covered by grasses or bushes. The fields around P1

received chemical fertilizers prior to planting each year in late March or early April.

Pond P2 is an excavated pond located on the University of Maryland's Wye Research and Education Center (WREC) which is located on Maryland's Eastern Shore. The pond receives water from a culvert at the northern end (location 12 on the map). The pond is surrounded by dense brush vegetation along the perimeter as well as by several trees planted further up from the banks. To the northern end of the pond is a small riparian area that surrounds an ephemeral creek while the southernmost portion has a wetland area that leads into the Wye River.

Along with each water sample that was collected, a YSI sonde was used to determine characteristics of the water quality in that sampling location. In 2016, a YSI MPS 556 (Yellow Springs Instruments, Yellow Springs Ohio) unit was used to measure dissolved oxygen (DO), pH (pH), specific conductance (SPC), and temperature (C) which were measured in-situ. At the laboratory, a Lamotte turbidimeter was used to measured turbidity (NTU) of the samples. In 2017, a YSI EXO 2 was used to measure all of the previously described water quality variables as well as the concentrations of chlorophyll (CHL), phycocyanin (PC), and fluorescent dissolved organic matter (*f*DOM). In 2018, the same YSI EXO 2 sonde was used but additional laboratory measurements were performed. These included ammonium (NH_4^+), orthophosphate (PO_4^-), total nitrogen (TN), and total carbon (TC). Ammonium was measured using an ion-selective probe (CleanGrow, United Kingdom) which was calibrated prior to analysis which occurred on the same day as sample collected. Orthophosphate was run on a SEAL AQ300 discrete nutrient analyzer (SEAL Analytical, Mequon, Wisconsin). Total carbon and TN were

analyzed on a Vario TOC cube (Elementar Hanau, Germany) using high temperature combustion and tandem TN and TC detectors.

E. coli enumeration was performed based on EPA method 1603 (US Environmental Protection Agency., 2005) which utilizes membrane filtration. Briefly, 100 ml of pond water was vacuum filtered through 0.45 µm membrane filters. Filters were then placed onto modified mTEC agar (BD Difco, Sparks, MD) and incubated for 2 h at 37 °C and then 22 h at 44.5 °C. After incubation, colonies that were purple in color were counted as *E. coli*. Each sample was duplicate plated and the resulting counts were then averaged.

6.2.2 Machine Learning Algorithms and Implementation

Several ML algorithms as well as a multiple linear regression (MLR) were compared in this work. The stochastic gradient boosting algorithm (SBG) builds the prediction model from an ensemble of weak models which in this case are decision trees (Friedman, 2002). Models are built in a step-wise fashion where at each step a weak model is fitted to a subsample of the training data drawn at random without replacement. The term "gradient boosting" comes from the fact the model is trying to minimize a loss function by tweaking parameters until a minimum value is reached. The R package "gbm" (Greenwell et al., 2020) was used develop SGB models. Parameters for the SGB algorithm are the number of trees (n.trees), the number of splits in the trees (interaction.depth), the learning rate (shrinkage), and minimum number of observations in terminal nodes of trees (n.minobsinnode).

The k-nearest neighbors (kNN) algorithm implements a non-parametric approach which computes distances from test datasets to the neighboring training

datasets and uses these distances to determine the predicted value for the test dataset. The distances are computed from predictor values for training and test datasets. The number of neighbors (k) used for the prediction is the single parameter for this algorithm. A Euclidean distance measure was used to determine nearest neighbors. The "kknn" package (Schliep et al., 2016) was used to develop kNN models.

The support vector machines (SVM) algorithm finds the global minimum in the predictor (Cristianini and Shawe-Taylor, 2000). Support vector machines automatically select their model size and prevent overfitting by using special form of the regression cost function that balances accuracy and flexibility (Vapnik et al., 1995). Support vector machines neglects small errors which makes it robust and computationally treatable. It employs mapping to use linear regression while the relationship between original (not mapped) predictor and output variables is non-linear. The "kernlab" package (Karatzoglou et al., 2019) was used to run the SVM algorithm with a radial basis function kernel which has two control parameters. The γ parameter defines how far the influence of a single training example reaches and the parameter C controls the overfitting prevention.

The random forest (RF) algorithm creates predictions by generating many decision trees and combining their predictions in a weighted average giving the final prediction. The RF algorithm also has a built-in mechanism for preventing overfitting by random selection of inputs for the individual trees. As implemented in the ranger package (Wright and Ziegler, 2017), the algorithm includes three control parameters. The mtry parameter controls overfitting by determining the number of variables to randomly select at each split in the trees. The min.node.size parameter sets the

minimum number of observations in a terminal node. The number of trees was kept at 500 to reduce computational intensity and because out-of-bag error did not appreciably change after this number of trees.

One of the outcomes of running RF algorithms is determining the most influential features that effect the model output. A random-forest based recursive feature elimination (RF-RFE) algorithm was applied to each dataset to determine the most influential predictors. The result of this procedure is to (i) find the subset of predictors with the minimum possible generalization error and (ii) to select the smallest possible subset of predictors which provide the optimal accuracy in model performance (Granitto et al., 2006). Within the algorithm, at each iteration feature importance is calculated based on the overall effect on the residual error and then the least important predictors are removed. The recursion is needed to address the problem that for some measures of relative importance, the results can change substantially over different subsets of the entire predictor list.

All ML algorithms as well as a MLR model were applied within the "caret" (classification and regression training) R package (Kuhn, 2008). This package contains functions that streamline training for complex ML regression and classification problems. The package facilitates the optimization and execution of ML algorithms and uses other R packages as functions for creating models. For each of the above-listed ML algorithms, we used the "caret" package to perform repeated cross-validation when fitting the ML models to the datasets. A default 10-fold cross-validation was performed with five repeats and then the results were averaged. The "caret" package was also applied to perform the recursive feature elimination.

Algorithm tuning was performed during cross-validation and tuning was performed to minimize the average root mean squared error RMSE. The "caret" package contains a grid search function for control parameter tuning which was utilized in this study. The optimal control parameters as well as most influential variables were identified during cross-validation.

6.2.3 Comparison of Algorithm Performance Metrics

The average root mean square error $\overline{RMSE}$, coefficient of determination $\overline{R^2}$, and mean absolute error $\overline{MAE}$ were the metrics used to evaluate algorithm performance in this study. Averaging of *RMSE*, R^2 and *MAE* was done across 50 values of those statistics obtained for all cross-validation folds and repeats. The probability of the averages being the same for a pair of algorithms was determined from the corrected Student statistics t_c:

$$t_{c,RMSE} = \frac{\overline{RMSE}_1 - \overline{RMSE}_2}{\sqrt{h\sigma^2_{RMSE_1 - RMSE_2}}}; \quad t_{c,MAE} = \frac{\overline{MAE}_1 - \overline{MAE}_2}{\sqrt{h\sigma^2_{MAE_1 - MAE_2}}}; \quad t_{c,R^2} = \frac{\overline{R_1^2} - \overline{R_2^2}}{\sqrt{h\sigma^2_{R_1^2 - R_2^2}}}; \tag{1}$$

where subscripts "1" and "2" refer to the first and the second compared algorithms, $\sigma^2_{RMSE_1 - RMSE_2}$, $\sigma^2_{MAE_1 - MAE_2}$, and $\sigma^2_{R_1^2 - R_2^2}$ are variances of differences between the values of *RMSE, MAE,* and R^2 , respectively, obtained for the same fold and repeat, h is the variance correction term proposed by Bouckaert and Frank (2004) to account for the fact that that values of the metrics in the 50 individual random samples are not independent as they are obtained by the random subsampling of the same dataset. The value of h is determined as

$$h = \frac{1}{k \cdot r} + \frac{n_{testing}}{n_{training}} \tag{2}$$

where k is the number of folds and r is the number of repeats, $n_{1training}$ instances are used for training, and the remaining n_{testing} instances for testing in each of runs. The t_c statistics in (1) have the Student t distribution with $k \cdot r - 1$ degrees of freedom. The value of the ratio $n_{\text{testing}}/n_{training}$ was 0.1 in this work as recommended by Bouckaert and Frank (2004). Having the value and knowing the distribution of t_c, one can estimate the probability of the differences between the average metrics from two models being equal to zero. Bouckaert and Frank (2004) referred to this test statistic as the "corrected repeated k-fold cv test".

Normalized root-mean-square-error (*NRMSE*) and mean absolute errors (*NMAE*) were also computed by dividing the $\overline{RMSE}$ and the $\overline{MAE}$, respectively, by the range of *E. coli* concentrations (e.g. maximum – minimum concentration) and then multiplying by 100 for each year and predictor set. *NRMSE* and *NMAE* show the percentage of algorithm error relative to the spread of data.

6.2.4 Data Preprocessing and Analysis

E. coli count data was log-transformed prior to statistical analysis. All observations of 0 CFU 100 mL^{-1} were assigned a value of 0.5 to facilitate the log-transformation (USEPA, 2015). Rows with missing values were removed prior to analysis. Data was not normalized or standardized before analysis. Preliminary findings showed that these operations did not substantially affect algorithms performance and in many cases resulted in poorer predictions.

To examine model performance and variable importance using different predictor sets we created three different predictor scenarios. These included set A which is DO, pH, SPC, NTU, and C, set AB which is set A plus *f*DOM, PC, and

CHL, and set ABC which is set AB plus $(PO_4)^{3-}$, NH_4^+, TN, and TC. Models with the Set A were developed for the individual years 2016 to 2018 and the combined 3-year dataset. Models with Set AB predictors were developed for 2017 and 2018, and models with the set ABC predictor were built only for 2018 where all 12 of the parameters were measured.

A separate study was performed to evaluate the models developed with the combined P1 and P2 datasets as opposed to models developed with separate P1 and P2 datasets. Only the predictor set A was evaluated in this exercise because these predictors were present in all years of observations and across ponds. The combined dataset was modeled with and without the introduction of a categorical variable "site" that labeled the data from different ponds.

6.3. Results

6.3.1 Summary of Monitoring Data

The P1 2016A, 2017A, and 2018A datasets contained 50, 126, and 138 samples, respectively, after row removal due to missing values. The P1 2017AB and 2018AB datasets were 126 and 138 samples, respectively, while the 2018ABC dataset was 92 samples after row removal. For P2, the sample set sizes were 97, 148, and 202 samples for 2016A, 2017A, and 2018A, respectively and the 2017AB and 2018AB had the same dimensions as the A scenario. The P2 2018ABC dataset contained 202 samples.

E. coli and other water quality variable concentration averages and standard errors are shown in Supplemental Table S7 (Appendix B). The two ponds contained similar concentration ranges of *E. coli* in general although the P1 2017 dataset

contained consistently higher concentrations. The 2016 datasets for each pond contained higher amounts of missing values of *E. coli* concentrations compared to the 2018 and 2018 datasets which had relatively few missing values (<5 %).

Values of most of the water quality parameters were similar between ponds with a few exceptions. Pond 2 in 2018 and 2018 had elevated DO concentrations compared to other instances. CHL and PC were also several times higher in P2 than in P1 in 2017 and 2018. The 2016 NTU concentrations at P2 were considerably higher than in other cases. Orthophosphate concentrations were about 16.5 times greater at P2 than in P1 in 2018. However, P1 had an average NH_4^+ concentration which was about three times greater than in P2. Average SPC values varied between 142.59 and 166.95 across the two ponds. Average *f*DOM values were higher in P2 in 2017 and 2018 than P1 for each corresponding year.

6.3.2 Evaluation of Machine Learning Algorithm Performance

The $\overline{RMSE}$ values and standard errors of RMSE are presented in the Table 6.1. As expected, the MLR performed substantially worse than the ML algorithms for both ponds and for all years and predictor sets. The differences between the performance of the ML algorithms were less substantial. Overall, the differences among average SVM, RF, and SGB $\overline{RMSE}$ for the same ponds and years were less than 10%. The kNN demonstrated relatively larger spread of differences between its RMSE and $\overline{RMSE}$ of other ML algorithms and the range of those differences was from 1.5% to 24.9%.

Table 6.1. Average root-mean-squared errors (RMSE) of logarithms of *E. coli* concentrations predicted with four machine learning algorithms and multiple linear regression.

	Predictor set A				Predictor set AB			Predictor set ABC
ML Algorithm	**2016**	**2017**	**2018**	**2016-2018**	**2017**	**2018**	**2017-2018**	**2018**
				Pond P1				
SGB	**0.247±0.011**	**0.250±0.012**	0.354±0.015	0.343±0.009	0.257±0.012	0.348±0.012	0.325±0.008	0.336±0.011
kNN	0.279±0.016	0.276±0.012	0.395±0.015	0.366±0.010	0.283±0.016	0.385±0.016	0.356±0.011	0.361±0.016
MLR	0.452±0.033	0.287±0.013	0.556±0.016	0.504±0.009	0.288±0.014	0.518±0.014	0.461±0.008	0.447±0.012
RF	0.255±0.015	**0.250±0.012**	**0.346±0.015**	**0.334±0.010**	**0.244±0.013**	**0.338±0.013**	**0.322±0.010**	**0.334±0.014**
SVM	0.269±0.013	0.255±0.012	0.384±0.013	0.356±0.009	0.260±0.012	0.382±0.014	0.344±0.009	0.371±0.014
				Pond P2				
SGB	0.332±0.011	0.422±0.013	0.381±0.007	0.402±0.007	0.428±0.015	0.375±0.008	0.403±0.007	0.314±0.009
kNN	0.370±0.015	0.416±0.015	0.405±0.008	0.423±0.008	0.424±0.012	0.401±0.009	0.408±0.009	0.396±0.009
MLR	0.421±0.016	0.463±0.012	0.434±0.008	0.506±0.008	0.467±0.012	0.418±0.009	0.506±0.006	0.391±0.010
RF	0.306±0.012	**0.416±0.014**	**0.344±0.009**	**0.381±0.007**	**0.418±0.014**	**0.343±0.008**	**0.385±0.007**	**0.304±0.008**
SVM	**0.288±0.012**	0.424±0.014	0.365±0.008	0.404±0.007	0.431±0.013	0.378±0.011	0.406±0.009	0.340±0.010

The ± separates the average from the standard error of the mean. The smallest RMSE are shown in bold. Machine learning algorithms: stochastic gradient boosting machines (SGB); k-nearest neighbor (kNN); multiple linear regression (MLR); random forest (RF); support vector machines (SVM). Predictor sets: A – °C, DO, pH, turbidity, and SPC; AB – all from A and PC, CHL, and *f*DOM ; ABC – all from AB and NH_4^+, PO_4^{3-}, TN, and TC.

6.3.3 Random Forest as the best performing algorithm

The RF algorithm provided the smallest $\overline{RMSE}$ value in 88% of cases. Only in 2016, the SGB and SVM algorithms, on average, provided lower RMSE values for P1 and P2, respectively. The SGB algorithm provided the second smallest $\overline{RMSE}$ in 75% of cases. Probabilities of $\overline{RMSE}$ being equal for RF and other algorithms are shown in Fig. 1. The probability ranges differed between the ponds and among algorithms. Whereas the probabilities of equal $\overline{RMSE}$ for SGB and RF were high for Pond 1, Pond 2 had a greater spread of probabilities that were generally lower. Similarly, the range of probabilities of equality of $\overline{RMSE}$ for RF and kNN was much wider for Pond 2 compared with Pond 1. Probabilities of equal $\overline{RMSE}$ for RF and SVM were relatively high in both ponds. Ranges of those probabilities were similar in both ponds, unlike with other algorithms. The two significant differences in model performance occurred between the RF and kNN algorithms with the 2018 A ($P < 0.043$) and the 2018 ABC ($P < 0.001$) P2 datasets.

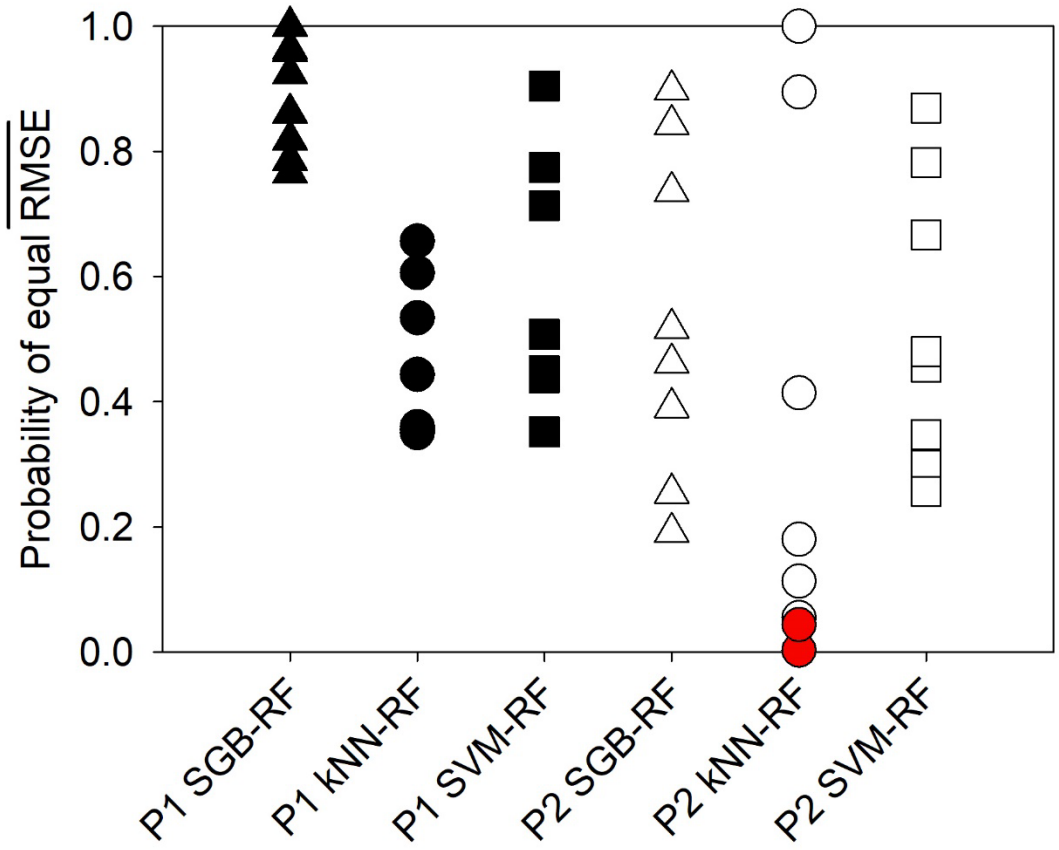


Figure 6.1. Probabilities for the mean RMSE value to be the same in the RF and other ML applications based on the corrected t-statistic. Symbols in red show statistical significance.

6.3.4 Inter-annual differences in algorithm performance

Probabilities of the absence of differences in $\overline{\text{RMSE}}$ for pairs of years varied by year, algorithm, and pond (Supplementary Table S8, Appendix B). Comparisons between 2016 and 2017 performance were not significant in the P1 dataset. For P2, the SVM, RF, and SGB algorithms performed significantly better on the 2017 than the 2016 set. The 2017 predictor A set showed significantly better performance than the 2018 A set in P1 but was not found to significantly differ for any model in P2.

Between the 2016 and 2018 predictor A sets, the SGB, kNN, and SVM algorithms performed significantly better for the 2016 dataset than for the 2018 dataset whereas the performance, while better for RF and MLR in 2016, did not significantly differ. The MLR model was the only model at P2 which showed significantly better performance in 2016 when compared to 2018.

6.3.5 Multiannual algorithm performance

When ML algorithms were applied to the three-year dataset from P1 with the predictor set A, the $\overline{\text{RMSE}}$ appeared to lie between the maximum and minimum $\overline{\text{RMSE}}$ obtained for individual years for the same pond and predictor set. For P2 and predictor set A, the $\overline{\text{RMSE}}$ of the three-year dataset were slightly higher than the largest of the individual year $\overline{\text{RMSE}}$. A similar pattern was observed with the predictor set AB and the two-year dataset. The RMSE for P1 was between the $\overline{\text{RMSE}}$s of individual years, whereas with P2 data, the RMSE of individual years were smaller than the RMSE of the two-year dataset.

6.3.6 Effect of the predictor set expansion

Expanding the predictor set size from A to AB in the 2017 resulted in a slight increase of $\overline{\text{RMSE}}$ in most cases with the one exception being for the RF model in P1. Transition from A to AB and then to ABC predictor sets generally led to the decrease of $\overline{\text{RMSE}}$ (Fig. 2). However, in some cases there was effectively no difference (i.e. RF in P1 between AB and ABC and kNN between all sets in P2). In P1, there was a gradual decrease in RMSE with increased number of predictors whereas for P2 the largest RMSE decreases were between AB and ABC with kNN being the exception.

In all cases the 12-predictor ABC set showed the best performance for all ML algorithms at both ponds.

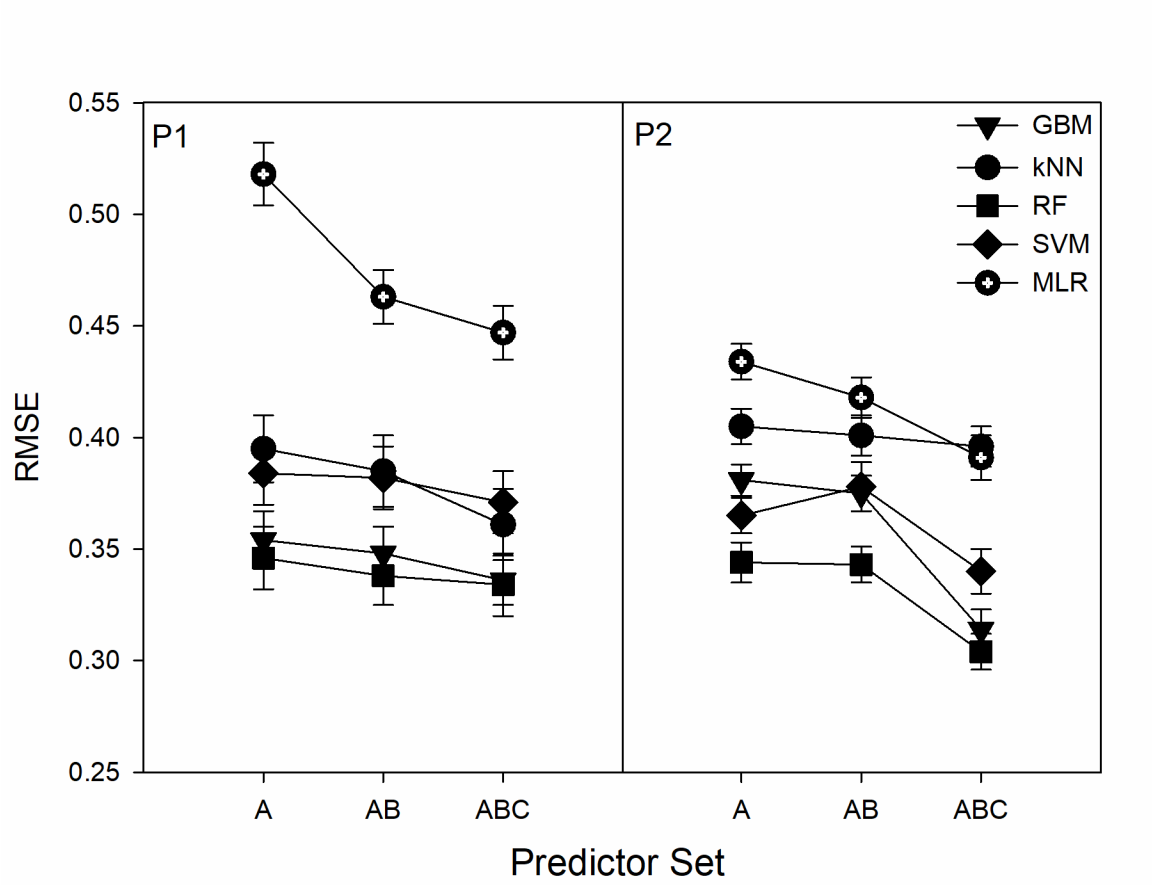


Figure 6.2. Dependence of the root-mean-squared error (RMSE) on the predictor set size. Predictor sets have 5, 8, and 12 predictors for A, AB, and ABC, respectively. The displayed results are from the 2018 datasets.

6.3.7 Other metrics of algorithm performance

The differences between algorithm performance measured with mean absolute error (MAE) and the determination coefficient (R^2) were similar to the differences found with RMSE values (Supplementary Tables S9 and S10, Appendix B). However, in a few instances the model with the best performance changed when analyzing different metrics. For example, the SGB model provided the best

performance in the 2018 ABC predictor set in P1 and the kNN provided the best result for the 2016-2018 predictor set A for P2 when using MAE as a metric. These two were predicted most accurately by the RF model when RMSE is used. Similarly, the SGB model in P1 2017-2018 AB predictor set in P1 provided the highest R^2 value and the kNN model was best in the 2017 A predictor set in P2 whereas both were predicted best by the RF model when using RMSE.

The RF algorithm provided either the best (12 cases) or the second best (4 cases) value of the average MAE. The SGB and SVM algorithms were the closest to the random forest (Supplementary Table S9). Similar to the results comparing RMSE, the 2016 and 2017A and AB sets were predicted with the lowest errors according to average MAE and errors were generally lower with P1 than P2. The only significant differences in ML model performances using the MAE metric was between the RF and kNN in the P2 2018 ABC dataset ($P = 0.004$) (Supplemental Figure S12, Appendix C). Based on R^2 values, the RF model was preferred in 12 of 16 cases with it being second in 2 cases and tied with the SGB model in two cases (Supplementary Table S14). Interestingly, R^2 values for the 2017 A and AB sets were the lowest across models despite this year having the smallest average RMSE and MAE values. Similar to results with RMSE and MAE, kNN was the only ML model to perform significantly worse ($P = 0.003$) than the RF model based on R^2 values which occurred for the 2017-2018 AB dataset (Supplemental Figure S13, Appendix C).

Normalized RMSE and MAE were calculated for all years and predictor sets for both ponds (Supplemental Tables S11 and S12, Appendix B). The preferred algorithm did not change from those presented in Table 6.1 and Supplemental Table

S9 for values of RMSE and MAE, since for a given dataset, RMSE and MAE values from all algorithm were divided by the same value. The percentages of the errors as shown by examination of NRMSE values (Supplemental Table S11) for the ML algorithms were generally around 10 % of the data range for the P1 2016A, 2017A, 2016 – 2018A, and 2017AB datasets. The P1 2018A, AB, and ABC datasets contained higher errors which were between 10 % and 20 %. The MLR algorithm provided consistently higher errors which were typically in the range of 15 to 20 % except for the 2017 datasets which had < 10% error. P2 contained greater relative error (15 – 20 %) than the P1 datasets with the exception being the 2018A, AB, and ABC data in which errors were in the same range between ponds. Values of NMAE were consistently lower than those of NRMSE and were in the range of 5.2 % to (RF P1 2017A) to 15.5 % (MLR P2 2018 A).

6.3.8 Algorithm performance for the combined Pond 1 and Pond 2 datasets

Results of combining the P1 and P2 datasets are shown in Supplemental Table S13. The SVM algorithm performed the best in terms of values of RMSE, R^2, and MAE in the combined 2016 dataset whereas the RF algorithm showed the best performance in the 2017, 2018, and 2016 – 2018 datasets. The algorithm performance for the combined P1 and P2 dataset was never better than the performance of both ponds run individually for certain years but was typically higher than one pond and lower for the other (Table 6.1; Supplemental Tables S9 and S10). Including 'site' as a categorical input generally improved the performance of the combined P1/P2 datasets across years (Supplemental Table S14).

6.3.9 Recursive feature elimination

In almost all scenarios the RF model was shown to provide the best $\overline{RMSE}$ values when compared to the other models. For this reason, a recursive feature elimination in "caret" was performed using the RF model to determine and rank the most important features for each pond and by year. Supplemental Figure S14 (Appendix C) shows the reduction in RMSE values for the RF models for each year. Examination of the graph shows that in most cases the models do not greatly benefit from more than 5 predictors but the specific predictors varied by pond and year. The predictor importance order for each year from the RF-RFE was recorded and each variable received a numeric value corresponding to the overall ranking (Table 6.2). In this way, the variables with the lowest scores corresponded to being picked as consistently most important.

Table 6.2. The top five important variable as determined by the recursive feature selection algorithm in caret with Random Forests. Variable set A uses importance metrics from 2016, 2017, and 2018. Variable set AB uses metrics from 2017 and 2018. Variable set ABC uses metrics only from 2018.

Pond 1						Pond 2					
Variable Set A		Variable Set AB		Variable Set ABC		Variable Set A		Variable Set AB		Variable Set ABC	
Variable	Average Rank	Variable	Average Rank	Variable	Rank	Variable	Average Rank	Variable	Average Rank	Variable	Rank
SPC	1.3	SPC	2.0	*f*DOM	1	C	2.0	SPC	1.5	*f*DOM	1
C	1.7	C	2.5	SPC	2	SPC	2.7	C	2.0	SPC	2
DO	3.7	*F*dom	3.5	CHL	3	pH	3.3	pH	2.5	CHL	3
pH	4.0	CHL	4.5	TN	4	NTU	3.3	NTU	5.0	TN	4
NTU	4.3	NTU	5.5	TC	5	DO	3.7	*f*DOM	5.5	TC	5

Variable set A was measured in 2016, 2017, and 2018. Variable set AB was measured in 2017 and 2018 and variable set ABC was measured in 2018 only. A = C, pH, NTU, SPC, AB = A + CHL, PC, fDOM. ABC = AB + NH_4^+, PO_4^{3-}, TN, and TC.

Within the various predictor set scenarios, there were some parameters that were consistently highly ranked as important features. These included SPC, C, *f*DOM, CHL and NTU. All other parameters showed largely inconsistent ranking. While SPC was consistently ranked in the top 2 variables for all scenarios, temperature was as well for variable sets A and AB but was ranked poorly in the ABC datasets. The PC value was consistently ranked low in the datasets that contained this parameter. Similarly, NH_4^+ and PO_4^{3-}, the only ionic nutrients measured in the study were also ranked low in the ABC set. Conversely, TN and TC were ranked 4th and 5th important in both ponds in the 2018 datasets. Interestingly, the 2018 ABC parameter sets had identical rankings of the 5 top important features and overall similar rankings of the remaining 7 parameters.

6.4. Discussion

While model performance generally did not significantly differ, the RF model was found to provide consistently better performance than any of the other models evaluated. Several publications focusing on ML model evaluation specifically for microbial water quality purposes have reached similar conclusions (Weller et al., 2021; Avila et al., 2018; Chen et al., 2020). The SGB model was found to provide the second-best performance across all three metrics from all datasets. In several empirical ML comparison studies, it has been stated that SGB models usually outperform RF models (Caruana and Niculescu-Mizil, 2005; Hastie et al., 2009; Maclin and Opitz, 1997) but other studies report the opposite (Manchada et al., 2007;

Khoshgoftlaar et al, 2010; Bauer and Kohavi, 1999). Therefore, the choice of "the best" model may be dataset-dependent which to some degree was evident in the results of this work (Table 6.1; Supplementary Tables S9 and S10, Appendix B). Also, the performance of both models is dependent on how they are tuned. SGB are considered harder to tune than RF models, contain greater sensitivity of tuning parameters with regard to the output, and have a greater number of tuning parameters (Freeman et al., 2016). Both RF and SGB algorithms are tree-based but the SGB algorithm incorporates a boosting procedure whereby model fitting is additive and each new tree is fit to the residuals of the previous tree with the goal of minimizing the most egregious errors according to a specific loss function such as MSE. Because SGBs are additive, they are more susceptible to over- or underfitting which the RF model is robust to because the individual trees are independent and are averaged to create the forest. Lastly, when bagging (RF) and boosting (SGB) type algorithms have been compared, bagging has been described to provide better results when datasets are noisy or there are class imbalances (Khoshgoftaar et al., 2010; Maclin and Opitz, 1997).

The performance of the SVM and kNN algorithms was generally poorer than the RF and SGB algorithms (Table 6.1) but in most cases the performance did not significantly differ (Figure 6.1). One possibility is that both SVM and kNN algorithms have been reported to not handle missing values, near-zero variance predictors, or noisy data as well as RF or SGB algorithms (Kuhn and Johnson, 2016). The SVM algorithm typically provided lower RMSE than the kNN algorithm and this may be because SVM is robust to outliers especially when non-linear kernels are

used. The RBF kernel was used in this study because it provided substantially better results than the linear or polynomial kernels (data not shown) which was also reported in the work by Weller et al. (2021) who used ML algorithms for the prediction of *E. coli* in NY streams. Several other water quality studies have also reported better performance of SVM than kNN when the two have been compared (Modaresi et al, 2014; Danades et al., 2016; Babbar and Babbar., 2017; Prakash et al. 2018; Chen et al., 2020). Finally, the kNN has been reported to not perform well with high dimensional or highly scattered datasets which is why centering and scaling is recommended. However, in our work, this pre-processing procedure did not affect results but may explain why SVM is preferred in other water quality datasets.

Different performance metrics in general agreed with each other, but in some cases contradicted. For example, while all algorithms showed the best $\overline{RMSE}$ on the 2017 P1 dataset (Table 6.1), this year was ranked the worst predicted by the R^2 metric (Supplemental Table S10, Appendix B). This is caused by the distribution of observed and predicted data along a 1:1 line. This example highlights the importance on the choice of performance metric reported in algorithm evaluation studies and the advantage of using multiple performance metrics.

We compared only 5 different algorithms in this study which were chosen due their popularity, but many more ML algorithms and their modifications exist and can be tested for regression-type application (Kuhn, 2008; Kuhn and Johnson, 2013; Weller et al., 2021). We chose not to run artificial neural networks (ANN) due to constraints of the dataset dimensions but other researchers have found success in applying ANN algorithms in the field of microbial water quality (Buyrukoglu et al.,

2021; Motamarri and Bocelli, 2012). Other promising algorithms for water quality determinations include those founded in Bayesian statistical methods such as Naïve Bayes or Bayesian Belief Networks (Panidhapu et al., 2020; Avila et al., 2018) and the use of ensemble or model stacking methods (Burukoglu et al., 2021).

Several predictor variables emerged as consistently important for both ponds and across years of observations. These included *f*DOM, SPC, C, and CHL, and NTU. Positive relationships between dissolved organic matter (*f*DOM) and concentrations of planktonic fecal bacteria in water have been reported (Bouteleux et al., 2004; Rincon and Pulgarin, 2004). The relationship is likely governed by the presence of suspended organic substances which may promote the growth and survival of *E. coli* in the present study by providing nutrients, an attachment surface, and decreasing direct cellular photo-inactivation (Katarzyte et al., 2018; Rincon and Pulgarin, 2004; Maraccini et al., 2016; Tamara-Armisen and Servais, 2009).

In both ponds and in every year NTU was positively correlated with *E. coli* concentrations (data not shown). Positive associations of fecal bacteria and NTU have been previously presented (Francey et al. 2013; Partyka et al., 2018; Weller et al., 2020) and can also be related to the level of suspended particulates which have been shown to enhance *E. coli* survival in water (Katarzyte et al., 2018). Additionally, elevated NTU levels may indicate recent disruption of bottom sediments either by bioturbation or runoff-related mixing which results in the resuspension of fecal bacteria contained in sediments (Stocker et al., 2016; Cho et al., 2010).

The presence of CHL in the lists of most important predictors apparently reflects mutualistic relationships between algae and *E. coli* have been reported and

attributed to solar shielding as well algae providing a source of labile organic nutrients which promote bacterial persistence (Vogeleer et al., 2014; Englebert et al., 2008). On average in both ponds, when chlorophyll-a levels were higher, *E. coli* concentrations were lower (Supplemental Table S7, Appendix B). There may exist a threshold level at which there is a mutualistic relationship between *E. coli* and algae and above this level there exists competition (Ansa et al., 2011).

The concentrations of SPC were determined as most important in the largest number of cases in the study. The concentrations of SPC in water are proportional to the ion concentrations present. Ionic nutrient concentrations in water have often shown positive relationships with the concentrations of *E. coli* present (Shelton et al., 2015; Ozkanca, 1993; Lim and Flint, 1989). Recent research has also demonstrated that *E. coli* survival rates in freshwater increase with conductivity by way of reducing osmotic stress and improving membrane stability but can be detrimental above certain levels (DeVilbiss et al., 2021). Runoff reaching waterways may either have a dilutional effect and lower water conductance or increase it (Baker et al., 2019). Rapid changes in SPC may be used as an indicator of when influent such as runoff or precipitation has reached water sources and thus may provide good indication of when *E. coli* concentrations can be expected to change within a waterbody.

Temperature effects on *E. coli* persistence in the environment are perhaps the most well-documented of any other variables in the literature but may also be the most inconsistent. Numerous review and meta-analysis articles indicate *E. coli* persistence is negatively influenced by higher temperatures (Blaustein et al., 2013; Stocker et al., 2014; Cho et al., 2016). However, others have reported positive

relationships (Truchado et al., 2018) while some have reported inconsistent direction of the relationship when multiple sites were included in the same study (Francy et al., 2013; McEgan et al., 2013). These diverse dependencies reveal the complexity of the relationships between *E. coli* and the predictors which govern the aquatic habitat and affect survival. Ultimately, ML algorithms are expected to handle the complex interactions and non-linear relationships better than traditional regression models in aquatic studies (Quetglas et al., 2011; Weller et al., 2021).

Through additional scenario testing (Table 6.1; Supplementary Tables S9 and S10, Appendix B) it was discovered that model performances did not substantially change when the 2017 and 2018 datasets were held at a lower number of predictors. This indicates while parameters introduced in later years of the study were found to be at times more important, the core 5 predictors utilized in 2016A, 2017A, and 2018A predictor sets (C, SPC, NTU, DO, and pH) were found to be largely suitable for predicting *E. coli* concentrations in agricultural pond waters. This finding is of special interest as each additional predictor introduces additional burden on water quality characterization program. Additionally, results of this study indicate that *E. coli* concentrations in irrigation ponds may be "now casted" by using relatively cheap deployable on-line sensor suites that are used for continuous monitoring. It must be acknowledged that this study utilized measurements of a total of twelve predictors. Many additional predictors exist (e.g. ORP, total suspended solids, or various nutrient concentrations such as nitrate or ammonia) which may further improve the predictive performance of the ML algorithms or lead to the creation of similar simple and effective sets of variables as those identified in this study.

We realize that the effect of redundancy of predictors was not fully elucidated in this work. There exist multiple suggestions on redundancy removal as a preprocessing step of regressions using correlations between predictors, variance inflation estimation, or principal component analysis as a basis for the removal of covarying predictors. Multiple methods are suggested in the literature to reduce the effects of the input reduction on variable importance determinations (Bovelstad et al., 2007). These methods tend to increase the reliability of the regression results (more data per coefficient), but at the same time, they may change the perception of the relative importance of input variables (Ransom et al., 2019). Applying these methods to several ML algorithms and assessing results presents an interesting research avenue. In this work, we limited the study by analyzing correlations between inputs. As expected, the only strong correlations were found between DO and pH (data not shown). The most likely mechanism for the observed co-linearity is photosynthetic activity in the ponds which consumes dissolved CO_2 (raising pH) and releases dissolved oxygen (raising DO). We cannot exclude the effect of this correlation on the occurrence and position of DO and pH in lists of important inputs.

The algorithm performance for the combined P1 and P2 dataset was never better than the performance of both ponds run individually for certain years but was typically higher than one pond and lower for the other (Table 6.1; Supplemental Table S13, Appendix B). Explanations for this may be site-specific responses of *E. coli* concentrations to differences in predictor levels in each pond which may in some cases be similar and in others dissimilar. For example, P2 had elevated levels of the photosynthetic pigments phycocyanin (PC) and chlorophyll-a (CHL) compared to P1.

Similarly, DO and pH levels in P2 were typically higher in P2 than in P1 (Supplemental Table S7, Appendix B). It is therefore possible that there were different extents to the effects of these predictor levels on *E. coli* concentrations which were unique for each pond. If monitoring datasets are available for multiple water bodies, one can pool these datasets together and compare the performance between site-specific models and those developed using pooled datasets across locations. Alternatively, one may use 'site' as categorical variable input which may preserve site-specific interactions between *E. coli* and predictors within a larger model. Indeed, in the present study, adding 'site' (e.g. Pond 1 or Pond 2) generally improved the performance of all algorithms on the P1/P2 combined datasets (Supplemental Tables S13 and S14, Appendix B).

The results of this work were gathered by studying only two irrigation ponds both in the state of Maryland and as such the scope of inference is limited. However, in the literature there is a lack of information regarding *E. coli* and water quality dynamics in irrigation sources let alone those involving ML algorithms. The current study suggests a framework for using ML algorithms for irrigation water quality determinations.

6.5. Conclusions

Overall, all ML algorithms performed well in predicting *E. coli* in the datasets. The RF algorithm predicted better in more cases than the other models when assessed in terms of average values of root-mean-squared-error, coefficient of determination, and mean absolute error. However, when those performance metrics

were treated as statistics, there was no significant difference between the ML algorithm performance in most cases. The MLR model consistently provided the worst results which demonstrated the non-linearity of the relationships between *E. coli* and its predictors. The recursive feature elimination exercise revealed similarities in important features across years and sites. Namely, SPC, NTU, C, CHL, and *f*DOM were found to be the most influential variables for the prediction of *E. coli* in the studied ponds. However, it was also shown that the algorithm performances were not substantially improved when predictor sets were expanded to 8 and 12 variables from the core 5 variable list (pH, DO, SPC, C, NTU). The performance of the RF model as well as its relatively simple set up and deployment indicate it may be a valuable tool for water quality managers and researchers to utilize when predicting the microbial quality of irrigation waters.

Chapter 7: Conclusions, Recommendations, and Future Work

Understanding the extent of the spatial and temporal variability of *E. coli* in agricultural irrigation ponds is vital for designing efficient and effective water quality monitoring regimes. Monitoring methods that ignore the spatial and temporal variability of microorganism concentrations in agricultural waters will not lead to an accurate assessment of microbial quality of water extracted for irrigation. As this work has shown, there can be a large degree of variation in *E. coli* concentrations even across relatively small farm ponds and this variation is three dimensional. Temporal variation in concentrations was shown to be substantial at the annual, inter-seasonal, and diurnal temporal scales as well as throughout the course of an irrigation event. Public health risks associated with using surface waters for irrigation cannot be reliably estimated without accounting for this variability.

7.1 Summary of Conclusions

It was shown in Chapter 3 of this work that stable spatial patterns of *E. coli* concentrations and water quality variables can develop in farm irrigation ponds. Sampling locations that were consistently below, about, and above the pond-wide average *E. coli* concentration were identified. These patterns were identified within a single year of observations and were found to persist over multiple years. Concentrations of *E. coli* in irrigation sources which show consistency in the deviations from the spatial average have major implications for microbial water

quality monitoring as microbial quality assessments may differ depending on the selection of sampling location. The Empirical Orthogonal Function (EOF) and Mean Relative Difference (MRD) analyses were used to identify these patterns and were in almost perfect agreement on the ranking of sampling locations by the *E. coli* concentration. The MRD analysis uses relative deviations from the average across the pond and is simpler than the EOF analysis that uses absolute concentrations. The EOF analysis offers more information in the case where there is more than one stable pattern present.

Recommendations stemming from this chapter include the advice to perform pilot surveys to assess the spatial and temporal variability of *E. coli* concentrations in farm ponds. Identification of sampling locations which are representative of the pond-wide average would be advantageous to assess the true risk of using a water source for irrigation. During pilot studies, effort should be made to identify and document the presence of physical features such as point sources, waterfowl congregation areas, or areas where sediments may be disturbed as these may affect the *E. coli* concentrations in the water. Pilot surveys should be performed over a single irrigation season and at least twice a month and should be conducted more than three days after rainfall. Finally, it is recommended to sample both the pond nearshore and interior locations as the observed concentrations have been shown to often significantly differ between these two areas.

Future work associated with Chapter 3 content includes monitoring of a greater number of farm ponds in the Mid-Atlantic and other regions of the USA to better understand both the degree of spatiotemporal variability of *E. coli* that may be

encountered and the environmental parameters which govern this variation. The effectiveness of composite sampling should also be tested as a method to approximate the average *E. coli* concentration across a farm pond while limiting the number of samples to process after sampling.

Chapter 4 of this dissertation focused on documenting the vertical distributions of *E. coli* in farm ponds and the effects of vertical concentrations gradients on the water which reaches the fields. *E. coli* concentrations were found to significantly vary with depth. Concentration hot spots were found at certain depths within some sampling locations but were not observed in the rest of the water column. The effect of vertical and horizontal gradients in *E. coli* concentrations on the water withdrawn for irrigation was demonstrated by using the Environmental Fluid Dynamic Code (EFDC) hydrodynamic modeling software. Spatial heterogeneity of concentrations across the pond translated to temporal variability at the intake. Wind speed and direction were shown to be substantial factors contributing to the movement of *E. coli* in the pond and to the intake. Finally, in this chapter it was also shown that the placement of the irrigation pipeline can be a factor which influences the concentration of *E. coli* in the irrigation water.

Recommendations from this chapter include advisability to sample at multiple depths within an irrigation pond. If this is not feasible, then sampling the irrigation water at the pump or at fields should be done several times throughout an irrigation event to capture the possible effects of the concentration heterogeneity in the pond. The optimal number of samples to take at the pump which best represents the spatial variability of *E. coli* in the pond, the frequency of this sampling, and the use of

composite samples also presents several interesting research questions. While we recommend the use of the EFDC model to gain insight into the factors which effect the dynamics of the *E. coli* concentrations in water which reaches the field, more work is needed to validate these models and assess prediction errors by comparing the results of simulations with periodic grab samples. Finally, the EFDC model can incorporate numerous water quality parameters such as dissolved oxygen and pH and thus it would be interesting to increase the number of input parameters and assess the effects on the *E. coli* concentrations.

In Chapter 5, diurnal variation of *E. coli* concentrations in three irrigation ponds was assessed over multiple sampling dates. In almost all cases, *E. coli* concentrations were the highest in the morning hours and declined throughout the day. On average for all ponds, concentrations decreased by a factor of 1.15 every hour between 9:00 and 12:00 and by factor of 1.11 every hour between 12:00 and 15:00. Physicochemical changes in the pond water, particularly with dissolved oxygen and pH levels, were shown to govern the extent of the decrease in concentrations. Significant diurnal variation was not observed at one of the study sites which had a constant influx of *E. coli* from an inflow along with small changes in dissolved oxygen and pH throughout the day. This indicates that diurnal variation in concentrations cannot be assumed to occur especially if there are opposing forces which neutralize the effects of changes in water properties or solar insolation. Regression trees were shown to be a valuable tool for assessing the environmental covariates which best split the *E. coli* datasets. Recommendations gained from the results of Chapter 5 are that sample collections must be conducted in the early

morning hours when concentrations are the highest. Such timing would provide the most conservative estimate of the actual risk of using an irrigation. There should be consistency in the timing of sampling so that results collected across a season are more comparable and are not confounded by diurnal effects. If multiple sites are being sampled within an irrigation pond, then samples must be collected in a way that minimizes the collection time.

Future work associated with Chapter 5 include performing research towards quantifying and explaining daily cycles and overnight restoration of *E. coli* populations in agricultural water sources. It would be advantageous to incorporate sampling at multiple water depths throughout a day to assess if similar die-off rates can be anticipated throughout the water column. Optimal irrigation regimes should be found that could take advantage of lower concentrations of microorganisms towards the afternoon or evening hours on one hand and the undesirability of irrigation during the maximum heat hours on other hand. It would be interesting to determine if there is a window in the evening when temperature and solar radiation have lowered and fecal microorganism concentrations have not yet begun to increase.

In the Chapter 6 of this dissertation, machine learning (ML) algorithms were applied to multi-year datasets to relate *E. coli* concentrations to the levels of easily measured physicochemical variables. Predictive models were developed using four ML algorithms. Of the 4 ML algorithms evaluated, the random forest algorithm provided the lowest error in almost all cases. It was also found that a simple suite of inexpensive sensors was sufficient for making predictions of *E. coli* concentrations using ML algorithms. The most influential variables varied by year and pond. These

were specific conductance, and concentrations of fluorescent dissolved organic matter, chlorophyll-a, total nitrogen, and total carbon.

Recommendations from Chapter 6 include supplementing microbial water quality monitoring with machine learning modeling. Concurrent measurement of water quality parameters and *E. coli* is highly advisable as it will create datasets and models for the express estimation of *E. coli* from in-situ sensing data. Collecting a large number of samples paired with in-situ data is advisable as the performance of machine learning models is directly related to the number of samples utilized in developing the models. While the random forest algorithm consistently provided the best results, algorithm comparisons should always be conducted as each dataset may contain unique associations between the predictors.

Future work associated with Chapter 6 include evaluating additional predictors which may be used to estimate *E. coli* concentrations such as various hydrological (e.g. horizontal or vertical mixing of water), meteorological (e.g. precipitation amounts or days since rainfall, wind speed, solar irradiance etc.) or physicochemical variables which were not measured in this study (e.g. ammonium, oxidation-reduction potential, salinity etc.). Multi-and hyperspectral imaging of ponds from aerial platforms or unmanned systems should also be considered as a source of predictors which would allow for the rapid collection of a large amount of data over survey areas. Determination of the site-specific usefulness of certain predictors would also be beneficial to able to prescribe an input set of variables which applies to a wide range of cases. Additional ML algorithms or variations on the ones utilized in this chapter need evaluation and may also provide satisfactory results for predicting

microbial water quality. Finally, gaining the ability to "now-cast" *E. coli* concentrations in ponds using sensor data would be highly valuable and would greatly reduce sampling burden and determination times.

7.2 Concluding Remarks

Overall, irrigation ponds appear to be highly complex ecosystems. *E. coli* population dynamics in these ponds are affected by multiple factors and the variability of these factors translates into spatial and temporal variability in the concentrations of *E. coli*. Multiple aspects of pond functioning, such as the role of sediments as a habitat for *E. coli* and the disturbance of these sediments during irrigation events, the effects of microplankton and aquatic vegetation, and competition with other microorganisms are promising avenues for future research which are waiting to be studied. The optimal water quality parameters to incorporate into studies, their site-specific importance, and the possibility to refine the number of parameters to reduce time and costs is also an important research topic.

In-situ sensing allows for the rapid prediction of *E. coli* concentrations when paired with machine learning modeling and proximal sensing using unmanned aerial vehicles (UAVs) equipped with multi-or hyperspectral cameras may further improve our ability to make rapid and accurate predictions (Kim et al., 2020; Morgan et al., 2021). Aquatic drones have the potential to expedite sampling, allow for access to hard-to-sample locations such as pond interiors or shallow waters, and sample at multiple depths without the use of a boat or specialized equipment.

Metagenomic surveys of agricultural waters through gene sequencing also present a promising opportunity for performing assessments on the presence or absence of specific virulence genes associated with foodborne pathogens (Tan et al., 2015; Chopyk et al., 2019). Recent advances in gene sequencing are providing opportunities to rapidly assess pathogen presence in irrigation waters in-field using rapid sequencing kits (Maguire et al., 2021). While sequencing alone does not provide a quantitative estimate of FIB or pathogen concentrations, it can be paired with quantitative PCR which can distinguish between culturable and non-culturable microorganisms (Khan e al., 2007).

The importance of agricultural ponds in providing clean and safe water for irrigation is high and is expected to increase. There is a need and feasibility to develop science-based monitoring designs, forecasting strategies, and models to reduce the incidences of using water of poor microbial quality for irrigation. While this dissertation has helped to elucidate several important factors and considerations regarding the microbial testing of irrigation ponds, an equal number, or greater, of questions have arisen stemming from the experimental chapters. Future work opportunities within this field are ripe as is the possibility to dramatically improve our ability to detect when unsafe water quality conditions are present. By combining improved sampling regimes, modeling, and emerging technologies we can minimize the risks of using surface waters for irrigation and protect the lives and livelihoods of countless people.

Appendices

Appendix A: Depth-dependent concentrations of *E. coli* in agricultural irrigation ponds

Introduction

Escherichia coli (*E. coli*) is a fecal indicator bacteria (FIB) which is measured to assess the microbial quality of waters used for irrigation. In 2016, the Food and Drug Administration's Produce Safety Rule (PSR) required that growers periodically sample their irrigation waters during the growing season to ensure adequate microbial water quality (FDA, 2016). However, FSMA gave growers little information on where or when to sample.

FIB and pathogens in environmental waters exhibit large spatial variability both laterally across a waterbody and with depth (USEPA, 2010). However, water samples to characterize the variability of microbial pathogens and FIB in agricultural ponds are usually collected only from a single depth close to the surface (0 to 30 cm). (McEgan et al., 2013; Topalcengiz et al., 2017; Solaiman et al., 2020). Only few studies incorporate samples taken from a greater depth. In particular, Luo et al. (2015) and Lee et al., (2018) sampled multiple irrigation ponds in Georgia but only collected 1 or 2 samples from the surface or a depth of 50 cm. Similarly, Antaki et al. (2016) sampled two ponds in Georgia but included a 100-cm depth in addition to a surface sample.

Irrigation ponds are typically shallow water bodies but this is dependent on the region in which they are located and the volume of water that is needed for

irrigation. The Natural Resource Conservation Service (NRCS) recommends that manmade irrigation ponds should range between 5 to 14ft (~1.5 to 4.3 m) with greater depths recommended for Western states with dryer climates and less precipitation (NRCS, 1997). Nevertheless, nonuniformity in fecal microorganism concentrations or the levels of water quality parameters which describe their habitat may be expected at different depths within farm ponds.

During irrigation events, the largest volume of water is withdrawn from the pond interior and at depths greater than the surface (~ 0 – 20 cm). For this reason, concentrations of microorganisms measured in samples taken near the surface may not be representative of the greater volume of water used for irrigation. There is ample literature available regarding the concentrations of pathogens or fecal indicator organisms in wastewater or maturation ponds which are engineered to reduce fecal microbe populations via biophysical processes, lengthened holding times, and/or chemical treatments (this is well reviewed in the work of Dias et al., 2017). However, there are very few reports which focus on ponds which directly provide irrigation water to crop fields. Substantial increases in indicator concentration with depth in non-tidal lakes or reservoirs can be expected due to the development of well-defined gradients in temperature, light, or dissolved oxygen (Davis et al., 2005; Partyka et al., 2016) but whether this can be expected in relatively shallow irrigation ponds is not known as sampling is often most often performed only at the water surface.

The objective of this work was to examine vertical distributions in *E. coli* concentrations at different times of the day in three irrigation ponds.

Materials and Methods

Water sampling was conducted at three agricultural irrigation ponds in Maryland between 2019 and 2021 (Figure 1). Detailed descriptions of P1 and P2 can be found in Stocker et al. (2021). Briefly, P1 is an embankment pond which is about 4000 m^2 and is located in central Maryland while P2 is a rectangular excavated pond located on the Eastern Shore of Maryland and is also about 4000 m^2. P1 and P2 have a similar average depth at roughly 270 cm. Both ponds also have ephemeral inflows which are marked in Figure 1.

Pond 3 is a retired irrigation pond located at the USDA's Beltsville Agricultural Research Center (BARC) in Beltsville, Maryland, USA. This man-made pond is 4,383 m^2 and used to provide irrigation water to the adjacent fields which are primarily corn fields. The Western perimeter of the pond is boarded by a 10-m grass buffer strip whereas the remainder is adjacent to a forested area. To the West and Southwest of the pond are separate fields planted with corn and orchard grass. The surrounding fields receive no nutrient amendments. The pond depth ranges from 30 cm to 130 cm with an average depth of 90 cm.

For P1 and P2, samples were taken from the interior at 4 depths: 0-cm, 50-cm, 100-cm, and 150-cm. The exception to this was sampling at P2 on two dates when the 150-cm depth was not sampled. Samples were not collected from P3 at a depth greater than 100-cm due to concern of disrupting the pond sediments. Samples were collected within 30 min around 9:00, 12:00, and 15:00 to assess potential diurnal effects on *E. coli* concentrations.

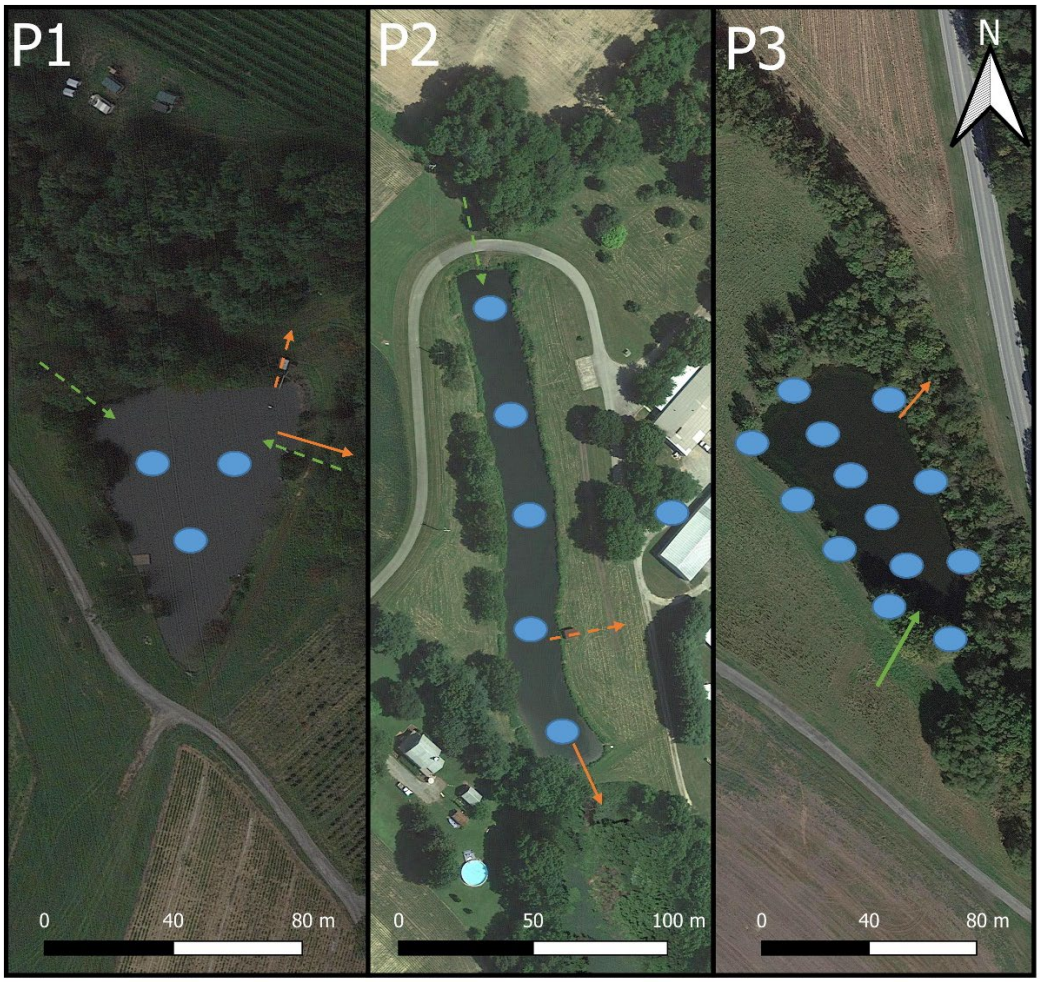


Figure 1. Sampling map of the three ponds in the study. Sampling sites are marked with blue dots. Ephemeral and constant inflows are marked with dashed and solid green arrows, respectively. Ephemeral and constant outflows are marked with dashed and solid orange arrows, respectively.

All water samples were collected from a boat with a Sigma 900 MAX all weather autosampler connected to an external battery (Hach, Loveland Colorado). This programmable peristaltic pump was fed a vinyl tube with a strainer at the end which was lowered to depths specified by pre-measured markings on the vinyl tube. The strainer was attached to the YSI EXO2 sensor guard such that the strainer openings were at the same level as the sensor tips. The vinyl tube was purged for 20

seconds with the water at the depth of interest prior to sample. The autosampler pumped water into pre-labeled 500-mL plastic sampling containers which were kept on ice until the boat returned to shore and samples were immediately removed from the cooler and placed in the dark in an ice-filled cooler. Sample processing in the lab occurred usually within 2-hr of the 15:00 sample collection time. At the beginning and end of sample collections the vinyl tubing was rinsed with tap water, then 70% ethanol which was allowed to sit for several minutes, followed by a rinse with deionized water.

E. coli enumeration was based on EPA method 1603 (U.S. EPA, 2014). Briefly, 50 to 150 mL of pond water was filtered, depending on turbidity, through a 0.45 μm filter (EMD Millipore) which was then aseptically transferred onto Petri dishes with modified mTEC media. Agar plates with filters were then incubated at 22 °C for 2 hours followed by incubation for 22 hours at 45 °C. After incubation, purple colonies were counted as *E. coli*. All samples were filtered in duplicate and counts were then averaged.

Data Analysis

The two-way analysis of variance (ANOVA) was performed to analyze the differences in mean concentrations of *E. coli* by time of day and depth layer and to assess the interactions between the factors. All statistical analysis was performed in the R programming language (R, Core, 2015) and statistical significance was assessed at the $\alpha = 0.05$ level. Figures were created with Sigmaplot v13 (Systat Software, California) and maps were made with QGIS v3.18.2

Results and Discussion

Figures 2 – 4 show the *E. coli* concentration at each of the three times of day and at sampling depths for Ponds P1, P2, and P3, respectively. The concentrations of *E. coli* in all three ponds were low relative to what has been observed in previous studies at these locations (Pachepsky et al., 2018; Stocker et al., 2019; Stocker et al., 2021) and were generally lower than 20 CFU 100 mL^{-1} on most sampling dates. The highest concentrations were recorded in P3 on the 8/12/2020 date which were several times higher than all other observations made at the other ponds throughout the study. Overall the concentrations measured in the surface layer of water at all ponds in this study fell within the ranges of values reported by other researchers of irrigation ponds in the Mid-Atlantic (Truitt et al., 2018; Solaiman et al., 2021) and Southeastern U.S. (Luo et al., 2015; Antaki et al., 2016). Several other researchers have reported much higher average concentrations of *E. coli* measured in irrigation ponds, however, in these studies the effects of recent precipitation were used as an explanation for this observation (Steele et al., 2006; Pahl et al., 2013; Topalcengiz et al., 2017). Explanations for very low concentrations reported in this study may be due to the intentional separation of sampling from recent precipitation events as we were trying to characterize the conditions in which irrigation was warranted. An additional factor may be that there were no domesticated animals or animal manures located on the adjacent lands to serve as direct fecal inputs.

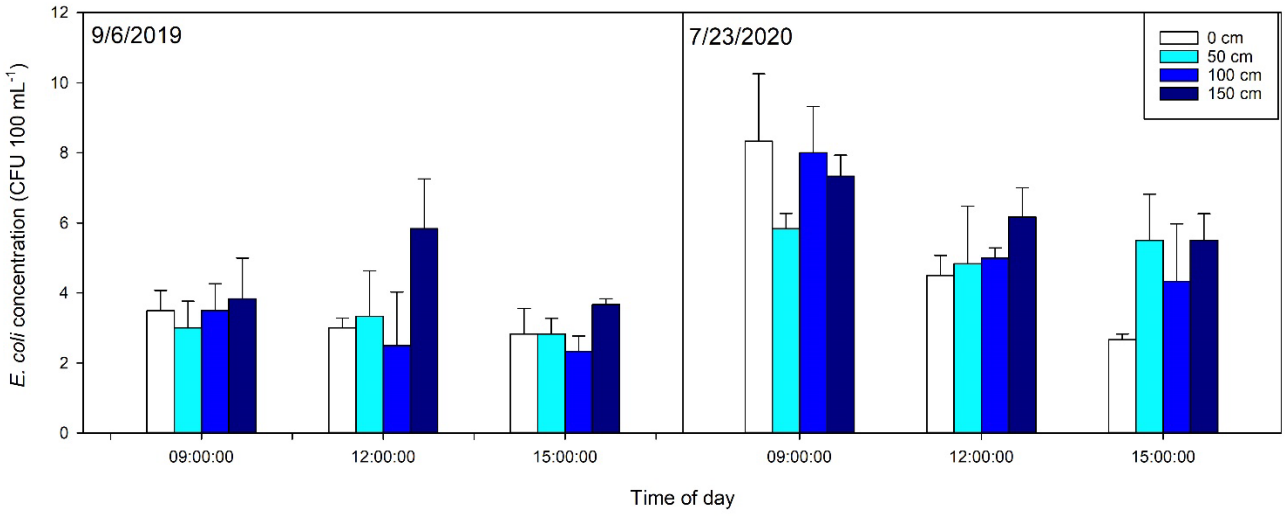


Figure 2. *E. coli* concentrations measured at Pond 1 on two observation dates. Error bars show standard errors.

E. coli concentrations at different sampling depths in P1 were very similar (Figure 2). ANOVA testing confirmed that depth was not a significant factor for either sampling date (Table 1). The sampling time was found to be a significant factor on 7/23/2020, and a steady decrease of concentrations was observed from 9:00 to 15:00. Significant diurnal reductions in fecal microorganism concentrations have been reported in the Great Lakes (Whitman et al., 2004: Ge et al., 2012), in streams (Traister and Anisfeld, 2006; Lenart-Boron et al., 2016; Stocker et al. 2016), mangroves (Becker et al., 2020), coastal waters (Boehm et al. 2002; Boehm and Weisberg, 2005), and irrigation canals (Lothrop et al., 2018). Common factors explaining observed diurnal trends are increase of solar radiation and temperature from morning to afternoon which promote the inactivation of fecal microorganisms in water (Maracinni et al., 2016; Lothrop et al., 2018).

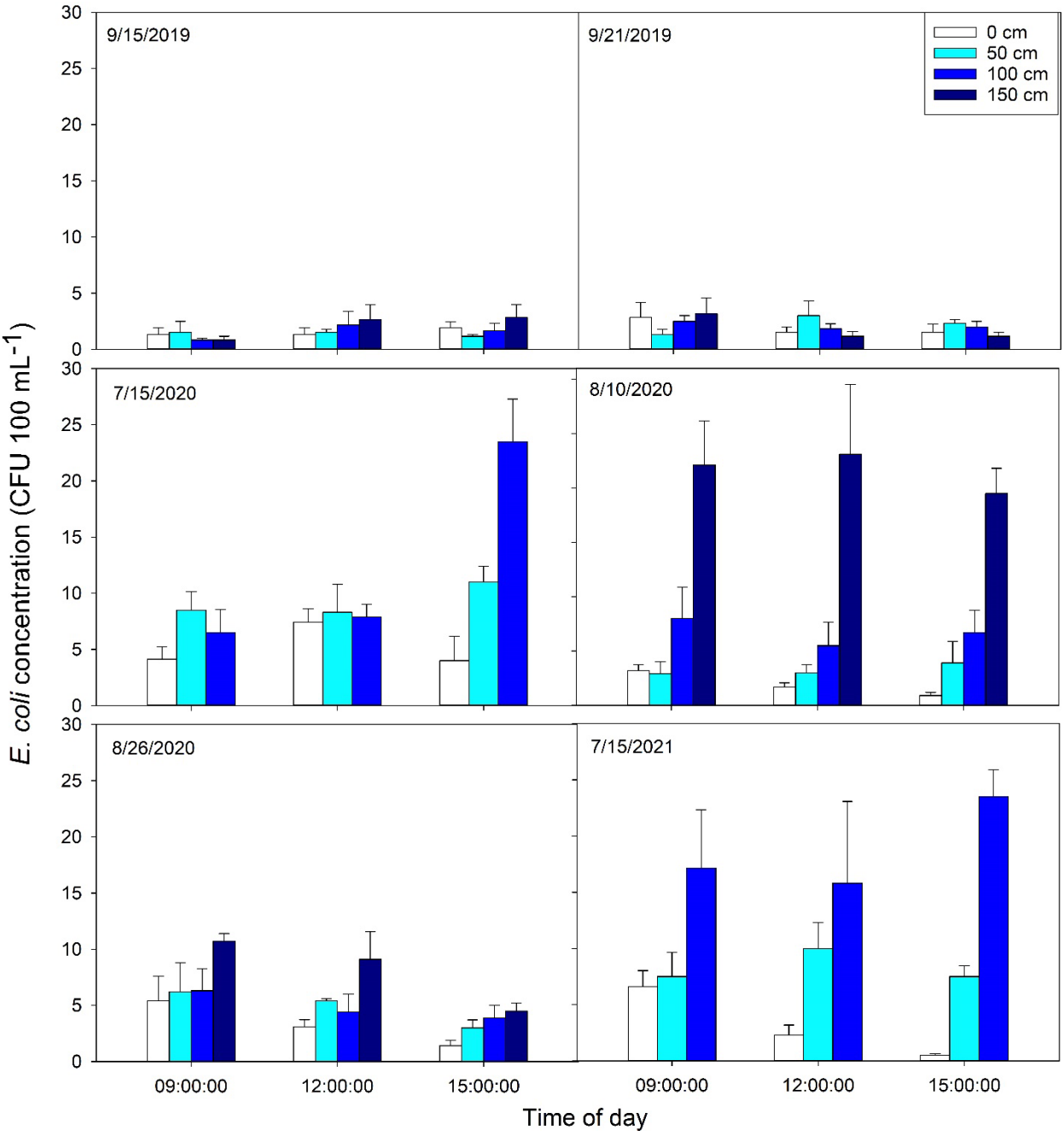


Figure 3. Concentrations of *E. coli* measured at Pond 2 during the observation dates. Error bars show the standard error of the mean.

At P2 during the 2019 sampling dates the concentrations of *E. coli* were near the detection limit (1 CFU 100 mL^{-1}) (Figure 3). At this study site, only one date

showed significant temporal trend (8/26/2020) of decrease in concentrations at each sampling time from 9:00 to 15:00. Depth was not a significant factor at P2 during the first three sampling dates although on 7/15/2020 during the 15:00 sampling time the concentration at the 100-depth was twice as high as the 50-cm depth and about 6 times as high as the surface water. On the 8/10/2020, 8/26/2020, and 7/15/2021 sampling dates at P2, depth was found to be a significant factor with *E. coli* concentrations increasing with sampling depth.

Information on differences in *E. coli* concentrations by depth in agricultural irrigation ponds is extremely scarce. In only two studies have authors taken samples from multiple depths and these studies were conducted in the same region of the United States. Antaki et al., (2016) sampled two irrigation ponds in Southern Georgia and found conflicting results. In one pond the levels of *E. coli* and *Salmonella* were roughly 31 and 4 times greater, respectively, at a 1 m depth compared to the surface water. At the second pond, *E. coli* and *Salmonella* were 0.5 and 6 times lower, respectively, at the surface depth than at the 1 meter depth. In the second study, Lee et al. (2018) sampled six irrigation ponds in Georgia and found nearly identical concentrations and prevalence of *Salmonella spp*. between the surface and 50-cm water sampling depth.

E. coli survival in water has been shown to be greater at depth than near the surface. For example, Jenkins et al. (2011) studied *E. coli* O157:H7, commensal *E. coli*, and intestinal enterococci persistence in in-situ microcosms placed in an irrigation pond. In their work, the time for a 10-fold decrease in the concentrations of all the studied microorganisms was significantly greater at the 50-cm depth than the

5-cm depth which the authors attributed primarily to higher solar insolation at the surface water depth. Similarly, Maiga et al. (2009) studied *E. coli* and enterococci survival in microcosms built to simulate wastewater ponds and found no difference of bacterial persistence at the 10-cm or 90-cm depths in dark conditions, however when the microcosms were exposed to sunlight, the inactivation at the 10-cm depth was significantly greater than at 90-cm. These findings agree with Mayo et al. (1989) who showed that the 10-fold decrease time in coliform concentrations in maturation pond water took 21, 90, and 150 hours at depths of 0-, 15-, and 100-cm, respectively.

Concentrations of *E. coli* at larger depths may be related to a closer proximity to pond sediments or sediment disruption. Sediments have been shown to be a reservoir of fecal microorganisms in aquatic systems (Jamieson et al., 2005; Cho et al., 2010). Additionally, hyporheic exchange of fecal microorganisms has been demonstrated in streams (Grant et al., 2011; Pachepsky et al., 2017) but so far there are no reports quantifying this flux in pond waters. Determining the influence of *E. coli* cells originating from the sediment on the microbial quality of pond waters would be an interesting avenue for research.

Table 1. Summary of two-way Analysis of Variance (ANOVA) testing with sampling depth and time of day as factors.

Site	Date	Time	Depth	Interaction
Pond 1	9/6/2019	0.321	0.368	0.861
Pond 1	7/23/2020	**0.000**	0.183	0.859
Pond 2	9/15/2019	0.087	0.274	0.128
Pond 2	9/21/2019	0.433	0.822	0.374
Pond 2	7/15/2020	0.494	0.336	0.987
Pond 2	8/10/2020	0.579	**0.000**	0.748
Pond 2	8/26/2020	**0.002**	**0.028**	0.788
Pond 2	7/15/2021	0.529	**0.001**	0.393
Pond 3	8/12/2020	0.632	0.621	0.271
Pond 3	7/7/2021	0.270	0.554	0.840
Pond 3	8/13/2021	0.182	0.098	0.733

Values in **bold** indicate statistical significance at the (P = 0.05) level.

The P3 pond did not show significant differences by depth or time of day on any of the three sampling dates (Table 1). However, in most cases concentrations were higher at the surface than at depth (Figure 4). He et al. (2015) reported uniformly distributed *E. coli* and fecal coliform concentrations in pond water at depths up to 1.2 m and found higher concentrations at depths exceeding this value. The authors attributed this to thermal gradients which may inhibit vertical movement of water and therefore establish the stratification of microorganism concentrations. It is possible that the shallow nature of this pond prevented gradient formation which may have been the case as well in the works by Antaki et al. (2016) and Lee et al. (2016). Is also possible, though, that these authors did not sample deep enough for samples to be effected by vertical gradients in parameters such as temperature, light penetration, or dissolved oxygen.

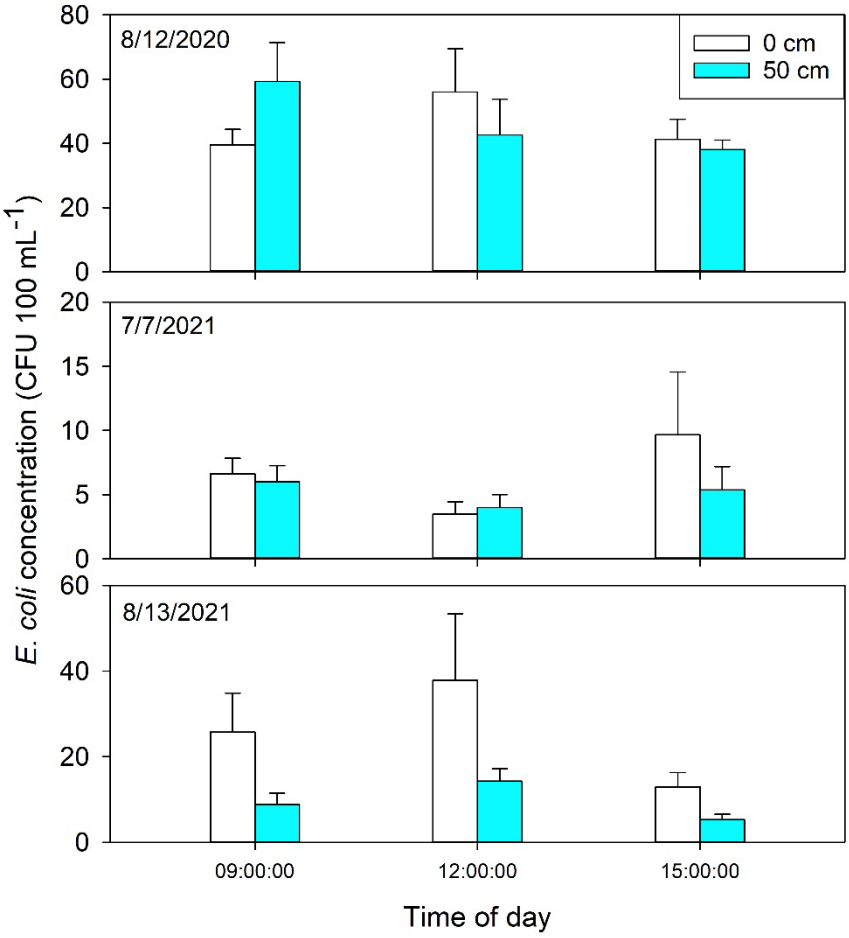


Figure 4. Concentrations of *E. coli* by sampling depth in Pond P3.

The results from studies focusing on wastewater, stabilization, or maturation ponds may apply to irrigation ponds whereby the factors that govern fecal microorganism concentrations are primarily pH, dissolved oxygen, temperature, and UV radiation (Ouali et al., 2013; Liu et al., 2020; Park et al., 2021). More work is needed to relate changes in water quality parameters controlling the *E. coli* habitat in irrigation ponds. Additionally, it would be worthwhile to identify and sample irrigation ponds which contain higher *E. coli* concentrations than those utilized in this

study to evaluate if concentration gradients proportional to those observed in this study can be expected. The *E. coli* concentrations encountered in this study were below the FSMA PSR thresholds and therefore it would also be interesting to see if higher concentrations in ponds by depth could affect regulatory compliance depending on where in the water column the sample is taken.

Vertical gradients of *E. coli* concentrations in ponds have been shown to affect the temporal variation in water withdrawn for irrigation (Stocker et al., 2020). Gaining more information on depth-specific concentration gradients will improve our ability to forecast the microbial quality of water withdrawn for irrigation as well as perform effective monitoring exercises which best characterize the true risk of using an irrigation source. Ultimately, the authors agree with the conclusions of Enns et al. (2012) who concluded sampling at different times of the day and at different depths can significantly impact microbial monitoring results and may lead to differing management decisions.

Conclusions

This study examined *E. coli* concentrations in irrigation ponds at more than two depths and throughout multiple time points of a day. Overall, low concentrations of *E. coli* were observed in each of the ponds and absolute differences in concentrations between different times of day or by depth were not substantial. However, statistical testing revealed significant differences in both the sampling time of day and by depth at various observation dates. Namely, in P2 three of the six sampling dates showed significant differences in *E. coli* concentrations by depth

while 1 of six showed time of day as a significant factor. In several observation dates at P2 the concentrations in the lowest layer were found to be several times higher than what was measured at the surface. In P1 and P3, significant differences in concentrations by depth were not observed but in one of the two P1 observation dates, time of day was found to be a significant factor. Results of this study demonstrate the possibility for the formation of vertical *E. coli* concentration gradients in irrigation ponds and suggest that monitoring campaigns should be expanded to sample at greater water depths which represent a larger volume of the pond.

Appendix B: Supplemental Tables

Supplemental Table S1. Observed levels of water quality parameters in P1 across the monitoring period. The average is separated from the standard error of the mean by the ± symbol.

2016						
Date	5/31	6/13	6/27	7/11	7/25	-
LEC	0.11 ± 0.10	0.59 ± 0.07	1.36 ± 0.07	1.00 ± 0.07	1.10 ± 0.06	-
C	26.56 ± 0.26	24.85 ± 0.10	25.94 ± 0.06	27.15 ± 0.28	26.48 ± 0.12	-
pH	8.56 ± 0.06	8.86 ± 0.05	8.49 ± 0.09	8.24 ± 0.12	9.10 ± 0.13	-
DO	8.42 ± 0.22	9.17 ± 0.34	9.04 ± 0.19	13.10 ± 0.34	9.94 ± 0.45	-
SPC	154.47 ± 3.73	137.20 ± 0.26	129.58 ± 2.23	143.58 ± 0.49	173.74 ± 1.18	-
NTU	NC	5.94 ± 0.18	6.58 ± 0.46	8.32 ± 0.43	NC	-

2017						
Date	6/7	6/21	7/5	7/18	8/1	8/16
LEC	1.61 ± 0.05	1.77 ± 0.06	1.52 ± 0.09	1.50 ± 0.06	1.83 ± 0.02	1.78 ± 0.05
C	22.73 ± 0.03	27.14 ± 0.08	28.00 ± 0.06	29.14 ± 0.10	25.44 ± 0.08	25.43 ± 0.09
pH	8.51 ± 0.06	8.70 ± 0.08	9.39 ± 0.02	8.86 ± 0.07	8.64 ± 0.05	8.79 ± 0.09
DO	10.52 ± 0.19	9.12 ± 0.22	11.92 ± 0.07	11.02 ± 0.23	10.16 ± 0.12	9.58 ± 0.08
SPC	156.40 ± 0.14	168.85 ± 0.19	170.13 ± 0.12	175.82 ± 0.27	150.74 ± 0.09	153.10 ± 0.11
NTU	2.88 ± 0.23	4.82 ± 1.02	4.06 ± 0.30	7.38 ± 0.77	6.28 ± 0.10	5.54 ± 0.31
CHL	1.88 ± 0.13	2.34 ± 0.78	6.25 ± 0.66	7.56 ± 1.90	1.68 ± 0.08	3.29 ± 0.44
PC	0.22 ± 0.02	0.32 ± 0.07	0.88 ± 0.12	2.34 ± 0.33	1.17 ± 0.06	1.94 ± 0.11
FDOM	3.42 ± 0.12	2.11 ± 0.30	4.82 ± 0.53	6.24 ± 0.24	36.65 ± 0.33	25.02 ± 0.32

2018						
Date	6/20	7/5	7/19	8/15	8/29	10/4
LEC	1.79 ± 0.07	1.04 ± 0.11	0.49 ± 0.13	1.46 ± 0.04	1.11 ± 0.06	0.67 ± 0.10
C	26.59 ± 0.06	29.57 ± 0.06	28.25 ± 0.14	26.83 ± 0.16	27.73 ± 0.21	22.97 ± 0.14
pH	8.46 ± 0.04	7.30 ± 0.02	7.25 ± 0.04	7.46 ± 0.03	8.06 ± 0.08	7.36 ± 0.03
DO	11.59 ± 0.16	7.89 ± 0.07	8.96 ± 0.06	10.50 ± 0.07	11.23 ± 0.11	10.23 ± 0.10
SPC	141.37 ± 0.09	156.44 ± 0.07	166.90 ± 0.15	146.49 ± 0.13	132.82 ± 9.10	148.95 ± 0.09
NTU	8.80 ± 1.00	3.76 ± 0.45	2.57 ± 0.47	2.47 ± 0.26	3.97 ± 0.39	3.60 ± 0.15
CHL	2.42 ± 0.14	1.12 ± 0.10	1.03 ± 0.17	4.55 ± 0.29	6.93 ± 0.48	0.73 ± 0.03
PC	2.51 ± 0.50	0.60 ± 0.12	0.28 ± 0.05	0.64 ± 0.10	0.91 ± 0.09	0.76 ± 0.06
FDOM	21.35 ± 0.59	7.11 ± 0.33	1.92 ± 0.20	29.94 ± 0.36	25.58 ± 0.50	38.39 ± 0.38

Parameters are abbreviated as C (temperature; °C), DO (dissolved oxygen; mg L^{-1}), SPC (specific conductivity; µS cm^{-1}), NTU (turbidity; NTU), PC (phycocyanin; RFU), CHL (chlorophyll-a; RFU), and FDOM (fluorescent dissolved organic matter; ug L^{-1}).

Supplemental Table S2. Observed levels of water quality parameters in P2 across the monitoring period. The average is separated from the standard error of the mean by the ± symbol.

2016						
Date	6/8	7/6	6/22*	7/20	8/4	8/10
LEC	1.43 ± 0.05	1.01 ± 0.09	3.07 ±1.76	0.73 ± 0.12	1.13 ± 0.04	0.80 ± 0.12
C	21.54 ± 0.34	27.44 ± 0.16	24.2 ± 0.17	28.94 ± 0.29	28.08 ± 0.07	28.12 ± 0.26
pH	8.83 ± 0.13	6.54 ± 0.11	6.71 ± 0.11	8.19 ± 0.21	7.25 ± 0.19	7.36 ± 0.18
DO	13.89 ± 0.52	7.04 ± 0.57	5.63 ± 0.59	10.16 ± 0.70	6.13 ± 0.22	10.54 ± 0.68
SPC	133.86 ± 5.03	164.91 ± 1.12	172.33 ± 1.51	151.29 ± 0.58	155.26 ± 1.54	163.96 ± 0.96
NTU	27.49 ± 3.29	NC	28.68 ± 3.80	14.51 ± 3.44	65.97 ± 22.89	NC

2017						
Date	5/31	6/13	6/27	7/11	7/25	8/8*
LEC	0.39 ± 0.10	0.57 ± 0.10	1.22 ± 0.06	0.75 ± 0.05	1.08 ± 0.05	3.06 ± 0.01
C	24.24 ± 0.21	30.69 ± 0.21	27.98 ± 0.11	29.58 ± 0.12	29.58 ± 0.06	22.14 ± 0.06
pH	9.29 ± 0.06	8.73 ± 0.08	8.05 ± 0.12	8.14 ± 0.11	8.08 ± 0.11	6.48 ± 0.03
DO	16.18 ± 0.62	13.79 ± 0.54	13.19 ± 0.68	11.57 ± 0.49	10.10 ± 0.57	5.32 ± 0.06
SPC	151.15 ± 0.33	160.15 ± 0.28	169.30 ± 0.24	174.99 ± 0.21	179.16 ± 0.32	141.77 ± 0.46
NTU	4.73 ± 0.53	9.58 ± 2.45	6.88 ± 1.21	4.16 ± 0.38	31.84 ± 2.03	5.11 ± 0.19
CHL	10.12 ± 1.47	17.09 ± 5.30	20.89 ± 5.03	13.25 ± 2.55	15.15 ± 1.49	2.42 ± 0.09
PC	1.11 ± 0.22	2.79 ± 1.08	2.33 ± 0.64	1.93 ± 0.46	11.20 ± 0.75	0.33 ± 0.01
FDOM	33.68 ± 0.51	25.18 ± 0.94	25.36 ± 0.59	23.32 ± 0.63	28.56 ± 0.62	42.84 ± 0.50

2018						
Date	6/14	6/26	7/10	8/7	8/23	9/20
LEC	0.72 ± 0.05	0.76 ± 0.06	0.51 ± 0.08	0.04 ± 0.06	0.93 ± 0.05	1.67 ± 0.03
C	25.45 ± 0.13	27.12 ± 0.11	28.64 ± 0.19	31.93 ± 0.38	25.92 ± 0.15	26.21 ± 0.27
pH	8.82 ± 0.04	8.49 ± 0.08	7.51 ± 0.06	8.50 ± 0.09	7.37 ± 0.11	8.32 ± 0.10
DO	20.73 ± 0.52	16.78 ± 0.71	11.11 ± 0.47	14.05 ± 0.53	11.46 ± 0.66	15.33 ± 0.47
SPC	136.37 ± 0.13	146.25 ± 0.14	162.08 ± 0.18	124.84 ± 1.91	139.56 ± 4.50	146.44 ± 0.21
NTU	9.30 ± 2.03	6.04 ± 0.96	10.17 ± 1.52	18.68 ± 3.70	18.52 ± 0.77	20.47 ± 5.89
CHL	36.64 ± 8.47	26.28 ± 5.85	12.90 ± 5.01	17.01 ± 1.68	13.69 ± 0.79	13.31 ± 1.56
PC	6.07 ± 1.38	4.16 ± 0.89	4.49 ± 1.16	4.72 ± 0.72	5.03 ± 0.24	4.69 ± 0.59
FDOM	39.56 ± 0.86	30.01 ± 0.80	27.51 ± 0.68	40.84 ± 1.17	38.75 ± 0.22	32.76 ± 1.10

Dates marked with an asterisk (*) indicate dates on or in very close proximity to rainfall dates which were not included in other analysis. Parameters are abbreviated as C (temperature; °C), DO (dissolved oxygen; mg L^{-1}), SPC (specific conductivity; µS cm^{-1}), NTU (turbidity; NTU), PC (phycocyanin; RFU), CHL (chlorophyll-a; RFU), and FDOM (fluorescent dissolved organic matter; ug L^{-1}).

Supplemental Table S3. Average values (± S.E.) of water quality parameters measured in ponds 1 and 2 over the study period. Averages and errors are pooled from all observation dates.

n = 5	Pond 1			n = 8	Pond 2		
Parameter	9:00	12:00	3:00	Parameter	9:00	12:00	3:00
LEC	0.90 ± 0.20	0.66 ± 0.27	0.64 ± 0.25	**LEC**	0.67 ± 0.08	0.37 ± 0.13	0.18 ± 0.09
C	26.44 ± 0.88	27.47 ± 0.97	29.43 ± 1.17	**C**	26.07 ± 1.03	28.18 ± 0.98	30.37 ± 0.97
DO	10.26 ± 1.74	10.76 ± 1.52	11.21 ± 1.31	**DO**	10.29 ± 0.69	15.55 ± 0.68	17.57 ± 1.16
SPC	150.94 ± 11.33	156.50 ± 6.99	157.63 ± 6.82	**SPC**	139.18 ± 15.99	139.70 ± 15.95	141.54 ± 15.80
PH	8.30 ± 0.39	8.46 ± 0.37	8.59 ± 0.35	**PH**	7.29 ± 0.15	8.42 ± 0.16	8.88 ± 0.13
NTU	4.99 ± 1.17	4.62 ± 1.31	5.67 ± 1.82	**NTU**	8.80 ± 1.05	8.06 ± 1.31	7.56 ± 1.56
PC	0.91 ± 0.25	0.81 ± 0.26	0.65 ± 0.20	**PC**	4.94 ± 0.73	3.28 ± 0.55	2.08 ± 0.53
CHL	5.23 ± 1.40	4.27 ± 1.21	3.59 ± 1.09	**CHL**	27.10 ± 4.95	15.97 ± 3.31	9.65 ± 2.55
FDOM	12.69 ± 1.88	11.08 ± 1.45	8.91 ± 1.11	**FDOM**	29.08 ± 2.83	28.36 ± 3.38	23.24 ± 3.25
TC	11.77 ± 0.56	11.18 ± 0.34	11.23 ± 0.40	**TC**	16.40 ± 1.84	14.92 ± 1.63	13.57 ± 1.34
TOC	5.38 ± 0.09	5.34 ± 0.23	5.41 ± 0.10	**TOC**	9.85 ± 1.18	8.84 ± 1.20	7.90 ± 0.84
TIC	6.36 ± 0.54	5.77 ± 0.46	5.71 ± 0.40	**TIC**	6.51 ± 0.74	6.05 ± 0.59	5.65 ± 0.56
TNB	3.06 ± 0.37	3.37 ± 0.44	3.42 ± 0.44	**TNB**	2.22 ± 0.33	2.36 ± 0.13	2.74 ± 0.12
PAR	794.88 ± 179.24	1418.23 ± 277.39	1681.62 ± 74.81	**PAR**	962.02 ± 108.51	1806.97 ± 75.52	1557.82 ± 70.97
sPAR	440.01 ± 90.20	919.87 ± 177.63	1070.47 ± 64.47	**sPAR**	436.02 ± 51.39	818.40 ± 113.21	853.86 ± 71.33
NH3	0.10 ± 0.03	0.10 ± 0.03	0.09 ± 0.03	**NH3**	0.18 ± 0.07	0.12 ± 0.04	0.12 ± 0.05
NO2	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	**NO2**	0.01 ± 0.00	0.20 ± 0.12	0.01 ± 0.00
NO3	1.16 ± 0.20	1.16 ± 0.22	1.17 ± 0.22	**NO3**	0.09 ± 0.02	0.07 ± 0.02	0.32 ± 0.06
OP	B.D.L.	B.D.L.	B.D.L.	**OP**	0.19 ± 0.04	0.19 ± 0.04	0.29 ± 0.03

B.D.L. indicates "below detection limit". (abbreviation;unit). LEC = $\log_{10}$ *E. coli* (log CFU 100 mL^{-1}), C = temperature (°C), DO = dissolved oxygen (mg L^{-1}), SPC = specific conductance (µS cm^{-1}), pH (unitless), PC = phycocyanin (relative fluorescence unit (RFU)), CHL = chlorophyll-a (RFU), FDOM = fluorescent dissolved organic matter (µg L^{-1}), TC = total carbon (mg L^{-1}), TIC = total inorganic carbon (mg L^{-1}), TOC = total organic carbon (mg L^{-1}), TNB = total bound nitrogen (mg L^{-1}), PAR = photosynthetic active radiation (µmol m^2 s^{-1}), sPAR = submerged photosynthetic active radiation (µmol m^2 s^{-1}), NH_3 = ammonia (mg L^{-1}), NO_2 = nitrite (mg L^{-1}), NO_3 = nitrate (mg L^{-1}), OP = orthophosphate (mg L^{-1}).

Supplemental Table S4. Average values (± S.E.) of water quality parameters measured in Pond 3 and the Pond 3 Inflow over the study period. Averages and errors are pool all observation dates.

n = 4	Pond 3			n = 2	Pond 3 Inflow		
Parameter	9:00	12:00	3:00	Parameter	9:00	12:00	3:00
LEC	0.92 ± 0.30	0.85 ± 0.36	0.82 ± 0.31	**LEC**	1.87 ± 0.30	1.71 ± 0.06	1.78 ± 1.02
C	29.20 ± 0.48	31.18 ± 0.35	33.00 ± 0.35	**C**	23.54 ± 0.06	24.20 ± 0.19	26.18 ± 0.32
DO	8.16 ± 0.72	9.64 ± 0.53	10.01 ± 0.40	**DO**	4.98 ± 4.15	5.45 ± 6.75	5.20 ± 5.55
SPC	89.28 ± 3.77	90.24 ± 3.79	91.24 ± 3.65	**SPC**	222.35 ± 0.15	221.75 ± 0.16	227.65 ± 0.43
PH	7.66 ± 0.34	8.22 ± 0.27	8.55 ± 0.15	**PH**	5.77 ± 1.99	5.91 ± 2.25	6.15 ± 1.22
NTU	5.66 ± 0.40	4.52 ± 0.68	3.82 ± 0.66	**NTU**	2.57 ± 0.03	3.08 ± 0.10	3.56 ± 0.08
PC	1.06 ± 0.15	0.55 ± 0.21	0.44 ± 0.19	**PC**	0.03 ± 0.14	0.10 ± 0.09	0.08 ± 0.05
CHL	6.80 ± 1.04	3.26 ± 0.71	2.97 ± 0.97	**CHL**	0.43 ± 37.18	0.62 ± 29.85	0.55 ± 34.04
FDOM	29.99 ± 3.84	28.23 ± 2.24	25.67 ± 1.62	**FDOM**	82.75 ± 34.32	66.10 ± 27.29	71.26 ± 28.04
TC	N.M.	N.M.	N.M.	**TC**	N.M.	N.M.	N.M.
TOC	N.M.	N.M.	N.M.	**TOC**	N.M.	N.M.	N.M.
TIC	N.M.	N.M.	N.M.	**TIC**	N.M.	N.M.	N.M.
TNB	N.M.	N.M.	N.M.	**TNB**	N.M.	N.M.	N.M.
PAR	611.60 ± 118.69	1483.31 ± 88.67	1105.38 ± 154.30	**PAR**	N.M.	N.M.	N.M.
sPAR	461.24 ± 64.24	1016.22 ± 136.42	739.64 ± 160.75	**sPAR**	N.M.	N.M.	N.M.
NH3	B.D.L	B.D.L	B.D.L	**NH3**	N.M.	N.M.	N.M.
NO2	B.D.L	B.D.L	B.D.L	**NO2**	N.M.	N.M.	N.M.
NO3	B.D.L	B.D.L	B.D.L	**NO3**	N.M.	N.M.	N.M.
OP	B.D.L	B.D.L	B.D.L	**OP**	N.M.	N.M.	N.M.

B.D.L. and N.M. indicate "below detection limit" and "not measured", respectively. (abbreviation;unit). LEC = $\log_{10}$ *E. coli* (log CFU 100 mL^{-1}), C = temperature (°C), DO = dissolved oxygen (mg L^{-1}), SPC = specific conductance (µS cm^{-1}), pH (unitless), PC = phycocyanin (relative fluorescence unit (RFU)), CHL = chlorophyll-a (RFU), FDOM = fluorescent dissolved organic matter (µg L^{-1}), TC = total carbon (mg L^{-1}), TIC = total inorganic carbon (mg L^{-1}), TOC = total organic carbon (mg L^{-1}), TNB = total bound nitrogen (mg L^{-1}), PAR = photosynthetic active radiation (µmol m^2 s^{-1}), sPAR = submerged photosynthetic active radiation (µmol m^2 s^{-1}), NH_3 = ammonia (mg L^{-1}), NO_2 = nitrite (mg L^{-1}), NO_3 = nitrate (mg L^{-1}), OP = orthophosphate (mg L^{-1}).

Table S5. Results of Kruskal-Wallis ANOVA testing with time of day as a factor.

P1		P2		P3	
Date	*p*	Date	*p*	Date	*P*
6/6/2019	0.837	6/4/2019	**0.002**	7/29/2020	0.918
7/2/2019	**0.001**	6/26/2019	**<0.001**	8/12/2020	0.426
8/8/2019	**<0.001**	8/30/2019	**<0.001**	7/7/2021	0.186
9/6/2019	0.678	9/21/2019	0.556	8/13/2021	0.404
7/23/2020	**0.007**	7/15/2020	0.598		
		8/10/2020	**0.004**		
		8/26/2020	**0.032**		
		7/15/2021	**0.002**		

Results in **bold** are statistically significant ($P < 0.05$)

Table S6. Values of the *E. coli* die-off rate constant (k h^{-1}) across ponds and for comparisons between different times of day.

		P1	
Date	9:00 and 12:00	9:00 and 15:00	12:00 and 15:00
6/6/2019	0.106	0.069	0.032
7/2/2019	0.502	0.210	-0.081
8/8/2019	0.090	0.124	0.161
9/9/2019	0.046	0.044	0.041
7/23/2020	0.193	0.182	0.170
$\bar{x}$	0.187	0.126	0.065
S.E.	0.082	0.032	0.046
		P2	
6/4/2019	0.799	0.311	-0.177
6/26/2019	0.147	0.200	0.253
8/30/2019	0.117	0.311	0.502
9/21/2019	0.203	0.117	0.032
7/15/2020	0.053	0.124	0.198
8/10/2020	0.219	0.230	0.240
8/26/2020	0.143	0.226	0.311
7/15/2021	0.447	0.364	0.283
$\bar{x}$	0.266	0.235	0.205
S.E.	0.087	0.031	0.071
		P3	
7/29/2020	0.005	0.018	0.030
8/12/2020	-0.076	-0.005	0.064
7/7/2021	0.166	0.037	-0.094
8/13/2021	0.012	0.090	0.168
$\bar{x}$	0.026	0.035	0.042
S.E.	0.051	0.020	0.054

Averages ($\bar{x}$) and standard error of the means (S.E.) are presented for each pond.

Supplementary Table S7. Average values of predictor variables measured in each year.

	Pond 1			Pond 2		
Variable	2016	2017	2018	2016	2017	2018
EC	0.90±0.05	1.67±0.03	1.12±0.05	1.03±0.04	0.83±0.04	0.81±0.04
pH	8.64±0.06	8.82±0.04	7.65±0.04	7.61±0.10	8.46±0.06	8.17±0.05
DO	10.04±0.23	10.39±0.10	10.07±0.12	9.45±0.33	12.97±0.30	14.91±0.32
SPC	147.89±1.90	162.51±0.81	148.83±1.75	154.46±1.30	166.95±0.79	142.59±1.13
C	26.24±0.11	26.31±0.18	26.99±0.18	26.95±0.23	28.41±0.19	27.54±0.18
NTU	7.05±0.27	5.16±0.26	4.20±0.28	36.34±8.33	11.44±1.05	13.86±1.30
CHL	N.M.	3.83±0.41	2.80±0.22	N.M.	15.30±1.61	19.97±2.02
PC	N.M.	1.15±0.09	0.95±0.11	N.M.	3.87±0.42	4.86±0.37
FDOM	N.M.	13.04±1.13	20.72±1.09	N.M.	27.22±0.41	34.90±0.50
TDN	N.M.	N.M.	6.44±0.11	N.M.	N.M.	4.31±0.05
TDC	N.M.	N.M.	13.96±0.11	N.M.	N.M.	15.97±0.08
NH_4^+	N.M.	N.M.	0.65±0.07	N.M.	N.M.	0.23±0.01
$[PO_4]^{3-}$ × 100	N.M.	N.M.	1.39±0.12	N.M.	N.M.	22.82±0.72

N.M. represents variables that were not measured in a specific year. The±symbol separates the average from the standard error. EC (log CFU 100 mL^{-1}), pH (unitless), DO (mg L^{-1}), SPC (µS cm^{-1}), C (°C), NTU (NTU), CHL (RFU), PC (RFU), FDOM (µg L^{-1}), TDN (mg L^{-1}), TDC (mg L^{-1}), NH_4^+ (mg L^{-1}), and $[PO_4]^{3-}$ (mg L^{-1}).

Supplemental Table S8. Probabilities for mean RMSE values to be the same over pairs of years.

	Predictor set A			Predictor set AB
Algorithm	2016-2017	2017-2018	2016-2018	2017-2018
		Pond P1		
GBM	0.940	**0.006**	**0.018**	**0.033**
kNN	0.951	**0.016**	**0.028**	0.072
MLR	0.063	**< 0.001**	0.255	**< 0.001**
RF	0.916	**0.045**	0.074	**0.042**
SVM	0.748	**0.012**	**0.022**	**0.009**
		Pond P2		
GBM	**0.044**	0.236	0.509	0.209
kNN	0.438	0.813	0.811	0.534
MLR	0.351	0.432	**0.008**	0.188
RF	**0.008**	0.052	0.963	0.063
SVM	**0.005**	0.139	0.614	0.210

Bold values indicate statistical significance.

Supplemental Table S9. Average mean absolute errors of logarithms of *E. coli* concentrations predicted with four machine learning algorithms and multiple linear regression.

	Predictor set A				Predictor set AB			Predictor set ABC
ML Algorithm	2016	2017	2018	2016-2018	2017	2018	2017-2018	2018
				Pond P1				
GBM	0.193±0.066	0.182±0.007	0.269±0.010	0.247±0.037	0.182±0.048	0.262±0.008	0.236±0.005	**0.268±0.061**
kNN	0.217±0.088	0.193±0.006	0.294±0.011	0.257±0.044	0.193±0.066	0.286±0.010	0.250±0.006	0.281±0.093
MLR	0.338±0.132	0.205±0.007	0.435±0.012	0.389±0.048	0.202±0.047	0.408±0.012	0.346±0.007	0.351±0.076
RF	**0.188±0.070**	0.182±0.007	**0.264±0.011**	**0.239±0.042**	**0.175±0.048**	**0.254±0.010**	**0.230±0.006**	0.269±0.073
SVM	0.214±0.073	**0.178±0.008**	0.288±0.009	0.248±0.044	0.185±0.045	0.288±0.009	0.245±0.006	0.287±0.072
				Pond P2				
GBM	0.265±0.070	0.324±0.009	0.293±0.006	0.305±0.035	0.323±0.078	0.289±0.006	0.310±0.006	0.235±0.054
kNN	0.291±0.076	0.318±0.011	0.315±0.008	0.320±0.041	**0.322±0.062**	0.310±0.008	0.365±0.008	0.306±0.065
MLR	0.317±0.074	0.356±0.009	0.347±0.007	0.393±0.040	0.362±0.062	0.334±0.007	0.394±0.005	0.305±0.059
RF	0.239±0.059	**0.319±0.010**	**0.263±0.007**	**0.289±0.029**	0.323±0.070	**0.266±0.007**	**0.294±0.005**	**0.226±0.046**
SVM	**0.233±0.062**	0.323±0.009	0.275±0.006	0.306±0.034	0.326±0.064	0.288±0.009	0.304±0.007	0.255±0.050

Values in bold show the model with the best performance.

Supplemental Table S10. Average coefficients of determination of logarithms of *E. coli* concentrations predicted with four machine learning algorithms and multiple linear regression.

	Predictor set A				Predictor set AB			Predictor set ABC
ML Algorithm	2016	2017	2018	2016-2018	2017	2018	2017-2018	2018
				Pond 1				
GBM	**0.720 ± 0.230**	0.381 ± 0.011	0.670 ± 0.025	0.628 ± 0.015	0.376 ± 0.032	0.680 ± 0.023	**0.677 ± 0.015**	0.749 ± 0.021
kNN	0.621 ± 0.114	0.304 ± 0.036	0.607 ± 0.028	0.585 ± 0.019	0.263 ± 0.035	0.611 ± 0.029	0.598 ± 0.022	0.722 ± 0.021
MLR	0.407 ± 0.298	0.244 ± 0.031	0.239 ± 0.023	0.210 ± 0.017	0.221 ± 0.029	0.332 ± 0.022	0.355 ± 0.018	0.587 ± 0.012
RF	0.696 ± 0.235	**0.384 ± 0.029**	**0.687 ± 0.026**	**0.647 ± 0.018**	**0.414 ± 0.028**	**0.697 ± 0.024**	0.672 ± 0.019	**0.751 ± 0.020**
SVM	0.631 ± 0.269	0.364 ± 0.033	0.619 ± 0.027	0.601 ± 0.020	0.325 ± 0.026	0.632 ± 0.024	0.625 ± 0.017	0.670 ± 0.030
				Pond 2				
GBM	0.614 ± 0.026	0.402 ± 0.029	0.600 ± 0.016	0.515 ± 0.017	0.394 ± 0.030	0.609 ± 0.018	0.511 ± 0.019	0.726 ± 0.017
kNN	0.503 ± 0.035	**0.449 ± 0.030**	0.562 ± 0.019	0.472 ± 0.018	0.425 ± 0.026	0.552 ± 0.018	0.492 ± 0.021	0.568 ± 0.019
MLR	0.397 ± 0.039	0.295 ± 0.025	0.486 ± 0.019	0.250 ± 0.016	0.279 ± 0.022	0.529 ± 0.020	0.225 ± 0.015	0.592 ± 0.021
RF	0.623 ± 0.035	0.436 ± 0.031	**0.669 ± 0.019**	**0.568 ± 0.015**	**0.424 ± 0.028**	**0.667 ± 0.019**	**0.551 ± 0.016**	**0.745 ± 0.014**
SVM	**0.693 ± 0.034**	0.411 ± 0.032	0.630 ± 0.017	0.519 ± 0.014	0.384 ± 0.029	0.609 ± 0.020	0.492 ± 0.022	0.675 ± 0.019

Values in bold show the model with the best performance.

Supplemental Table S11. Average normalized root-mean-squared errors (NRMSE) of logarithms of *E. coli* concentrations predicted with four machine learning algorithms and multiple linear regression.

	Predictor set A				Predictor set AB			Predictor set ABC
ML Algorithm	2016	2017	2018	2016-2018	2017	2018	2017-2018	2018
				Pond P1				
SGB	**0.097 ± 0.004**	**0.071 ± 0.003**	0.126 ± 0.005	0.098 ± 0.003	0.073 ± 0.003	0.124 ± 0.004	0.093 ± 0.002	0.120 ± 0.004
kNN	0.110 ± 0.006	0.079 ± 0.003	0.141 ± 0.005	0.105 ± 0.003	0.081 ± 0.005	0.138 ± 0.006	0.102 ± 0.003	0.129 ± 0.006
MLR	0.178 ± 0.013	0.082 ± 0.004	0.199 ± 0.006	0.144 ± 0.003	0.082 ± 0.004	0.185 ± 0.005	0.132 ± 0.002	0.160 ± 0.004
RF	0.100 ± 0.006	**0.071 ± 0.003**	**0.124 ± 0.005**	**0.095 ± 0.003**	**0.070 ± 0.004**	**0.121 ± 0.005**	**0.092 ± 0.003**	**0.119 ± 0.005**
SVM	0.106 ± 0.005	0.073 ± 0.003	0.137 ± 0.005	0.102 ± 0.003	0.074 ± 0.003	0.136 ± 0.005	0.098 ± 0.003	0.133 ± 0.005
				Pond P2				
SGB	0.111 ± 0.004	0.156 ± 0.005	0.166 ± 0.003	0.149 ± 0.003	0.159 ± 0.006	0.163 ± 0.003	0.149 ± 0.003	0.137 ± 0.004
kNN	0.123 ± 0.005	**0.154 ± 0.006**	0.176 ± 0.003	0.157 ± 0.003	0.157 ± 0.004	0.174 ± 0.004	0.151 ± 0.003	0.172 ± 0.004
MLR	0.140 ± 0.005	0.171 ± 0.004	0.189 ± 0.003	0.187 ± 0.003	0.173 ± 0.004	0.182 ± 0.004	0.187 ± 0.002	0.170 ± 0.004
RF	0.102 ± 0.004	**0.154 ± 0.005**	**0.150 ± 0.004**	**0.141 ± 0.003**	**0.155 ± 0.005**	**0.149 ± 0.003**	**0.143 ± 0.003**	**0.132 ± 0.003**
SVM	**0.096 ± 0.004**	0.157 ± 0.005	0.159 ± 0.003	0.150 ± 0.003	0.160 ± 0.005	0.164 ± 0.005	0.150 ± 0.003	0.148 ± 0.004

The ± separates the average from the standard error of the mean. The smallest RMSE are shown in bold. Machine learning algorithms: stochastic gradient boosting machines (SGB); k-nearest neighbor (kNN); multiple linear regression (MLR); random forest (RF); support vector machines (SVM). Predictor sets: A – temperature, DO, pH, turbidity, and SPC; AB – all from A and PC, CHL, and *f*DOM ; ABC – all from AB and NH_4^+, $[PO4]^{3-}$, TN, and TC.

Supplemental Table S12. Average normalized mean absolute errors (NMAE) of logarithms of *E. coli* concentrations predicted with four machine learning algorithms and multiple linear regression.

	Predictor set A				Predictor set AB			Predictor set ABC
ML Algorithm	2016	2017	2018	2016-2018	2017	2018	2017-2018	2018
				Pond P1				
SGB	0.076 ± 0.004	0.052 ± 0.002	0.096 ± 0.004	0.071 ± 0.011	0.052 ± 0.014	0.094 ± 0.003	0.067 ± 0.001	**0.096 ± 0.022**
kNN	0.085 ± 0.005	0.055 ± 0.002	0.105 ± 0.004	0.073 ± 0.013	0.055 ± 0.019	0.102 ± 0.004	0.071 ± 0.002	0.100 ± 0.033
MLR	0.133 ± 0.007	0.059 ± 0.002	0.155 ± 0.004	0.111 ± 0.014	0.058 ± 0.013	0.146 ± 0.004	0.099 ± 0.002	0.125 ± 0.027
RF	**0.074 ± 0.004**	0.052 ± 0.002	**0.094 ± 0.004**	**0.068 ± 0.012**	**0.050 ± 0.014**	**0.091 ± 0.004**	**0.066 ± 0.002**	0.096 ± 0.026
SVM	0.084 ± 0.004	**0.051 ± 0.002**	0.103 ± 0.003	0.071 ± 0.013	0.053 ± 0.013	0.103 ± 0.003	0.070 ± 0.002	0.103 ± 0.026
				Pond P2				
SGB	0.088 ± 0.003	0.120 ± 0.003	0.127 ± 0.003	0.113 ± 0.013	0.120 ± 0.029	0.126 ± 0.003	0.115 ± 0.002	0.102 ± 0.023
kNN	0.097 ± 0.004	0.118 ± 0.004	0.137 ± 0.003	0.119 ± 0.015	**0.119 ± 0.023**	0.135 ± 0.003	0.135 ± 0.003	0.133 ± 0.028
MLR	0.106 ± 0.003	0.132 ± 0.003	0.151 ± 0.003	0.146 ± 0.015	0.134 ± 0.023	0.145 ± 0.003	0.146 ± 0.002	0.133 ± 0.026
RF	0.080 ± 0.003	**0.118 ± 0.004**	**0.114 ± 0.003**	**0.107 ± 0.011**	0.120 ± 0.026	**0.116 ± 0.003**	**0.109 ± 0.002**	**0.098 ± 0.020**
SVM	**0.078 ± 0.003**	0.120 ± 0.003	0.120 ± 0.003	0.113 ± 0.013	0.121 ± 0.024	0.125 ± 0.004	0.113 ± 0.003	0.111 ± 0.022

The ± separates the average from the standard error of the mean. The smallest RMSE are shown in bold. Machine learning algorithms: stochastic gradient boosting machines (SGB); k-nearest neighbor (kNN); multiple linear regression (MLR); random forest (RF); support vector machines (SVM). Predictor sets: A – temperature, DO, pH, turbidity, and SPC; AB – all from A and PC, CHL, and *f*DOM ; ABC – all from AB and NH_4^+, $[PO4]^{3-}$, TN, and TC.

Supplemental Table S13**.** Average root-mean-squared errors (RMSE), coefficients of determination (R^2), and mean absolute errors (MAE) of logarithms of *E. coli* concentrations predicted with four machine learning algorithms and multiple linear regression for the combined datasets from each pond over the observation period using predictor set A.

RMSE				
ML Algorithm	2016	2017	2018	2016-2018
SGB	0.318 ± 0.012	0.397 ± 0.009	0.382 ± 0.008	0.394 ± 0.005
kNN	0.354 ± 0.011	0.400 ± 0.012	0.407 ± 0.009	0.418 ± 0.007
MLR	0.441 ± 0.012	0.505 ± 0.010	0.529 ± 0.009	0.543 ± 0.006
RF	0.307 ± 0.009	**0.383 ± 0.012**	**0.356 ± 0.008**	**0.377 ± 0.006**
SVM	**0.295 ± 0.011**	0.394 ± 0.010	0.383 ± 0.007	0.398 ± 0.005
R2				
ML Algorithm	2016	2017	2018	2016-2018
SGB	0.584 ± 0.029	0.608 ± 0.016	0.626 ± 0.013	0.591 ± 0.010
kNN	0.509 ± 0.029	0.596 ± 0.020	0.581 ± 0.016	0.546 ± 0.012
MLR	0.265 ± 0.028	0.357 ± 0.018	0.302 ± 0.023	0.233 ± 0.014
RF	0.629 ± 0.022	**0.636 ± 0.022**	**0.677 ± 0.015**	**0.630 ± 0.010**
SVM	**0.655 ± 0.023**	0.610 ± 0.019	0.626 ± 0.013	0.586 ± 0.012
MAE				
ML Algorithm	2016	2017	2018	2016-2018
SGB	0.245 ± 0.008	0.296 ± 0.007	0.295 ± 0.006	0.293 ± 0.003
kNN	0.271 ± 0.008	0.288 ± 0.008	0.303 ± 0.006	0.306 ± 0.004
MLR	0.339 ± 0.009	0.391 ± 0.008	0.427 ± 0.007	0.424 ± 0.005
RF	0.229 ± 0.006	**0.278 ± 0.008**	**0.268 ± 0.005**	**0.278 ± 0.004**
SVM	**0.225 ± 0.008**	0.284 ± 0.007	0.288 ± 0.005	0.293 ± 0.004

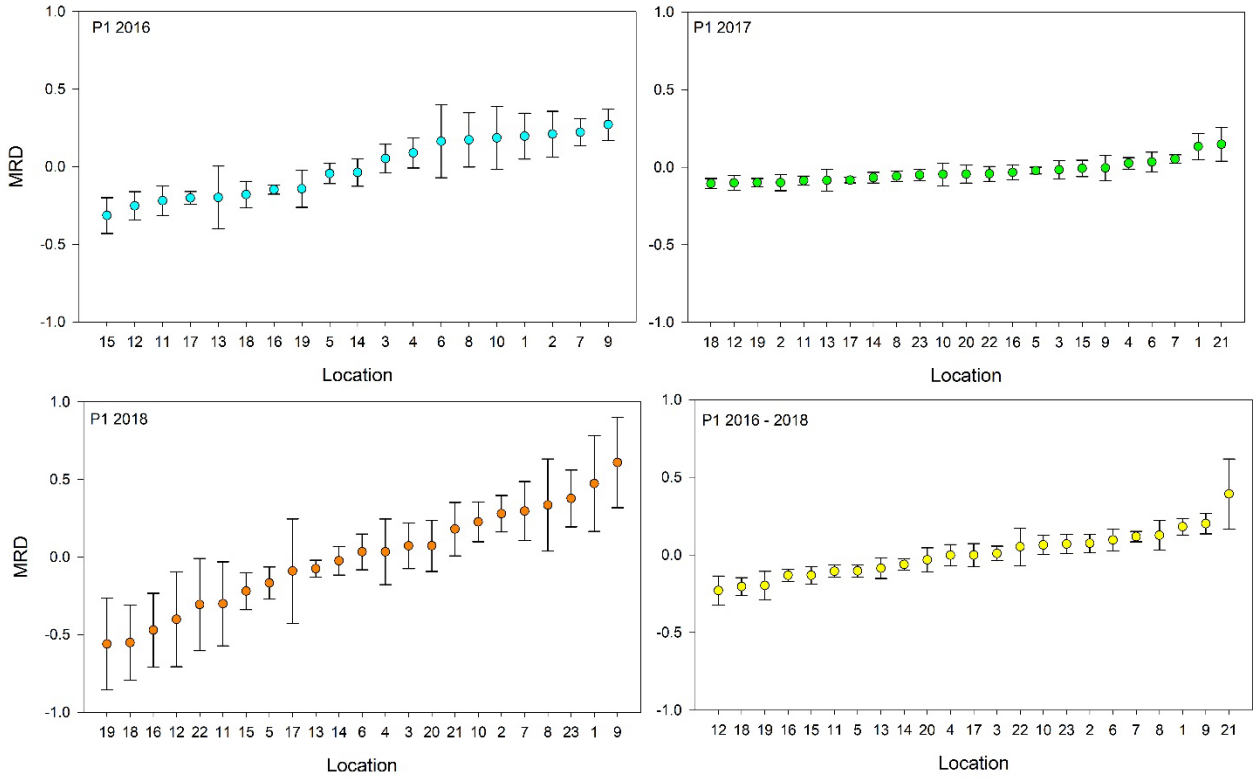
Values highlighted in **bold** show the model with the best performance.

Supplemental Table S14**.** Average root-mean-squared errors (RMSE), coefficients of determination (R^2), and mean absolute errors (MAE) of logarithms of *E. coli* concentrations predicted with four machine learning algorithms and multiple linear regression for the combined datasets from each pond over the observation period using predictor set A and including 'site' (e.g. Pond 1 or Pond 2) as a categorical variable.

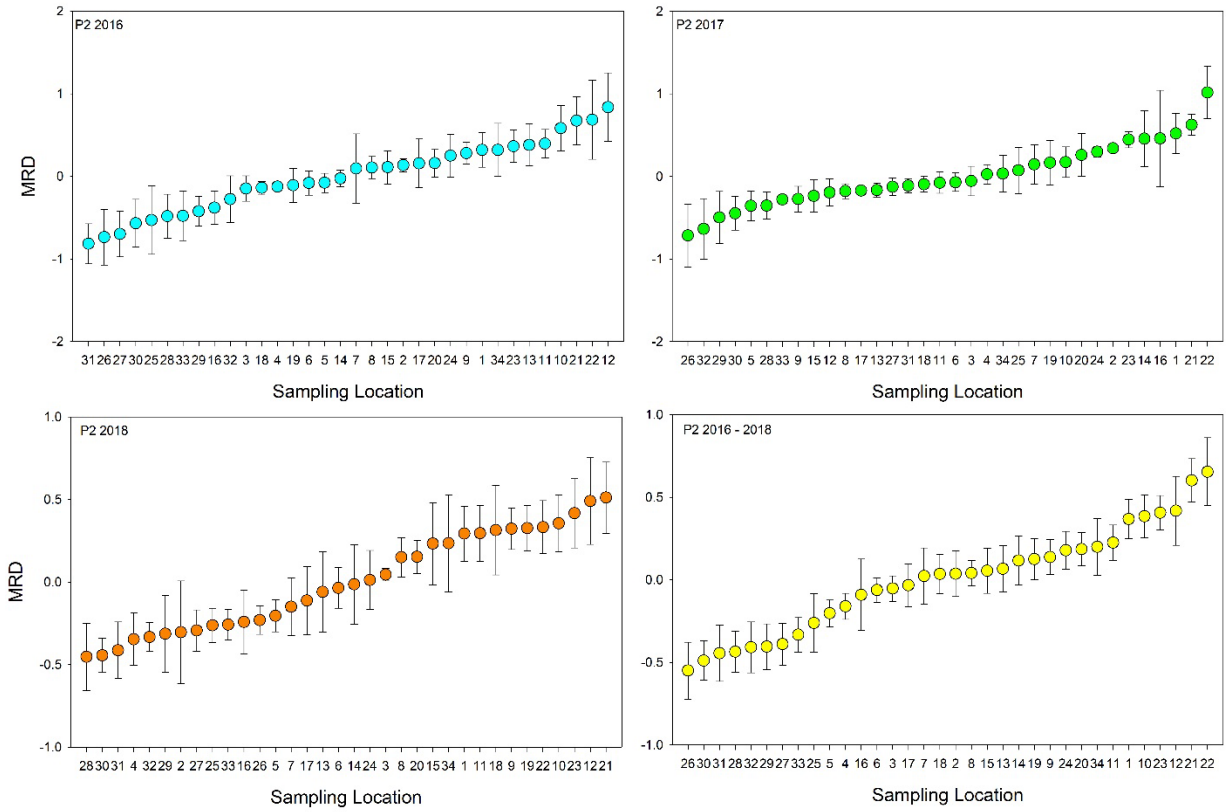
RMSE				
ML Algorithm	2016	2017	2018	2016-2018
SGB	0.316 ± 0.013	0.365 ± 0.010	0.376 ± 0.007	0.386 ± 0.006
kNN	0.354 ± 0.012	0.394 ± 0.010	0.403 ± 0.010	0.393 ± 0.007
MLR	0.467 ± 0.029	0.405 ± 0.008	0.526 ± 0.008	0.530 ± 0.007
RF	0.301 ± 0.011	**0.359 ± 0.009**	**0.352 ± 0.008**	**0.365 ± 0.005**
SVM	**0.295 ± 0.009**	0.369 ± 0.009	0.378 ± 0.008	0.384 ± 0.006
R2				
ML Algorithm	2016	2017	2018	2016-2018
SGB	0.604 ± 0.026	0.667 ± 0.016	0.639 ± 0.013	0.606 ± 0.012
kNN	0.518 ± 0.028	0.618 ± 0.019	0.586 ± 0.018	0.599 ± 0.012
MLR	0.262 ± 0.031	0.593 ± 0.014	0.294 ± 0.019	0.265 ± 0.012
RF	**0.640 ± 0.022**	**0.674 ± 0.014**	**0.681 ± 0.014**	**0.651 ± 0.009**
SVM	0.634 ± 0.025	0.661 ± 0.014	0.636 ± 0.015	0.616 ± 0.010
MAE				
ML Algorithm	2016	2017	2018	2016-2018
SGB	0.245 ± 0.010	0.271 ± 0.006	0.291 ± 0.004	0.287 ± 0.004
kNN	0.271 ± 0.009	0.283 ± 0.007	0.302 ± 0.007	0.289 ± 0.004
MLR	0.352 ± 0.014	0.301 ± 0.006	0.422 ± 0.006	0.406 ± 0.004
RF	**0.227 ± 0.007**	**0.263 ± 0.006**	**0.264 ± 0.005**	**0.269 ± 0.003**
SVM	**0.227 ± 0.007**	0.264 ± 0.005	0.278 ± 0.006	0.281 ± 0.004

Values highlighted in **bold** show the model with the best performance.

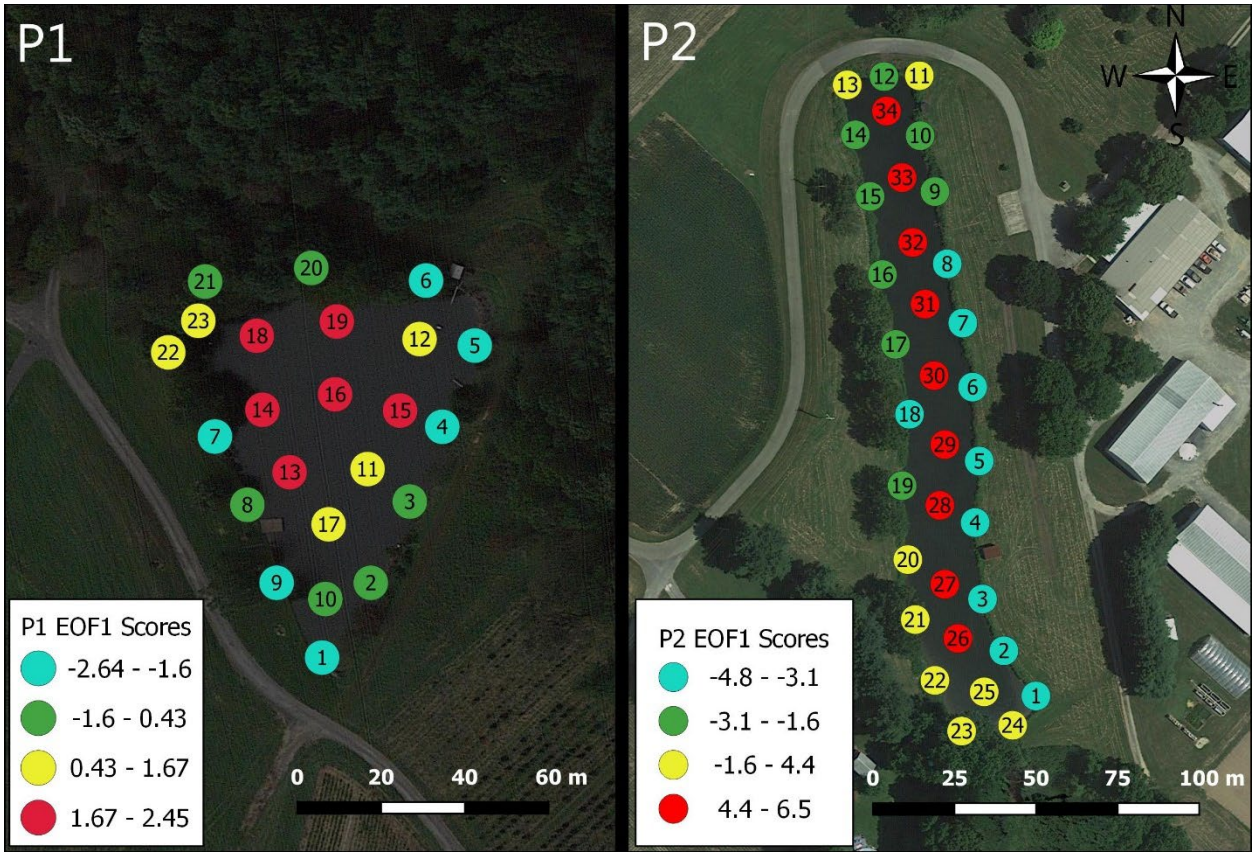
Appendix C: Supplemental Figures



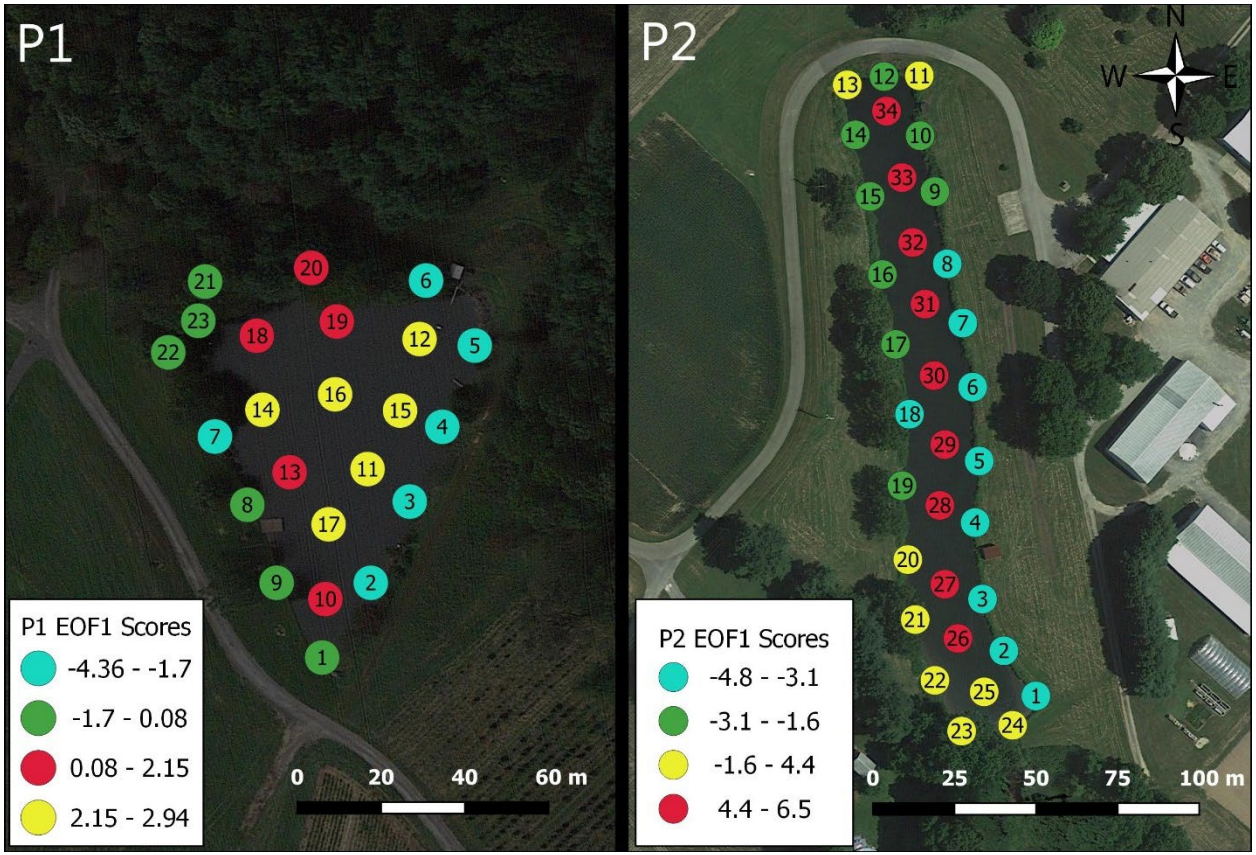
Supplemental Figure S1. Mean relative difference values for each year at P1 as well as the 3-year average.



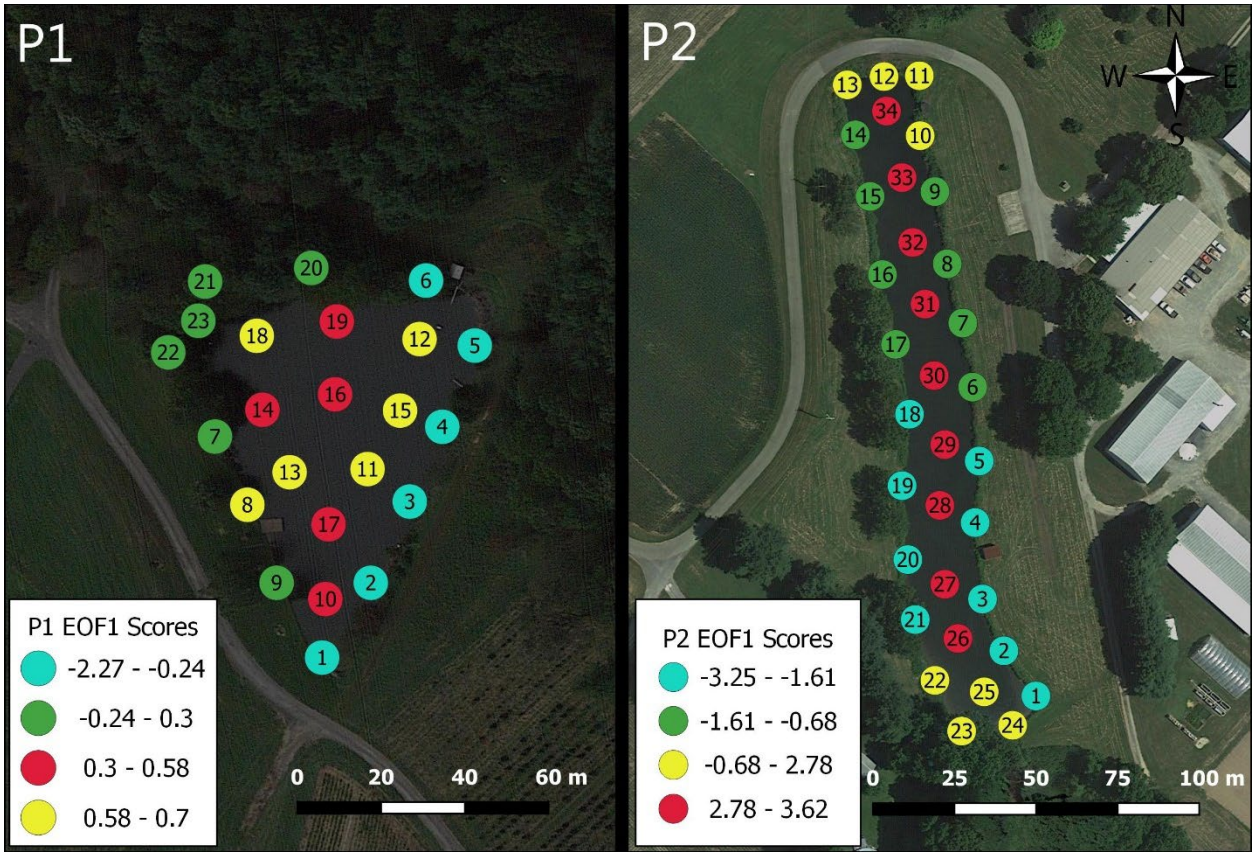
Supplemental Figure S2. Mean relative difference values for each year at P2 as well as the 3-year average.



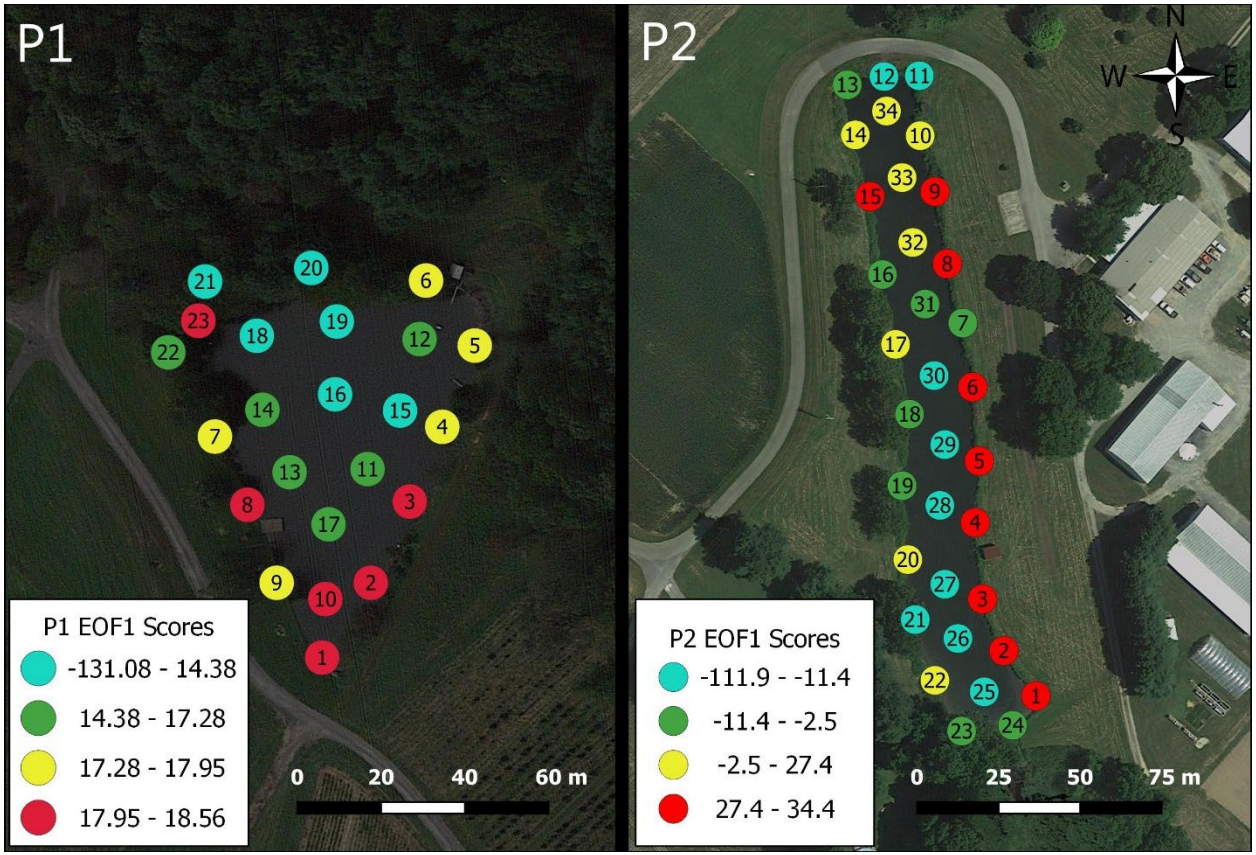
Supplemental Figure S3. Spatial distribution of EOF_1 values of temperature across the two observation ponds between 2016 and 2018.



Supplemental Figure S4. Spatial distribution of EOF_1 values of dissolved oxygen (DO) across the two observation ponds between 2016 and 2018.

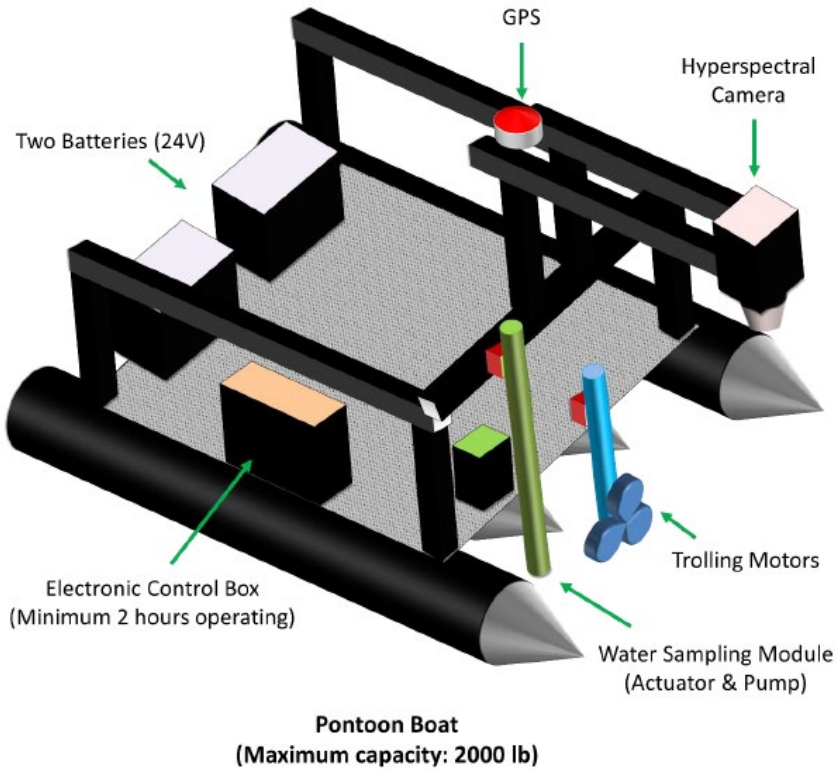


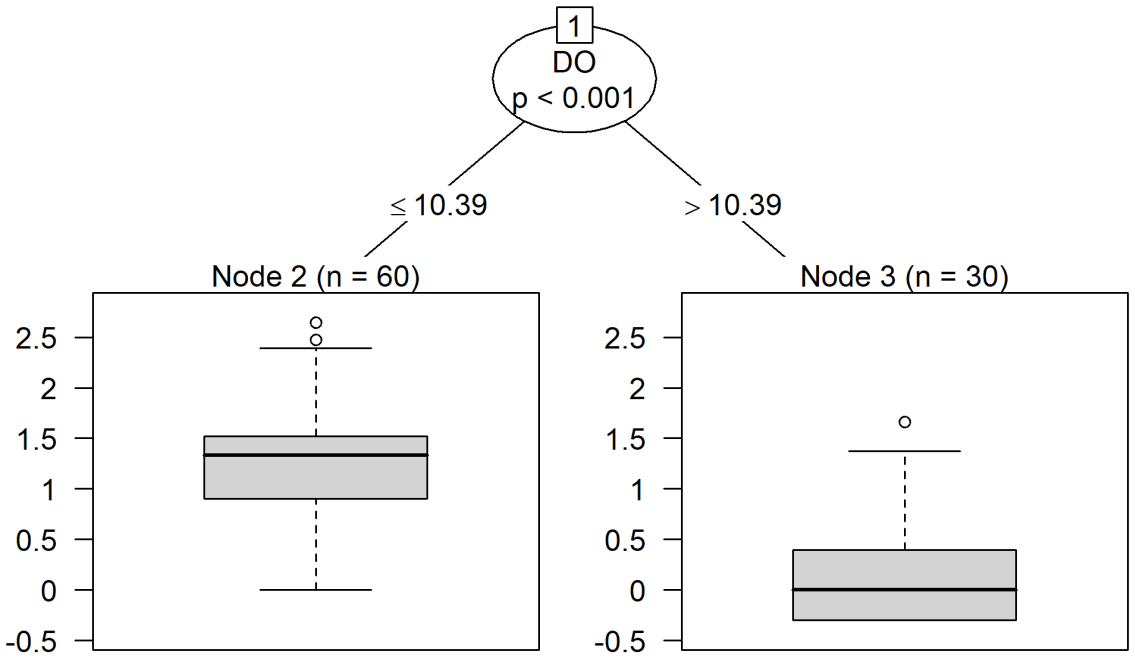
Supplemental Figure S5. Spatial distribution of EOF_1 values of pH across the two observation ponds between 2016 and 2018.



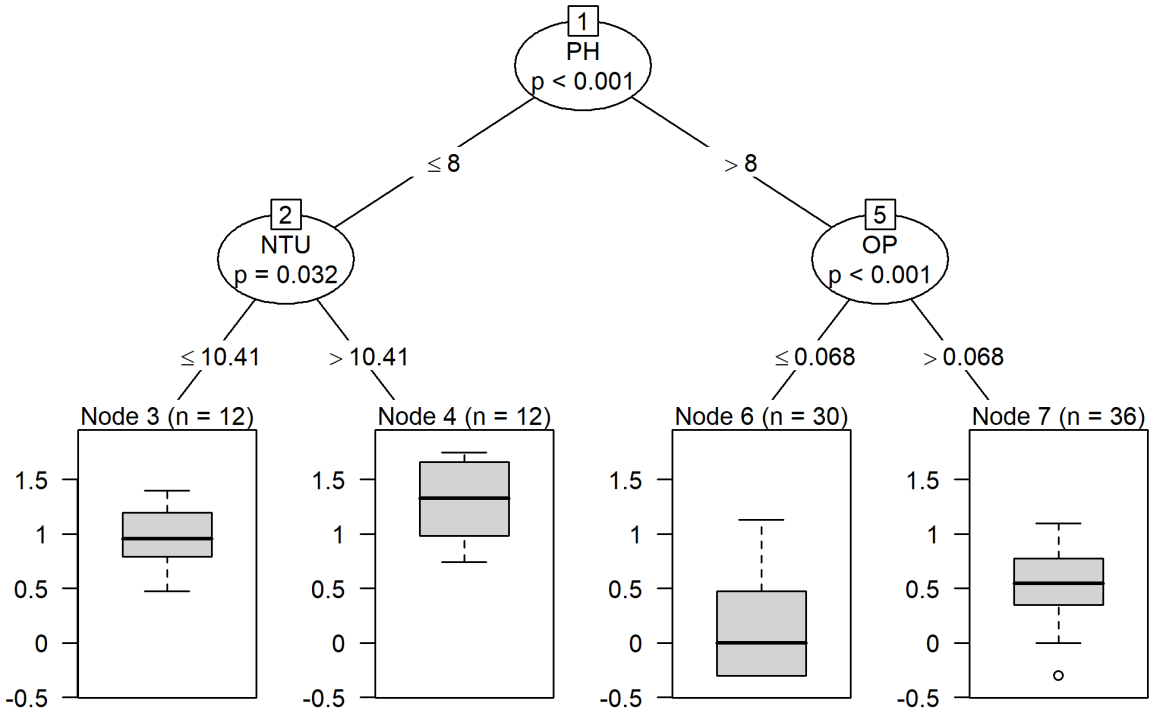
Supplemental Figure S6. Spatial distribution of EOF_1 values of specific conductivity (SPC) across the two observation ponds between 2016 and 2018.

Supplemental Figure S7. Schematics of the Mobile, Sampling, Sensing, and Hyperspectral Imaging Platform (MSSHIP).

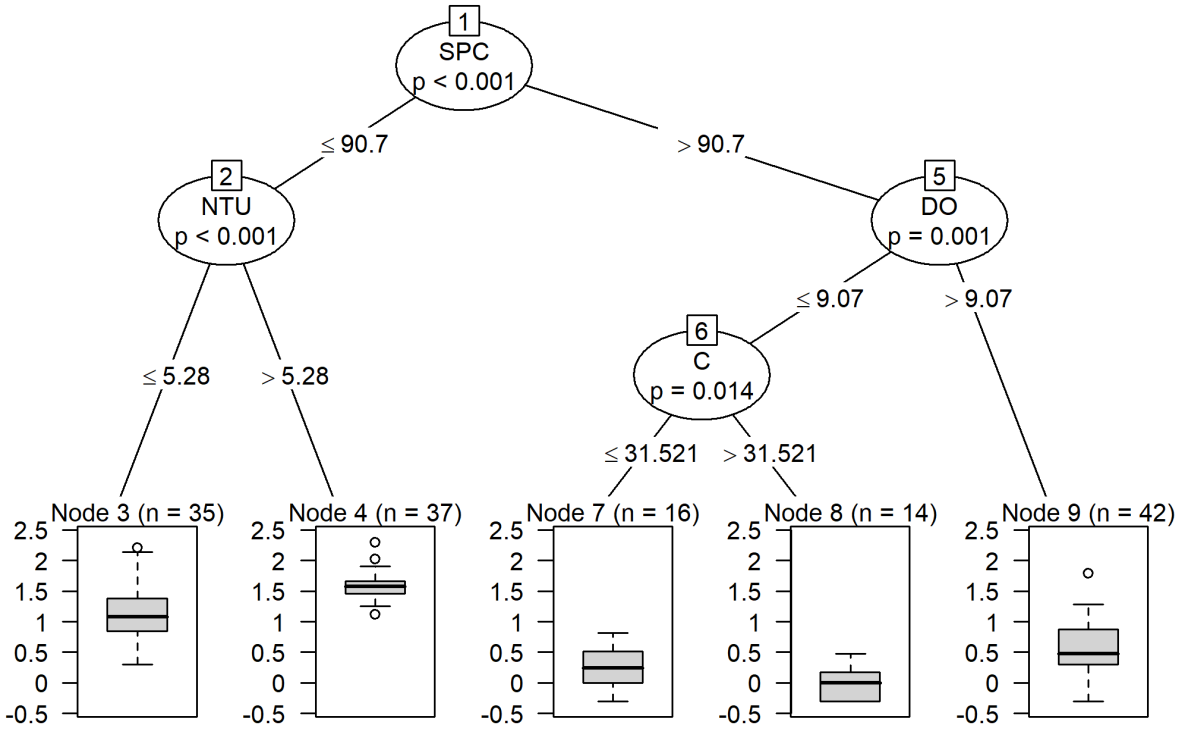




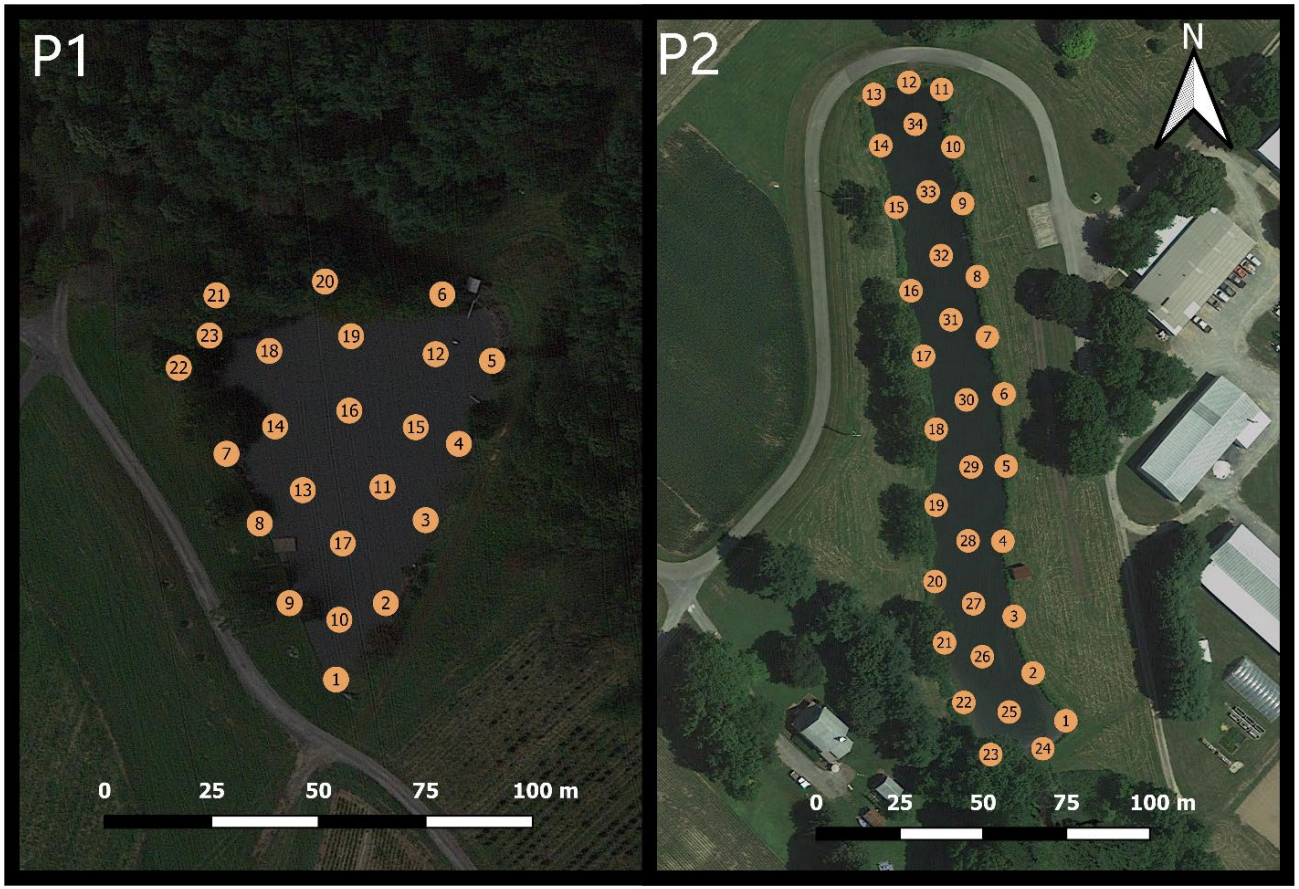
Supplemental Figure S8. The full conditional inference tree (cTREE) for Pond 1. The y-axis of the terminal nodes shows the base$_{10}$ logarithm of the observed *E. coli* concentration. The solid line in the box and whiskers show the median while the bottom and top boarders of the box display the 1st (Q1) and third (Q3) quartiles also known as the interquartile range (IQR), respectively. Outliers (< Q1-1.5*IQR or > Q3+1.5*IQR) of the concentration distribution are shown as dots. Dissolved oxygen is abbreviated by DO and the splitting concentration of 10.39 is in units mg L^{-1}. The value *n* next to the node number indicates the number of observations in that split.



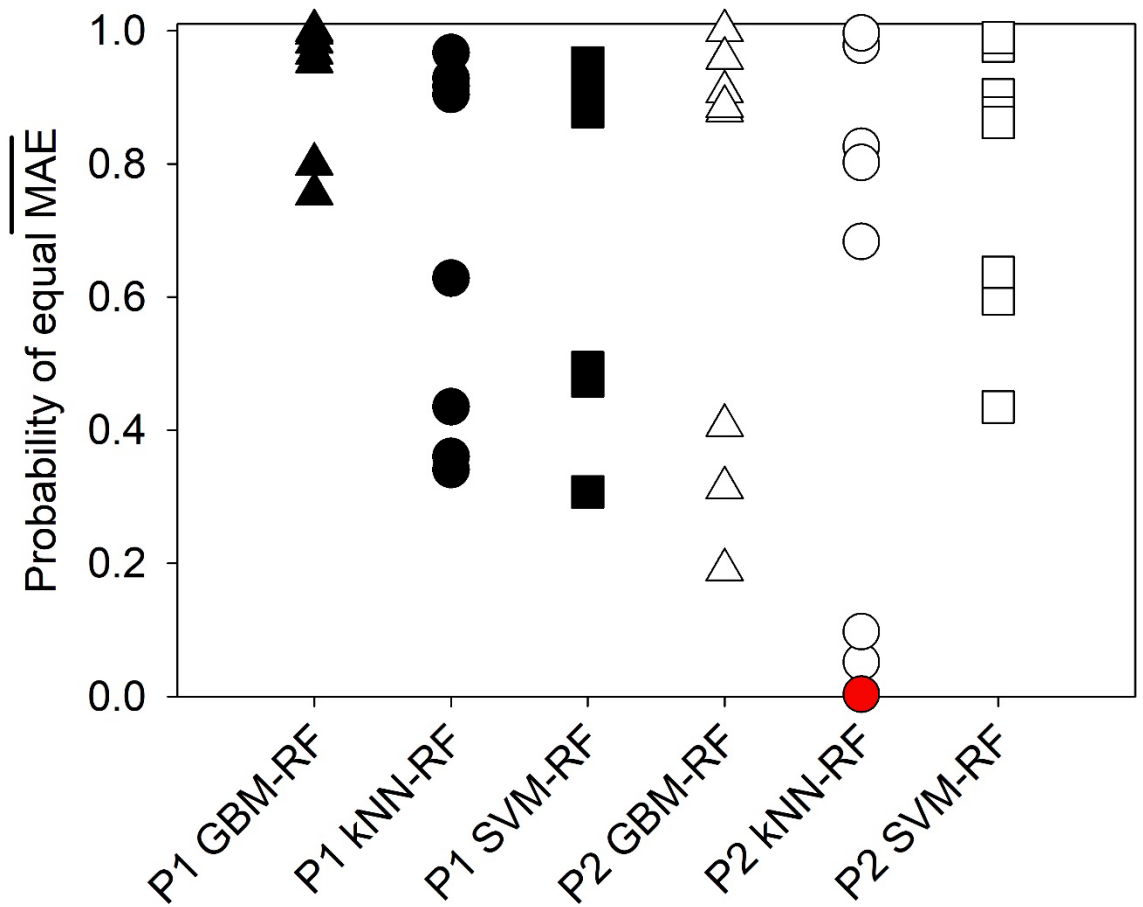
Supplemental Figure S9. The full conditional inference tree (cTREE) for Pond 2. The y-axis of the terminal nodes shows the base$_{10}$ logarithm of the observed *E. coli* concentration. The solid line in the box and whiskers show the median while the bottom and top boarders of the box display the 1st (Q1) and third (Q3) quartiles also known as the interquartile range (IQR), respectively. Outliers (< Q1-1.5*IQR or > Q3+1.5*IQR) of the concentration distribution are shown as dots. pH is abbreviated as pH (unitless); turbidity is abbreviated as NTU (NTU), and OP is the abbreviation for the orthophosphate concentration (mg L^{-1}). The value *n* next to the node number indicates the number of observations in that split.



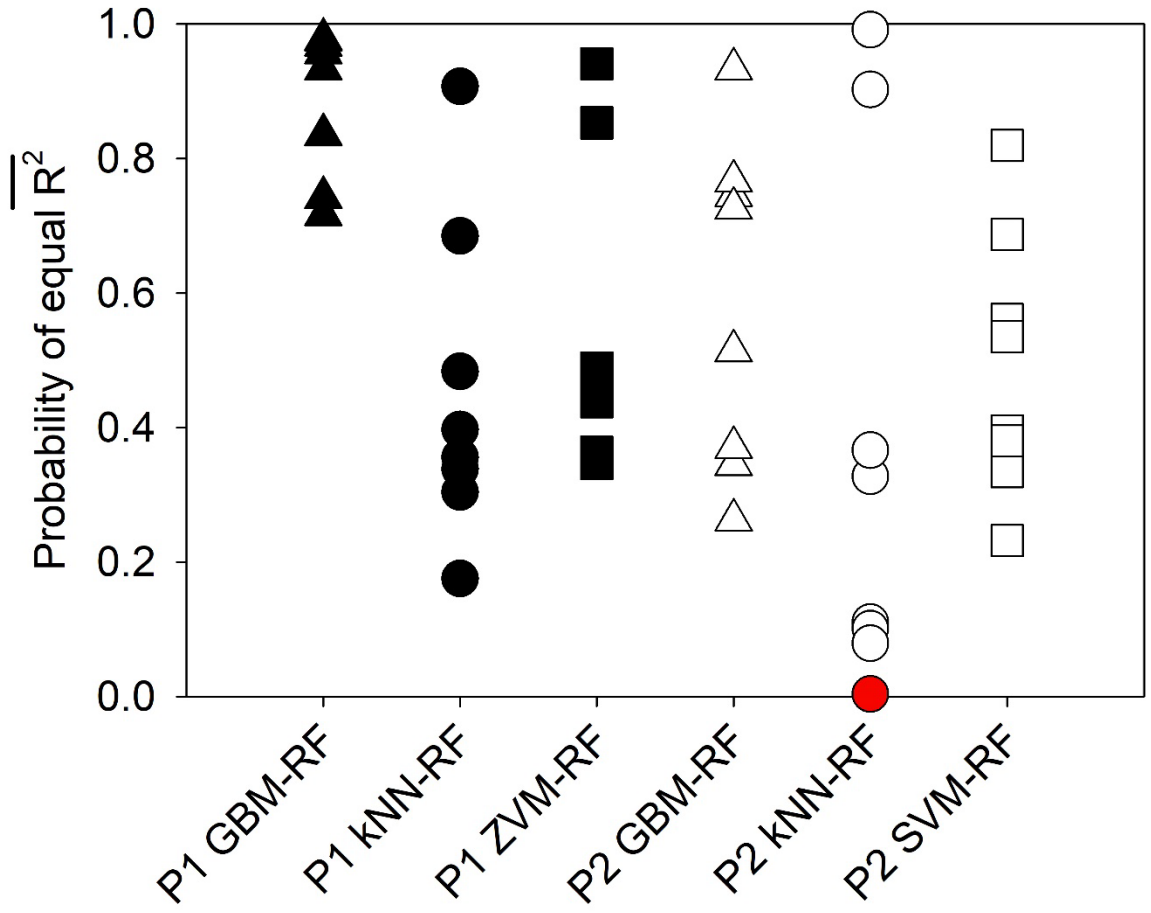
Supplemental Figure S10. The full conditional inference tree (cTREE) for Pond 3. The y-axis of the terminal nodes shows the $base_{10}$ logarithm of the observed *E. coli* concentration. The solid line in the box and whiskers show the median while the bottom and top boarders of the box display the 1st (Q1) and third (Q3) quartiles also known as the interquartile range (IQR), respectively. Outliers (< Q1-1.5*IQR or > Q3+1.5*IQR) of the concentration distribution are shown as dots. Specific conductance is abbreviated as SPC (µS cm^{-1}); turbidity is abbreviated as NTU (NTU), dissolved oxygen is abbreviated by DO (mg L^{-1}), and temperature is abbreviated by (C;°C). The value *n* next to the node number indicates the number of observations in that split.



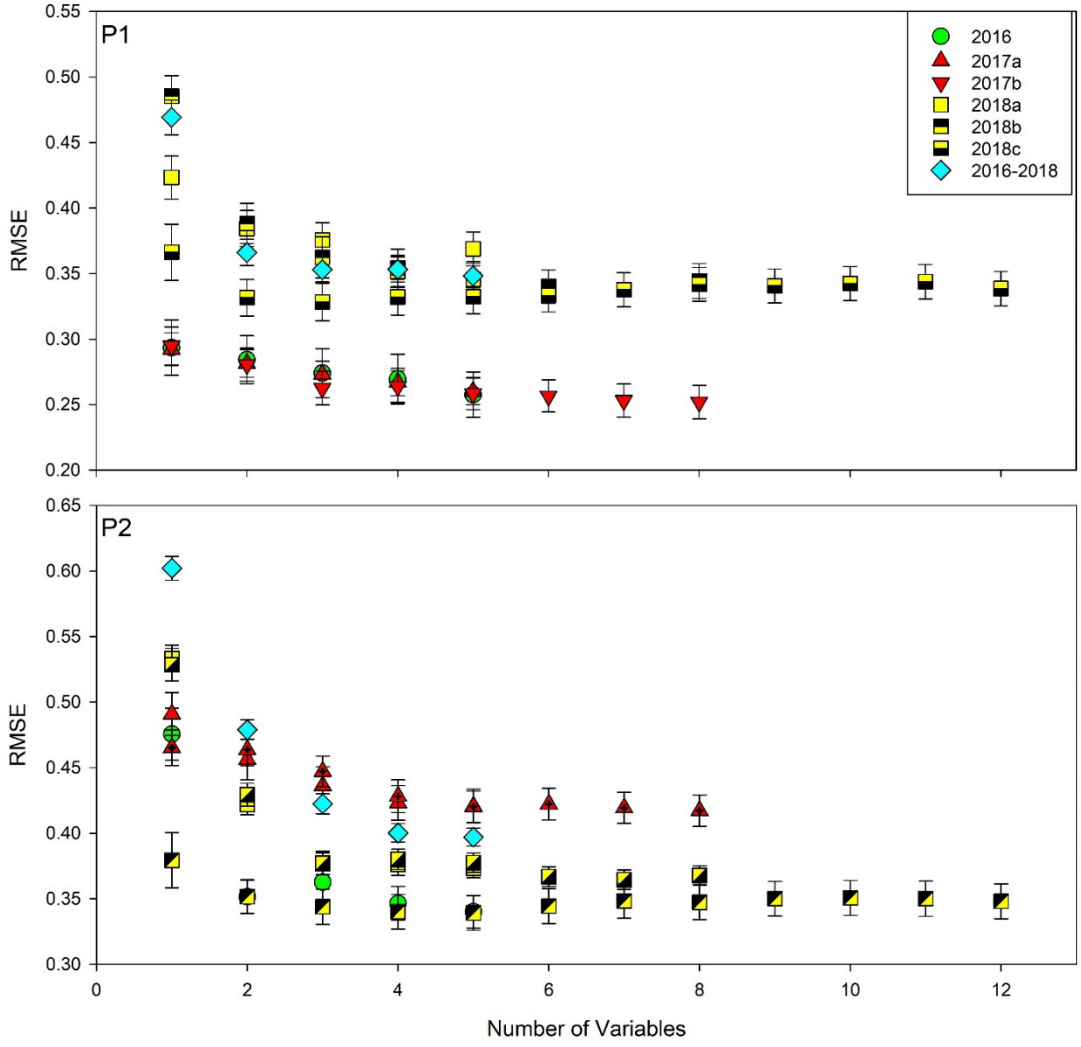
Supplementary Figure S11. Site maps for the two studied ponds. P1 had 23 sampling locations and is pictured on the left. P2 had 34 sampling locations and is pictured on the right



Supplementary Figure S12. Probabilities of MAE between the RF and other ML models. Red symbols indicate statistically significant differences.



Supplementary Figure S13. Probabilities of R^2 between the RF and other ML models. Red symbols indicate statistically significant differences.



Supplementary Figure S14**.** RMSE reduction for each pond based on the results from the recursive feature elimination exercise.

Bibliography

1. 1 Abebe, E., Gugsa, G., & Ahmed, M. (2020). Review on major food-borne zoonotic bacterial pathogens. Journal of tropical medicine, 2020.
2. Abimbola, O. P., Mittelstet, A. R., Messer, T. L., Berry, E. D., Bartelt-Hunt, S. L., & Hansen, S. P. (2020). Predicting Escherichia coli loads in cascading dams with machine learning: An integration of hydrometeorology, animal density and grazing pattern. Science of the Total Environment, 722 doi:10.1016/j.scitotenv.2020.137894
3. Ackers M. L., Mahon B. E., Leahy E., Goode B., Damrow T., Hayes P. S. (1998). An outbreak of Escherichia coli O157:H7 infections associated with leaf lettuce consumption. J. Infect. Dis. 177 1588–1593.
4. Adalja, A., & Lichtenberg, E. (2018). Implementation challenges of the food safety modernization act: Evidence from a national survey of produce growers. Food Control, 89, 62-71.
5. Aiyedun, S. O., Onarinde, B. A., Swainson, M., & Dixon, R. A. (2021). Foodborne outbreaks of microbial infection from fresh produce in Europe and North America: a systematic review of data from this millennium. International Journal of Food Science & Technology, 56(5), 2215-2223.
6. Akinde, S. B., Sunday, A. A., Adeyemi, F. M., Fakayode, I. B., Oluwajide, O. O., Adebunmi, A. A., & Adebooye, C. O. (2016). Microbes in Irrigation Water and Fresh Vegetables: Potential Pathogenic Bacteria Assessment and Implications for Food Safety. Applied Biosafety, 21(2), 89-97.
7. Allende, A., & Monaghan, J. (2015). Irrigation water quality for leafy crops: a perspective of risks and potential solutions. International journal of environmental research and public health, 12(7), 7457-7477.
8. Allende, A., Castro-Ibáñez, I., Lindqvist, R., Gil, M. I., Uyttendaele, M., & Jacxsens, L. (2017). Quantitative contamination assessment of Escherichia coli in baby spinach primary production in Spain: effects of weather conditions and agricultural practices. International journal of food microbiology, 257, 238-246.
9. Allende, A., Datta, A. R., Smith, W. A., Adonis, R., MacKay, A., & Adell, A. D. (2018). Implications of new legislation (US FSMA) and guidelines (EC) on the establishment of management systems for agricultural water. Food microbiology, 75, 119-125.
10. Allende, A., Truchado, P., Lindqvist, R., & Jacxsens, L. (2018). Quantitative microbial exposure modelling as a tool to evaluate the impact of contamination level of surface irrigation water and seasonality on fecal hygiene indicator *E. coli* in leafy green production. Food microbiology, 75, 82-89.

11. Álvarez-Cabria, M., Barquín, J., & Peñas, F. J. (2016). Modelling the spatial and seasonal variability of water quality for entire river networks: Relationships with natural and anthropogenic factors. Science of the Total Environment, 545, 152-162.
12. An, Y., Kampbell, D., Breidenbach, G.P. (2002). Escherichia coli and total coliforms in water and sediments at lake marinas. Environmental Pollution, 120, 771-778.
13. Ansa, E. D. O., Lubberding, H. J., Ampofo, J. A., & Gijzen, H. J. (2011). The role of algae in the removal of Escherichia coli in a tropical eutrophic lake. Ecological Engineering, 37(2), 317-324.
14. Antaki, E. M., Vellidis, G., Harris, C., Aminabadi, P., Levy, K., & Jay-Russell, M. T. (2016). Low concentration of Salmonella enterica and generic Escherichia coli in farm ponds and irrigation distribution systems used for mixed produce production in Southern Georgia. Foodborne pathogens and disease, 13(10), 551-558.
15. Avery, L. M., Hill, P., Killham, K., & Jones, D. L. (2004). Escherichia coli O157 survival following the surface and sub-surface application of human pathogen contaminated organic waste to soil. Soil Biology and Biochemistry, 36(12), 2101-2103.
16. Avila, R., Horn, B., Moriarty, E., Hodson, R., & Moltchanova, E. (2018). Evaluating statistical model performance in water quality prediction. Journal of environmental management, 206, 910-919.
17. Babbar, R., & Babbar, S. (2017). Predicting river water quality index using data mining techniques. Environmental Earth Sciences, 76(14), 1-15.
18. Baker, M. E., Schley, M. L., & Sexton, J. O. (2019). Impacts of expanding impervious surface on specific conductance in urbanizing streams. Water Resources Research, 55(8), 6482-6498.
19. Ballesté, E., Belanche-Muñoz, L. A., Farnleitner, A. H., Linke, R., Sommer, R., Santos, R., Blanch, A. R. (2020). Improving the identification of the source of faecal pollution in water using a modelling approach: From multi-source to aged and diluted samples. Water Research, 171 doi:10.1016/j.watres.2019.115392
20. Banach, J. L., & Van der Fels-Klerx, H. J. (2020). Microbiological reduction strategies of irrigation water for fresh produce. Journal of food protection, 83(6), 1072-1087.
21. Barcina, I., Arana, I., Iriberri, J., & Egea, L. (1986). Factors affecting the survival of E. coli in a river. Hydrobiologia, 141(3), 249-253.
22. Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine learning, 36(1), 105-139.

23. Belias, A., Brassill, N., Roof, S., Rock, C., Wiedmann, M., & Weller, D. (2021). Cross-validation indicates predictive models may provide an alternative to indicator organism monitoring for evaluating pathogen presence in southwestern US agricultural water. Frontiers in Water, 87.
24. Benjamin, L. A., Jay-Russell, M. T., Atwill, E. R., Cooley, M. B., Carychao, D., Larsen, R. E., & Mandrell, R. E. (2015). Risk factors for Escherichia coli O157 on beef cattle ranches located near a major produce production region. Epidemiology & Infection, 143(1), 81-93.
25. Benjamin, L., Atwill, E. R., Jay-Russell, M., Cooley, M., Carychao, D., Gorski, L., & Mandrell, R. E. (2013). Occurrence of generic Escherichia coli, *E. coli* O157 and Salmonella spp. in water and sediment from leafy green produce farms and streams on the Central California coast. International Journal of Food Microbiology, 165(1), 65-76.
26. Bennett, S. D., Sodha, S. V., Ayers, T. L., Lynch, M. F., Gould, L. H., & Tauxe, R. V. (2018). Produce-associated foodborne disease outbreaks, USA, 1998–2013. Epidemiology & Infection, 146(11), 1397-1406.
27. Berry, E. D., & Foegeding, P. M. (1997). Cold temperature adaptation and growth of microorganisms. Journal of Food Protection, 60(12), 1583-1594.
28. Bien, J., Sokolova, O., & Bozko, P. (2012). Role of uropathogenic Escherichia coli virulence factors in development of urinary tract infection and kidney damage. International journal of nephrology, 2012.
29. Bintsis, T. (2017). Foodborne pathogens. AIMS microbiology, 3(3), 529.
30. Bitton, G. (2005). Microbial indicators of fecal contamination: application to microbial source tracking. Report submitted to the Florida Stormwater Association, 719, 4-6.
31. Björnsson, H. and Venegas, S.A. (1997). "A manual for EOF and SVD analyses of climate data", McGill University, CCGCR Report No. 97-1, Montréal, Québec, 52pp.
32. Blaustein, R. A., Pachepsky, Y. A., Shelton, D. R., & Hill, R. L. (2015). Release and removal of microorganisms from land-deposited animal waste and animal manures: a review of data and models. Journal of environmental quality, 44(5), 1338-1354.
33. Blaustein, R. A., Pachepsky, Y., Hill, R. L., Shelton, D. R., & Whelan, G. (2013). Escherichia coli survival in waters: temperature dependence. Water research, 47(2), 569-578.
34. Boehm, A. B., & Sassoubre, L. M. (2014). Enterococci as indicators of environmental fecal contamination. In Enterococci: From Commensals to Leading Causes of Drug Resistant Infection [Internet]. Massachusetts Eye and Ear Infirmary.
35. Boehm, A. B., Grant, S. B., Kim, J. H., Mowbray, S. L., McGee, C. D., Clark, C. D., & Wellman, D. E. (2002). Decadal and shorter period variability of surf

zone water quality at Huntington Beach, California. Environmental science & technology, 36(18), 3885-3892.

36. Boehm, A. B., Yamahara, K. M., Love, D. C., Peterson, B. M., McNeill, K., & Nelson, K. L. (2009). Covariation and photoinactivation of traditional and novel indicator organisms and human viruses at a sewage-impacted marine beach. Environmental science & technology, 43(21), 8046-8052.
37. Bolton, N. F., Cromar, N. J., Buchanan, N. A., & Fallowfield, H. J. (2011). Variations in sunlight attenuation in waste stabilisation ponds and environmental waters. In Proceedings of the 9th IWA Specialist Group Conference on Waste Stabilisation Ponds (pp. 1-3).
38. Bosshard, F., Bucheli, M., Meur, Y., & Egli, T. (2010). The respiratory chain is the cell's Achilles' heel during UVA inactivation in Escherichia coli. Microbiology, 156(7), 2006-2015.
39. Bouckaert, R. R., & Frank, E. (2004, May). Evaluating the replicability of significance tests for comparing learning algorithms. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 3-12). Springer, Berlin, Heidelberg.
40. Bouteleux, C., Saby, S., Tozza, D., Cavard, J., Lahoussine, V., Hartemann, P., & Mathieu, L. (2005). Escherichia coli behavior in the presence of organic matter released by algae exposed to water treatment chemicals. Applied and environmental microbiology, 71(2), 734-740.
41. Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, Ø., Frigessi, A., & Lingjærde, O. C. (2007). Predicting survival from microarray data - A comparative study. Bioinformatics, 23(16), 2080-2087. doi:10.1093/bioinformatics/btm305
42. Bradshaw, J. K., Snyder, B. J., Oladeinde, A., Spidle, D., Berrang, M. E., Meinersmann, R. J., ... & Molina, M. (2016). Characterizing relationships among fecal indicator bacteria, microbial source tracking markers, and associated waterborne pathogen occurrence in stream water and sediments in a mixed land use watershed. Water research, 101, 498-509.
43. Brennan, F. P., O'Flaherty, V., Kramers, G., Grant, J., & Richards, K. G. (2010). Long-term persistence and leaching of Escherichia coli in temperate maritime soils. Applied and environmental microbiology, 76(5), 1449-1455.
44. Brissaud, F., Tournoud, M. G., Drakides, C., & Lazarova, V. (2002). Mixing and its impact on faecal coliform removal in a stabilisation pond. Water Science and Technology, 48, 75-80.
45. Brookes, J.D., Antenucci, J., Hipsey, M., Burch, M.D., Ashbolt, N.J., Ferguson, C. (2004). Fate and transport of pathogens in lakes and reservoirs. Environment International, 30, 741–759. https://doi.org/10.1016/j.envint.2003.11.006

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

47. Burke, T.C. (2018). Investigating the Influence of Abiotic Variability on Escherichia coli Population Dynamics in Three Denver, CO Springs. Regis University Student Publications. 854. https://epublications.regis.edu/theses/854

48. Burkov, A. (2020). Machine Learning Engineering. True Positive Incorporated. http://themlbook.com/ . Accessed January 2022.

49. Buyrukoğlu, G., Buyrukoğlu, S., & Topalcengiz, Z. (2021). Comparing Regression Models with Count Data to Artificial Neural Network and Ensemble Models for Prediction of Generic Escherichia coli Population in Agricultural Ponds Based on Weather Station Measurements. Microbial Risk Analysis, 100171.

50. Carstens, C. K., Salazar, J. K., & Darkoh, C. (2019). Multistate outbreaks of foodborne illness in the United States associated with fresh produce from 2010 to 2017. Frontiers in microbiology, 10, 2667.

51. Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168).

52. CDC, 2018. Estimates of Foodborne Illness in the United States. Centers for Disease Control and Prevention. U.S. Department of Health and Human Services. https://www.cdc.gov/foodborneburden/index.html . Accessed January 2022.

53. CDC, 2021. CDC and Food Safety. Centers for Disease Control and Prevention. U.S. Department of Health and Human Services. https://www.cdc.gov/foodsafety/pdfs/cdc-and-food-safety.pdf. Accessed January 2022.

54. Ceuppens, S., Johannessen, G. S., Allende, A., Tondo, E. C., El-Tahan, F., Sampers, I., ... & Uyttendaele, M. (2015). Risk factors for Salmonella, shiga toxin-producing Escherichia coli and Campylobacter occurrence in primary production of leafy greens and strawberries. International Journal of Environmental Research and Public Health, 12(8), 9809-9831.

55. Chandrasekaran, R., Hamilton, M. J., Wang, P., Staley, C., Matteson, S., Birr, A., & Sadowsky, M. J. (2015). Geographic isolation of Escherichia coli genotypes in sediments and water of the Seven Mile Creek—a constructed riverine watershed. Science of the Total Environment, 538, 78-85.

56. Chen, H. J., & Chang, H. (2014). Response of discharge, TSS, and *E. coli* to rainfall events in urban, suburban, and rural watersheds. Environmental Science: Processes & Impacts, 16(10), 2313-2324.

57. Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zou, X., Wang, J., Zhang, Y., Chen, D., Chen, X., Deng, Y., & Ren, H. (2020).

Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water research, 171, 115454.

58. Cheng, J., Niu, S., & Kim, Y. (2013). Relationship between water quality parameters and the survival of indicator microorganisms–Escherichia coli–in a stormwater wetland. Water Science and Technology, 68(7), 1650-1656.

59. Cho, K. H., Pachepsky, Y. A., Kim, J. H., Guber, A. K., Shelton, D. R., & Rowland, R. (2010). Release of Escherichia coli from the bottom sediment in a first-order creek: Experiment and reach-specific modeling. Journal of Hydrology, 391(3-4), 322-332.https://www.cdc.gov/foodsafety/outbreaks/index.html. Accessed January 2022.

60. Cho, K. H., Pachepsky, Y. A., Oliver, D. M., Muirhead, R. W., Park, Y., Quilliam, R. S., & Shelton, D. R. (2016). Modeling fate and transport of fecally-derived microorganisms at the watershed scale: state of the science and future opportunities. Water research, 100, 38-56.

61. Choi, C., Kim, J., Kim, J., Kim, D., Bae, Y., & Kim, H. S. (2018). Development of heavy rain damage prediction model using machine learning based on big data. Advances in meteorology, 2018.

62. Chopyk, J., Allard, S., Nasko, D. J., Bui, A., Mongodin, E. F., & Sapkota, A. R. (2018). Agricultural freshwater pond supports diverse and dynamic bacterial and viral populations. Frontiers in microbiology, 9, 792.

63. Chudoba, Elizabeth A., Michael A. Mallin, Lawrence B. Cahoon, and Stephen A. Skrabal. (2013). Stimulation of Fecal Bacteria in Ambient Waters by Experimental Inputs of Organic and Inorganic Phosphorus. Water Research, 47, no. 10, 3455–66. https://doi.org/10.1016/j.watres.2013.03.047

64. Colombara, D. V., Khalil, I. A. M., Rao, P. C., Troeger, C., Forouzanfar, M. H., Riddle, M. S., & Mokdad, A. H. (2016). Chronic health consequences of acute enteric infections in the developing world. The American Journal of Gastroenterology Supplements, 3(2), 4.

65. Conboy, M. J., & Goss, M. J. (2000). Natural protection of groundwater against bacteria of fecal origin. Journal of contaminant hydrology, 43(1), 1-24.

66. Cooley, M. B., Quiñones, B., Oryang, D., Mandrell, R. E., & Gorski, L. (2014). Prevalence of shiga toxin producing Escherichia coli, Salmonella enterica, and Listeria monocytogenes at public access watershed sites in a California Central Coast agricultural region. Frontiers in cellular and infection microbiology, 4, 30.

67. Cooley, M., Carychao, D., Crawford-Miksza, L., Jay, M. T., Myers, C., Rose, C., ... & Mandrell, R. E. (2007). Incidence and tracking of Escherichia coli O157: H7 in a major produce production region in California. PloS one, 2(11), e1159.

68. Cosh, M.H., Jackson, T.J., Moran, S., Bindlish, R. (2008). Temporal persistence and stability of surface soil moisture in a semi-arid watershed. Remote Sensing of Environment, 112, 304–313. https://doi.org/10.1016/j.rse.2007.07.001

69. Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

70. Curtis, T. P., Mara, D. D., & Silva, S. A. (1992). Influence of pH, oxygen, and humic substances on ability of sunlight to damage fecal coliforms in waste stabilization pond water. Applied and environmental microbiology, 58(4), 1335-1343.

71. Danades, A., Pratama, D., Anggraini, D., & Anggriani, D. (2016, October). Comparison of accuracy level K-nearest neighbor algorithm and support vector machine algorithm in classification water quality status. In 2016 6th International Conference on System Engineering and Technology (ICSET) (pp. 137-141). IEEE.

72. Davies-Colley, R. J., Donnison, A. M., & Speed, D. J. (1997). Sunlight wavelengths inactivating faecal indicator microorganisms in waste stabilisation ponds. Water Science and Technology, 35(11-12), 219-225.

73. Davies-Colley, R. J., Donnison, A. M., Speed, D. J., Ross, C. M., & Nagels, J. (1999). Inactivation of faecal indicator micro-organisms in waste stabilisation ponds: interactions of environmental factors with sunlight. Water research, 33(5), 1220-1230.

74. Davies-Colley, R., Valois, A., & Milne, J. (2018). Faecal contamination and visual clarity in New Zealand rivers: correlation of key variables affecting swimming suitability. Journal of water and health, 16(3), 329-339.

75. Davies-Colley, R.J., Donnison, A.M., Speed, D.J. (2000). Towards a mechanistic understanding of pond disinfection. Water Science and Technology, 42, 149–158. https://doi.org/10.2166/wst.2000.0630

76. Davis, K., Anderson, M., Yates, M.V. (2005). Distribution of indicator bacteria in Canyon Lake, California. Water Research, 39, 1277-1288.

77. Delbeke, S., Ceuppens, S., Hessel, C. T., Castro, I., Jacxsens, L., De Zutter, L., & Uyttendaele, M. (2015). Microbial safety and sanitary quality of strawberry primary production in Belgium: risk factors for Salmonella and Shiga toxin producing Escherichia coli (STEC) contamination. Applied and environmental microbiology, AEM-03930.

78. Delmas, J., Dalmasso, G., & Bonnet, R. (2015). Escherichia coli: The good, the bad and the ugly. Clinical Microbiology: Open Access.

79. Derlet, R. W., Richards, J. R., Tanaka, L. L., Hayden, C., Ger, K. A., & Goldman, C. R. (2012). Impact of summer cattle grazing on the Sierra Nevada watershed: aquatic algae and bacteria. Journal of Environmental and Public Health, 2012.

80. Desai, A. M., & Rifai, H. S. (2013). Escherichia coli concentrations in urban watersheds exhibit diurnal sag: Implications for water-quality monitoring and assessment. Journal of the American Water Resources Association, 49(4), 766-779.

81. DeVilbiss, S. E., Steele, M. K., Krometis, L. A. H., & Badgley, B. D. (2021). Freshwater salinization increases survival of Escherichia coli and risk of bacterial impairment. Water Research, 191, 116812.

82. Dias, D. F. C., Passos, R. G., & Von Sperling, M. (2017). A review of bacterial indicator disinfection mechanisms in waste stabilisation ponds. Reviews in Environmental Science and Bio/Technology, 16(3), 517-539.

83. Dias, D. F., & von Sperling, M. (2017). Solar radiation (PAR, UV-A, UV-B) penetration in a shallow maturation pond operating in a tropical climate. Water Science and Technology, 76(1), 182-191.

84. Dias, D. F., & von Sperling, M. (2018). Vertical profiling and modelling of Escherichia coli decay in a shallow maturation pond operating in a tropical climate. Environmental technology, 39(6), 759-769.

85. Dieter, C.A., Maupin, M.A., Caldwell, R.R., Harris, M.A., Ivahnenko, T.I., Lovelace, J.K., Barber, N.L., and Linsey, K.S. (2018). Estimated use of water in the United States in 2015. U.S. Geological Survey Circular 1441, 65 p., https://doi.org/10.3133/cir1441

86. Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation, 10(7), 1895-1923.

87. Draper, A. D., Doores, S., Gourama, H., & LaBORDE, L. F. (2016). Microbial survey of Pennsylvania surface water used for irrigating produce crops. Journal of food protection, 79(6), 902-912.

88. Durham, B.W., Porter, L., Webb, A., Thomas, J. (2016). Seasonal influence of environmental variables and artificial aeration on Escherichia coli in small urban lakes. Journal of Water and Health, 14, 929–941. https://doi.org/10.2166/wh.2016.020

89. Duris JW, Reif AG, Krouse DA, Isaacs NM. (2013). Factors related to occurrence and distribution of selected bacterial and protozoan pathogens in Pennsylvania streams. Water Res 47:300–14.

90. Eblen, B. S., Walderhaug, M. O., Edelson-Mammel, S., Chirtel, S. J., De Jesus, A., Merker, R. I., & Miller, A. J. (2004). Potential for internalization, growth, and survival of Salmonella and Escherichia coli O157: H7 in oranges. Journal of food protection, 67(8), 1578-1584.

91. Englebert, E. T., McDermott, C., & Kleinheinz, G. T. (2008). Impact of the alga Cladophora on the survival of *E. coli*, Salmonella, and Shigella in laboratory microcosm. Journal of Great Lakes Research, 34(2), 377-382.

92. Entry, J. A., Leytem, A. B., & Verwey, S. (2005). Influence of solid dairy manure and compost with and without alum on survival of indicator bacteria in soil and on potato. Environmental Pollution, 138(2), 212-218.

93. Fallowfield, H. J., N. J. Cromar, and L. M. Evison. (1996). Coliform Die-off Rate Constants in a High Rate Algal Pond and the Effect of Operational and Environmental Variables. Water Science and Technology, 34 (11), 141–47. https://doi.org/10.1016/S0273-1223(96)00831-1.

94. Faust, M. A., Aotaky, A. E., & Hargadon, M. T. (1975). Effect of physical parameters on the in situ survival of Escherichia coli MC-6 in an estuarine environment. Applied microbiology, 30(5), 800-806.

95. Food and Drug Administration (FDA). (1998). Guidance for Industry: Guide to Minimize Microbial Food Safety Hazards for Fresh Fruits and Vegetables. Food and Drug Administration. Federal Register 63 FR 58055.

96. Food and Drug Administration (FDA). (2008). Guidance for Industry: Guide to Minimize Microbial Food Safety Hazards of Fresh-cut Fruits and Vegetables. Fed. Regist. 73 FR 10037.

97. Food and Drug Administration (FDA). (2009). Guidance for industry: guide to minimize microbial food safety hazards of melons. http://www.fda.gov/Food/GuidanceComplianceRegulatoryInformation/GuidanceDocuments/ProduceandPlanProducts/ucm174171.htm, Accessed January 2022.

98. Food and Drug Administration (FDA). (2015). Standards for the growing, harvesting, packing, and holding of produce for human consumption; final rule. Fed. Regist. 80:74354–74568.

99. Food and Drug Administration (FDA). (2016). FSMA Final Rule on Produce Safety: Standards for the Growing, Harvesting, Packing, and Holding of Produce for Human Consumption. Accessed from: http://www.fda.gov/Food/GuidanceRegulation/FSMA/ucm334114.htm.

100. Food and Drug Administration (FDA). (2017). FSMA Final Rule on Produce Safety. Standards for the Growing, Harvesting, Packing, and Holding of Produce for Human Consumption. https://www.fda.gov/food/food-safety-modernization-act-fsma/fsma-final-rule-produce-safety. Accessed January 2022.

101. Food and Drug Administration (FDA). (2021). FSMA Compliance Dates. Food and Drug Administration. https://www.fda.gov/food/food-safety-modernization-act-fsma/fsma-compliance-dates. Accessed January 2022.

102. Food and Drug Administration (FDA). 2013. Standards for the growing, harvesting, packing, and holding of produce for human consumption; proposed rule; correction. Fed. Regist. 78:3503–3646.

103. Food and Drug Adminsitration (FDA). 2018. https://www.fda.gov/food/foodborneillnesscontaminants/peopleatrisk/ucm352830.htm.

104. Food and Drug Adminsitration (FDA). 2019. FSMA Final Rule on Produce Safety. Available at https://www.fda.gov/food/food-safety-modernization-act-fsma/fsma-final-rule-produce-safety. Accessed on 2/21/2021

105. Francy, D. S., Brady, A. M., Cicale, J., Dalby, H. D., & Stelzer, E. (2020). Nowcasting methods for determining microbiological water quality at recreational beaches and drinking-water source waters. Journal of Microbiological Methods, 175.

106. Francy, D. S., Stelzer, E. A., Duris, J. W., Brady, A. M. G., Harrison, J. H., Johnson, H. E., et al. (2013). Predictive models for Escherichia coli concentrations at inland lake beaches and relationship of model variables to pathogen detection. Appl. Environ. Microbiol. 79, 1676–1688. doi: 10.1128/AEM.02995-12

107. Franz, C. M. A. P., den Besten, H. M., Boehnlein, C., Gareis, M., Zwietering, M. H., & Fusco, V. (2018). Microbial food safety in the 21st century: Emerging challenges and foodborne pathogenic bacteria. Trends in Food Science & Technology, Keeping Food Safety on the Agenda for, 15, 34-37.

108. Freeman, E. A., Moisen, G. G., Coulston, J. W., & Wilson, B. T. (2016). Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. Canadian Journal of Forest Research, 46(3), 323-339.

109. Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367-378.

110. García-Alba, J., Bárcena, J. F., Ugarteburu, C., & García, A. (2019). Artificial neural networks as emulators of process-based models to analyse bathing water quality in estuaries. Water research, 150, 283-295.

111. Garcia-Armisen, T., & Servais, P. (2009). Partitioning and Fate of Particle-Associated *E. coli* in River Waters. Water Environment Research, 81(1), 21-28.

112. Ge, Z., Whitman, R. L., Nevers, M. B., Phanikumar, M. S., & Byappanahalli, M. N. (2012). Nearshore hydrodynamics as loading and forcing factors for Escherichia coli contamination at an embayed beach. Limnology and Oceanography, 57(1), 362-381.

113. Gelting, R. J., Baloch, M. A., Zarate-Bermudez, M., Hajmeer, M. N., Yee, J. C., Brown, T., & Yee, B. J. (2015). A systems analysis of irrigation water quality in an environmental assessment of an *E. coli* O157: H7 outbreak in the United States linked to iceberg lettuce. Agricultural Water Management, 150, 111-118.

114. Gerba, C. P., & Smith, J. E. (2005). Sources of pathogenic microorganisms and their fate during land application of wastes. Journal of environmental quality, 34(1), 42-48.

115. Ginn, T. R., Wood, B. D., Nelson, K. E., Scheibe, T. D., Murphy, E. M., & Clement, T. P. (2002). Processes in microbial transport in the natural subsurface. Advances in Water Resources, 25(8-12), 1017-1042.

116. Givens, C. E., Kolpin, D. W., Borchardt, M. A., Duris, J. W., Moorman, T. B., & Spencer, S. K. (2016). Detection of hepatitis E virus and other livestock-related pathogens in Iowa streams. Science of the Total Environment, 566, 1042-1051.

117. Goldberg D, Kroupitski Y, Belausov E, Pinto R, Sela S. (2011). Salmonella Typhimurium internalization is variable in leafy vegetables and fresh herbs. Intl J Food Microbiol. 2011;145:250–7.

118. Goodburn, C., & Wallace, C. A. (2013). The microbiological efficacy of decontamination methodologies for fresh produce: a review. Food Control, 32(2), 418-427.

119. Gorski, L., Parker, C. T., Liang, A., Cooley, M. B., Jay-Russell, M. T., Gordus, A. G., & Mandrell, R. E. (2011). Prevalence, distribution, and diversity of Salmonella enterica in a major produce region of California. Applied and Environmental Microbiology, 77(8), 2734-2748.

120. Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometrics and intelligent laboratory systems, 83(2), 83-90.

121. Grant, S.B., Litton Mueller, R.M., and Ahn, J.H.. 2011. Measuring and modeling the flux of fecal bacteria across the sediment water interface in a turbulent stream. Water Resour. Res. 47 (5): W05517.

122. Grbčić, L., Družeta, S., Mauša, G., Lipić, T., Lušić, D. V., Alvir, M., & Kranjčević, L. (2021). Coastal water quality prediction based on machine learning with feature interpretation and spatio-temporal analysis. arXiv preprint arXiv:2107.03230.

123. Greene, S. K., E. R. Daly, E. A. Talbot, L. J. Demma, S. Holzbauer, N. J. Patel, T. A. Hill. (2008). Recurrent Multistate Outbreak of Salmonella Newport Associated with Tomatoes from Contaminated Fields, 2005. Epidemiology & Infection 136 (2), 157–65. https://doi.org/10.1017/S095026880700859X.

124. Greenwell, B., Boehmke, B., Cunningham, J., and Developers, G.B.M. (2020). gbm: Generalized Boosted Regression Models. R package version 2.1.8. https://CRAN.R-project.org/package=gbm

125. Gu, G., Luo, Z., Cevallos-Cevallos, J. M., Adams, P., Vellidis, G., Wright, A., & van Bruggen, A. H. (2012). Factors affecting the occurrence of Escherichia coli O157 contamination in irrigation ponds on produce farms in the Suwannee River Watershed. Canadian journal of microbiology, 59(3), 175-182.

126. Gu, G., Strawn, L. K., Ottesen, A. R., Ramachandran, P., Reed, E. A., Zheng, J., ... & Rideout, S. L. (2021). Correlation of Salmonella enterica and Listeria monocytogenes in irrigation water to environmental factors, fecal indicators, and bacterial communities. *Frontiers in microbiology*, 3426.

127. Guan, H., Dong, X., Yan, G., Searls, T., Bourque, C. P. A., & Meng, F. R. (2021). Conditional inference trees in the assessment of tree mortality rates in the transitional mixed forests of Atlantic Canada. PloS one, 16(6), e0250991.

128. Guo X, van Iersel MW, Chen J, Brackett RE, and Beuchat LR. 2002. Evidence of association of salmonellae with tomato plants grown hydroponically in inoculated nutrient solution. Appl. Environ. Microbiol. 68:3639–3643.

129. Gutierrez-Rodriguez, E., & Adhikari, A. (2018). Preharvest farming practices impacting fresh produce safety. Preharvest Food Safety, 19-46.

130. Hamid Babamoradi, Frans van den Berg, Åsmund Rinnan, Bootstrap based confidence limits in principal component analysis — A case study, Chemometrics and Intelligent Laboratory Systems, Volume 120, 15 January 2013, Pages 97-105

131. Hammer, D.A.T., Ryan, P.D., Hammer, Ø., Harper, D.A.T. (2001). Past: Paleontological Statistics Software Package for Education and Data Analysis, Palaeontologia Electronica.

132. Hamrick, J. M. (1992). A three-dimensional environmental fluid dynamics computer code: Theoretical and computational aspects.

133. Han, Lu, Kaidi Wang, Lina Ma, Pascal Delaquis, Susan Bach, Jinsong Feng, and Xiaonan Lu. (2020). Viable but Nonculturable Escherichia coli O157:H7 and Salmonella Enterica in Fresh Produce: Rapid Determination by Loop-Mediated Isothermal Amplification Coupled with a Propidium Monoazide Treatment. Applied and Environmental Microbiology 86, no. 7 (n.d.): e02566-19. https://doi.org/10.1128/AEM.02566-19.

134. Harris, C. S., Tertuliano, M., Rajeev, S., Vellidis, G., & Levy, K. (2018). Impact of storm runoff on Salmonella and Escherichia coli prevalence in irrigation ponds of fresh produce farms in southern Georgia. Journal of Applied Microbiology, 124(3), 910-921.

135. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction 2nd Edition. NY:Springer. doi:10.1007/b94608.

136. Hatvani, I. G., Kirschner, A. K., Farnleitner, A. H., Tanos, P., & Herzig, A. (2018). Hotspots and main drivers of fecal pollution in Neusiedler See, a large shallow lake in Central Europe. Environmental Science and Pollution Research, 25(29), 28884-28898.

137. Havelaar, A. H., Vazquez, K. M., Topalcengiz, Z., Muñoz-Carpena, R., & Danyluk, M. D. (2017). Evaluating the US Food Safety Modernization Act

Produce Safety Rule standard for microbial quality of agricultural water for growing produce. Journal of food protection, 80(11), 1832-1841.

138. He, G., Fang, H., Bai, S., Liu, X., Chen, M., & Bai, J. (2011). Application of a three-dimensional eutrophication model for the Beijing Guanting Reservoir, China. Ecological Modelling, 222(8), 1491-1501.

139. He, J., Vale, C., & Chu, A. (2015). Variation in Water Quality of a Stormwater Pond from Diurnal Thermal Stratification. Journal of Water Resource and Hydraulic Engineering, 4(2), 181–190. https://doi.org/10.5963/jwrhe0402008

140. Hintz LD, Boyer RR, Ponder MA, Williams RC, and Rideout SL. (2010). Recovery of Salmonella enterica Newport introduced through irrigation water from tomato (Lycopersicum esculentum) fruit, roots, stems, and leaves. HortScience 45:675–678.

141. Hoagland, L., Ximenes, E., Ku, S., & Ladisch, M. (2018). Foodborne pathogens in horticultural production systems: Ecology and mitigation. Scientia Horticulturae, 236, 192-206.

142. Hoffman, S., Maculloch, B., & Batz, M. (2015). Economic burden of major foodborne illnesses acquired in the United States (No. 1476-2016-120935).

143. Hoffmann, S., & Ahn, J. W. (2021). Updating Economic Burden of Foodborne Diseases Estimates for Inflation and Income Growth (No. 1962-2021-3214).

144. Hosetti, B. B., and Patil, H. S. Performance of wastewater stabilization ponds at different depths. Water, Air, and Soil Pollution 34: 191–198. 1987.

145. Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. The Journal of Machine Learning Research, 16(1), 3905-3909.

146. Huang, Z., Miao, H.-T., Liu, Y., Tian, F.-P., He, H., Shen, W., López-Vicente, M., Wu, G.-L. (2018). Soil water content and temporal stability in an arid area with natural and planted grasslands. Hydrological Processes, 32, 3784–3792. https://doi.org/10.1002/hyp.13289

147. Hughes, K. A. (2003). Influence of seasonal environmental variables on the distribution of presumptive fecal coliforms around an Antarctic research station. Applied and Environmental Microbiology, 69(8), 4884-4891.

148. Huot, Y., Jeffrey, W. H., Davis, R. F., & Cullen, J. J. (2000). Damage to DNA in Bacterioplankton: A Model of Damage by Ultraviolet Radiation and its Repair as Influenced by Vertical Mixing¶. Photochemistry and Photobiology, 72(1), 62-74.

149. Hussain, M. A., & Dawson, C. O. (2013). Economic impact of food safety outbreaks on food businesses. Foods, 2(4), 585-589.

150. Hutchison, M. L., Walters, L. D., Moore, A., Crookes, K. M., & Avery, S. M. (2004). Effect of length of time before incorporation on survival of pathogenic

bacteria present in livestock wastes applied to agricultural soil. Applied and environmental microbiology, 70(9), 5111-5118.

151. Ishii, S., & Sadowsky, M. J. (2008). Escherichia coli in the environment: implications for water quality and human health. Microbes and environments, 23(2), 101-108.

152. Islam, M., Doyle, M.P., Phatak, S.C., Millner, P., Jiang, X.P. (2004a.) Persistence of enterohemoorhagic Escherichia coli O157:H7 in soil and on leaf lettuce and parsley grown in fields treated with contaminated manure composts or irrigation water. Journal of Food Protection, 67(2004), 1365-1370.

153. Islam, M., Morgan, J., Doyle, M.P., Phatak, S.C., Millner, P., Jiang, X.P. (2004b). Persistence of Salmonella enterocia serovar Typhimuririum on lettuce and parsley and in soils on which they were gornw in fields treated with contaminated manure composts or irrigation water. Foodborne Pathogen Diseases, 1, 27-35.

154. Ismail, N. S., Dodd, H., Sassoubre, L. M., Horne, A. J., Boehm, A. B., & Luthy, R. G. (2015). Improvement of urban lake water quality by removal of Escherichia coli through the action of the bivalve Anodonta californiensis. Environmental science & technology, 49(3), 1664-1672.

155. Jamieson, R. C., Joy, D. M., Lee, H., Kostaschuk, R., & Gordon, R. J. (2005). Resuspension of sediment-associated Escherichia coli in a natural stream. Journal of Environmental Quality, 34(2), 581-589.

156. Jenkins, M. B., Adams, M. P., Endale, D. M., Fisher, D. S., Lowrance, R., Newton, G. L., & Vellidis, G. (2015). Storm flow dynamics and loads of fecal bacteria associated with ponds in southern piedmont and coastal plain watersheds with animal agriculture. Agricultural Water Management, 148, 97–105. https://doi.org/10.1016/j.agwat.2014.09.022

157. Jenkins, M. B., Fisher, D. S., Endale, D. M., & Adams, P. (2011). Comparative die-off of Escherichia coli 0157: H7 and fecal indicator bacteria in pond water. Environmental science & technology, 45(5), 1853-1858.

158. Jenkins, Michael B., Dwight S. Fisher, Dinku M. Endale, and Paige Adams. (2011). Comparative Die-off of Escherichia coli 0157:H7 and Fecal Indicator Bacteria in Pond Water. Environmental Science & Technology 45 (5), 1853–58. https://doi.org/10.1021/es1032019.

159. Jeon, D. J., Pachepsky, Y., Coppock, C., Harriger, M. D., Zhu, R., & Wells, E. (2020). Temporal stability of *E. coli* and Enterococci concentrations in a Pennsylvania creek. Environmental Science and Pollution Research, 27(4), 4021-4031.

160. Jeon, D. J., Pachepsky, Y., Harriger, D., Coppock, C, Wells, E., Hong, E. M. 2019. Analysis of Escherichia coli and enterococci concentrations patterns in a Pennsylvania creek using empirical orthogonal functions. Journal of Environmental Quality 48(6) 48(6), 1703-1710

161. Jokinen, C.C., Hillman, E., Tymensen, L. (2019). Sources of generic Escherichia coli and factors impacting guideline exceedances for food safety in an irrigation reservoir outlet and two canals. Water Research, 156, 148–158. https://doi.org/10.1016/j.watres.2019.03.013

162. Jones, L. A., Worobo, R. W., & Smart, C. D. (2014). Plant-pathogenic oomycetes, Escherichia coli strains, and Salmonella spp. frequently found in surface water used for irrigation of fruit and vegetable crops in New York State. Appl. Environ. Microbiol., 80(16), 4814-4820.

163. Jongman, M., Korsten, L. (2016). Microbial quality and suitability of roof-harvested rainwater in rural villages for crop irrigation and domestic use. Journal of Water and Health, 14, 961–971. https://doi.org/10.2166/wh.2016.058

164. Karatzoglou, A., Smola, A., Hornik, K., and Karatzoglou, M. A. (2019). Package 'kernlab'. R package version 0.9-29. https://cran.r-project.org/web/packages/kernlab/index.html

165. Kataržytė, M., Mėžinė, J., Vaičiūtė, D., Liaugaudaitė, S., Mukauskaitė, K., Umgiesser, G., & Schernewski, G. (2018). Fecal contamination in shallow temperate estuarine lagoon: Source of the pollution and environmental factors. Marine pollution bulletin, 133, 762-772.

166. Kearney, J. (2010). Food consumption trends and drivers. Philosophical transactions of the royal society B: biological sciences, 365(1554), 2793-2807.

167. Khan, I. U., Gannon, V., Kent, R., Koning, W., Lapen, D. R., Miller, J., ... & Edge, T. A. (2007). Development of a rapid quantitative PCR assay for direct detection and quantification of culturable and non-culturable Escherichia coli from agriculture watersheds. Journal of microbiological methods, 69(3), 480-488

168. Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 41(3), 552-568.

169. Kim, G., Baek, I., Stocker, M. D., Smith, J. E., Van Tassell, A. L., Qin, J., ... & Kim, M. S. (2020). Hyperspectral Imaging from a Multipurpose Floating Platform to Estimate Chlorophyll-a Concentrations in Irrigation Pond Water. Remote Sensing, 12(13), 2070.

170. Kim, S., Daughtry, C., Russ, A., Pedrera-Parrilla, A., & Pachepsky, Y. (2020). Analysis of spatiotemporal variability of corn yields using empirical orthogonal functions. Water (Switzerland), 12(12) doi:10.3390/w12123339

171. Kleinheinz, G. T., McDermott, C. M., Hughes, S., & Brown, A. (2010). Effects of rainfall on *E. coli* concentrations at Door County, Wisconsin beaches. International Journal of Microbiology, 2009.

172. Kleinheinz, G.T., C.M. McDermott, M.C. Leewis, and E. Englebert. 2006. Influence of sampling depth on Escherichia coli concentrations in beach monitoring. Water Research 40(20):3831-3837.

173. Kohn, T., & Nelson, K. L. (2007). Sunlight-mediated inactivation of MS2 coliphage via exogenous singlet oxygen produced by sensitizers in natural waters. Environmental science & technology, 41(1), 192-197.

174. Korres, W., Koyama, C. N., Fiener, P., & Schneider, K. (2010). Analysis of surface soil moisture patterns in agricultural landscapes using Empirical Orthogonal Functions. Hydrology and Earth System Sciences, 14(5), 751-764.

175. Kovačić, A., Tripković, I., Galov, A., & Žitko, T. (2011). Distribution of microbiological indicators of fecal pollution in the riverine substrates. Environmental monitoring and assessment, 172(1), 623-630.

176. Kozak, G. K., MacDonald, D., Landry, L., & Farber, J. M. (2013). Foodborne outbreaks in Canada linked to produce: 2001 through 2009. Journal of food protection, 76(1), 173-183.

177. Kuhn, M. (2008). Building predictive models in R using the caret package. Journal of statistical software, 28(1), 1-26.

178. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26, p. 13). New York: Springer.

179. Kutser, T., Hedley, J., Giardino, C., Roelfsema, C., & Brando, V. E. (2020). Remote sensing of shallow waters–a 50 year retrospective and future directions. Remote Sensing of Environment, 240, 111619.

180. Lal, A., Hales, S., French, N., & Baker, M. G. (2012). Seasonality in human zoonotic enteric diseases: a systematic review. PLoS one, 7(4), e31883.

181. Laureano-Rosario, A. E., Duncan, A. P., Symonds, E. M., Savic, D. A., & Muller-Karger, F. E. (2019). Predicting culturable enterococci exceedances at Escambron Beach, San Juan, Puerto Rico using satellite remote sensing and artificial neural networks. Journal of water and health, 17(1), 137-148.

182. Lee, D., Tertuliano, M., Vellidis, G., Harris, C., Grossman, M. K., Rajeev, S., & Levy, K. (2018). Evaluation of grower-friendly, science-based sampling approaches for the detection of Salmonella in ponds used for irrigation of fresh produce. Foodborne pathogens and disease, 15(10), 627-636.

183. Lee, H., Song, H. -., & Sohn, B. -. (2021). Possible mechanisms of long-term trend of june rainfall over the korean peninsula. Asia-Pacific Journal of Atmospheric Sciences, 57(1), 51-62. doi:10.1007/s13143-019-00170-5

184. LGMA, 2015. California Leafy Green Products Handler Marketing Agreement. State of California Department of Food and Agriculture. https://lgma-assets.sfo2.digitaloceanspaces.com/content-images/2015-0504-LGMA-Marketing-Agreement.pdf. Accessed January 2022.

185. LGMA, 2021. Food Safety Practices. Leafy Green Marketing Agreement. https://lgmatech.com/wp-content/uploads/2020/08/Current-version_August-

2021-CA-LGMA-Metrics_FINAL-v20210818-clean-1262021.pdf. Accessed January 2022.

186. Li, M., Zhang, H., Lemckert, C., Roiko, A., & Stratton, H. (2018). On the hydrodynamics and treatment efficiency of waste stabilisation ponds: From a literature review to a strategic evaluation framework. Journal of Cleaner Production, 183, 495-514.

187. Li, Y., Wang, X., Zhao, Z., Han, S., & Liu, Z. (2020). Lagoon water quality monitoring based on digital image analysis and machine learning estimators. Water Research, 172 doi:10.1016/j.watres.2020.115471

188. Lim, C. H., & Flint, K. P. (1989). The effects of nutrients on the survival of Escherichia coli in lake water. Journal of applied bacteriology, 66(6), 559-569.

189. Liu, B., D'Sa, E. J., Maiti, K., Rivera-Monroy, V. H., & Xue, Z. (2021). Biogeographical trends in phytoplankton community size structure using adaptive sentinel 3-OLCI chlorophyll a and spectral empirical orthogonal functions in the estuarine-shelf waters of the northern gulf of mexico. Remote Sensing of Environment, 252 doi:10.1016/j.rse.2020.112154

190. Liu, L., Hall, G., & Champagne, P. (2016). Effects of environmental factors on the disinfection performance of a wastewater stabilization pond operated in a temperate climate. Water, 8(1), 5.

191. Liu, Lubo, Mantha S. Phanikumar, Stephanie L. Molloy, Richard L. Whitman, Dawn A. Shively, Meredith B. Nevers, David J. Schwab, and Joan B. Rose. "Modeling the Transport and Inactivation of *E. coli* and Enterococci in the Near-Shore Region of Lake Michigan." Environmental Science & Technology 40 (16), 5022–28. https://doi.org/10.1021/es060438k.

192. Liu, Y., Li, Y., Li, L., & Chen, C. (2019). Spatiotemporal variability of monthly and annual snow depths in xinjiang, china over 1961-2015 and the potential effects. Water (Switzerland), 11(8)

193. Lothrop, N., Bright, K. R., Sexton, J., Pearce-walker, J., Reynolds, K. A., & Verhougstraete, M. P. (2018). Optimal strategies for monitoring irrigation water quality. Agricultural Water Management, 199, 86–92. https://doi.org/10.1016/j.agwat.2017.12.018

194. Luo Z, Gu G, Ginn A, Giurcanu MC, Adams P, Vellidis G, van Bruggen AH, Danyluk MD, Wright AC. (2015). Distribution and characterization of Salmonella enterica isolates from irrigation ponds in the Southeastern United States. Appl Environl Microbiol 2015;81:4376–4387.

195. Luo, X., & Li, X. (2018). Using the EFDC model to evaluate the risks of eutrophication in an urban constructed pond from different water supply strategies. Ecological modelling, 372, 1-11.

196. Lušić, D. V., Kranjčević, L., Maćešić, S., Lušić, D., Jozić, S., Linšak, Ž., ... & Bilajac, N. (2017). Temporal variations analyses and predictive modeling of microbiological seawater quality. *Water research*, *119*, 160-170.

197. Maclin, R., & Opitz, D. (1997). An empirical evaluation of bagging and boosting. AAAI/IAAI, 1997, 546-551.

198. Maguire, M., Kase, J. A., Brown, E. W., Allard, M., Musser, S., & González-Escalona, N. (2021). Metagenomic survey of agricultural water using long read sequencing: considerations for a successful analysis.

199. Maïga, Y., Wethe, J., Denyigba, K., & Ouattara, A. S. (2009). The impact of pond depth and environmental conditions on sunlight inactivation of Escherichia coli and enterococci in wastewater in a warm climate. Canadian journal of microbiology, 55(12), 1364-1374.

200. Manchanda, S., Dave, M., & Singh, S. B. (2007). An Empirical Comparison Of Supervised Learning Processes. International Journal of Engineering, 1(1), 21.

201. Maraccini, P. A., Mattioli, M. C. M., Sassoubre, L. M., Cao, Y., Griffith, J. F., Ervin, J. S., ... & Boehm, A. B. (2016). Solar inactivation of enterococci and Escherichia coli in natural waters: effects of water absorbance and depth. Environmental science & technology, 50(10), 5068-5076.

202. Marine, S. C., Pagadala, S., Wang, F., Pahl, D. M., Melendez, M. V., Kline, W. L., ... & Micallef, S. A. (2015). The growing season, but not the farming system, is a food safety risk determinant for leafy greens in the mid-Atlantic region of the United States. Applied and environmental microbiology, 81(7), 2395-2407.

203. Markland, S. M., Ingram, D., Kniel, K. E., & Sharma, M. (2017). Water for agriculture: the convergence of sustainability and safety. Microbiology spectrum, 5(3), 5-3.

204. Martinez, B., Stratton, J., Bianchini, A., Wegulo, S., & Weaver, G. (2015). Transmission of Escherichia coli O157: H7 to internal tissues and its survival on flowering heads of wheat. Journal of food protection, 78(3), 518-524.

205. Mas, D.M.L., Ahlfeld, D.P., 2007. Comparing artificial neural networks and regression models for predicting faecal coliform concentrations. Hydrol. Sci. J. 52, 713–731.

206. Mayo, A. W., & Noike, T. (1994). Response of mixed cultures of Chlorella vulgaris and heterotrophic bacteria to variation of pH. Water Science and Technology, 30(8), 285.

207. McEgan, R., Mootian, G., Goodridge, L. D., Schaffner, D. W., & Danyluk, M. D. (2013). Predicting Salmonella populations from biological, chemical, and physical indicators in Florida surface waters. Applied and Environmental Microbiology, 79(13), 4094-4105.

208. Meays, C. L., Broersma, K., Nordin, R., Mazumder, A., & Samadpour, M. (2006). Diurnal variability in concentrations and sources of Escherichia coli in three streams. Canadian journal of microbiology, 52(11), 1130-1135.

209. Mendonca, A. F., Amoroso, T. L., & Knabel, S. J. (1994). Destruction of gram-negative food-borne pathogens by high pH involves disruption of the cytoplasmic membrane. Applied and Environmental Microbiology, 60(11), 4009-4014.

210. Micallef, S. A., Goldstein, R. E. R., George, A., Kleinfelter, L., Boyer, M. S., McLaughlin, C. R., ... & Sapkota, A. R. (2012). Occurrence and antibiotic resistance of multiple Salmonella serotypes recovered from water, sediment and soil on mid-Atlantic tomato farms. *Environmental research*, *114*, 31-39.

211. Modaresi, F., & Araghinejad, S. (2014). A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification. Water resources management, 28(12), 4095-4111.

212. Mohammed, H., Hameed, I. A., & Seidu, R. (2018). Comparative predictive modelling of the occurrence of faecal indicator bacteria in a drinking water source in Norway. Science of the Total Environment, 628, 1178-1190.

213. Mohammed, H., Tornyeviadzi, H. M., & Seidu, R. (2021). Modelling the impact of weather parameters on the microbial quality of water in distribution systems. Journal of Environmental Management, 284, 111997.

214. Monaghan, J. M., & Hutchison, M. L. (2012). Distribution and decline of human pathogenic bacteria in soil after application in irrigation water and the potential for soil-splash-mediated dispersal onto fresh produce. Journal of applied microbiology, 112(5), 1007-1019.

215. Morgan, B. J., Stocker, M. D., Pachepsky, Y. A., & Kim, M. S. (2021, April). Accuracy and reliability of predictions of *E. coli* concentrations in water of irrigation ponds from drone-based imagery as affected by parameters of the random forest algorithm. In Sensing for Agriculture and Food Quality and Safety XIII (Vol. 11754, p. 117540F). International Society for Optics and Photonics.

216. Motamarri, S., & Boccelli, D. L. (2012). Development of a neural-based forecasting tool to classify recreational water quality using fecal indicator organisms. Water research, 46(14), 4508-4520.

217. Moyne, A.-L., Sudarshana, M.R., Blessington, T., Koike, S.T., Cahn, M.D. and Harris, L.J. (2011) Fate of Escherichia coli O157:H7 in field-inoculated lettuce. Food Microbiol 28, 1417–1425.

218. Mudd, Dennis, Tamara Anan'eva, and Julie Kinzelman. (2012). Examination of Diurnal Variation at a Non-Sewage Impacted Beach via QPCR and Culture Based Methods." Journal of Environmental Protection 03 (10) 1310–17. https://doi.org/10.4236/jep.2012.310149.

219. Muirhead, R. W. (2009). Soil and faecal material reservoirs of Escherichia coli in a grazed pasture. New Zealand Journal of Agricultural Research, 52(1), 1-8.

220. Muirhead, R. W., & Monaghan, R. M. (2012). A two reservoir model to predict Escherichia coli losses to water from pastures grazed by dairy cows. Environment international, 40, 8-14.

221. Muirhead, R. W., Collins, R. P., & Bremer, P. J. (2006). The association of E. coli and soil particles in overland flow. Water Science and Technology, 54(3), 153-159.

222. Nadeau, Claude, and Yoshua Bengio. "Inference for the generalization error." Machine learning 52, no. 3 (2003): 239-281.

223. Nagels, J. W., Davies-Colley, R. J., Donnison, A. M., & Muirhead, R. W. (2002). Faecal contamination over flood events in a pastoral agricultural stream in New Zealand. Water Science and Technology, 45(12), 45-52.

224. Navarra, A., & Simoncini, V. (2010). A guide to empirical orthogonal functions for climate data analysis. Springer Science & Business Media.

225. Nguyen, M. T., Jasper, J. T., Boehm, A. B., & Nelson, K. L. (2015). Sunlight inactivation of fecal indicator bacteria in open-water unit process treatment wetlands: modeling endogenous and exogenous inactivation rates. Water research, 83, 282-292.

226. NRCS, 1997. Ponds – Planning, Design, Construction. Natural Resource Conservation Service. United States Department of Agriculture. https://nrcspad.sc.egov.usda.gov/distributioncenter/product.aspx?ProductID=115 . Accessed January 2022.

227. Nygård, K., Lassen, J., Vold, L., Andersson, Y., Fisher, I., Löfdahl, S., Threlfall, J, Luzzi, I, Peters, T, Hampton, M., Torpdahl, M, Kapperud, G & Aavitsland, P. (2008). Outbreak of Salmonella Thompson infections linked to imported rucola lettuce. Foodborne Pathogens and Disease, 5(2), 165-173.

228. O'Callaghan, P., Kelly-Quinn, M., Jennings, E., Antunes, P., O'Sullivan, M., Fenton, O., & Ó hUallacháin, D. (2018). Impact of Cattle Access to Watercourses: Literature Review on Behalf of the COSAINT Project. Environmental Protection Agency of Ireland: Wexford, Ireland.

229. O'Keefe, T.C., Elliot, S.R., and Naiman, R.J. Introduction to Watershed Ecology. United States Environmental Protection Agency. https://cfpub.epa.gov/watertrain/moduleFrame.cfm?module_id=20&parent_object_id=516&object_id=516 . Accessed January 2022.

230. Odonkor, S. T., & Ampofo, J. K. (2013). Escherichia coli as an indicator of bacteriological quality of water: an overview. Microbiology research, 4(1), 5-11.

231. Olaimat, A. N., & Holley, R. A. (2012). Factors influencing the microbial safety of fresh produce: a review. Food microbiology, 32(1), 1-19.

232. Oliveira, M., Usall, J., Viñas, I., Solsona, C., & Abadias, M. (2011). Transfer of Listeria innocua from contaminated compost and irrigation water to lettuce leaves. Food microbiology, 28(3), 590-596.

233. Oliveira, M., Viñas, I., Usall, J., Anguera, M., & Abadias, M. (2012). Presence and survival of Escherichia coli O157: H7 on lettuce leaves and in soil treated with contaminated compost and irrigation water. International journal of food microbiology, 156(2), 133-140.

234. Overland, J. E., & Preisendorfer, R. W. (1982). A significance test for principal components applied to a cyclone climatology. Monthly Weather Review, 110(1), 1-4.

235. Ozkanca, R. (1993). Survival and physiological status of Escherichia coli in lake water under different nutrient conditions (Doctoral dissertation, University of Warwick).

236. Pachepsky, Y. A., & Shelton, D. R. (2011). Escherichia coli and fecal coliforms in freshwater and estuarine sediments. Critical reviews in environmental science and technology, 41(12), 1067-1110.

237. Pachepsky, Y. A., Blaustein, R. A., Whelan, G. E. N. E., & Shelton, D. R. (2014). Comparing temperature effects on Escherichia coli, Salmonella, and Enterococcus survival in surface waters. Letters in applied microbiology, 59(3), 278-283.

238. Pachepsky, Y., Kierzewski, R., Stocker, M., Sellner, K., Mulbry, W., Lee, H., Kim, M. (2018). Temporal stability of Escherichia coli concentrations in waters of two irrigation ponds in Maryland. Applied and Environmental Microbiology, 84, 1876–1893. https://doi.org/10.1128/AEM.01876-17

239. Pachepsky, Y., Shelton, D. R., McLain, J. E. T., Patel, J., Mandrell, R. (2014). Irrigation Waters as a Source of Pathogenic Microorganisms in Produce. A Review. Advances in Agronomy, 113:73-138 . 2011. Environmental Science and Pollution Research.27:4021-4031.

240. Pachepsky, Y., Shelton, D., Dorner, S., & Whelan, G. (2016). Can *E. coli* or thermotolerant coliform concentrations predict pathogen presence or prevalence in irrigation waters?. Critical reviews in microbiology, 42(3), 384-393.

241. Pachepsky, Y., Stocker, M., Saldaña, M. O., & Shelton, D. (2017). Enrichment of stream water with fecal indicator organisms during baseflow periods. Environmental monitoring and assessment, 189(2), 51.

242. Pagadala, S., Marine, S. C., Micallef, S. A., Wang, F., Pahl, D. M., Melendez, M. V., ... & Buchanan, R. L. (2015). Assessment of region, farming system, irrigation source and sampling time as food safety risk factors for tomatoes. International Journal of Food Microbiology, 196, 98-108.

243. Pahl, D. M., Telias, A., Newell, M., Ottesen, A. R., & Walsh, C. S. (2013). Comparing source of agricultural contact water and the presence of fecal indicator organisms on the surface of 'Juliet'grape tomatoes. Journal of food protection, 76(6), 967-974.

244. Painter, J. A., Hoekstra, R. M., Ayers, T., Tauxe, R. V., Braden, C. R., Angulo, F. J., & Griffin, P. M. (2013). Attribution of foodborne illnesses,

hospitalizations, and deaths to food commodities by using outbreak data, United States, 1998–2008. Emerging infectious diseases, 19(3), 407.

245. Pan, X., Li, X., Cheng, G., Chen, R., & Hsu, K. (2017). Impact analysis of climate change on snow over a complex mountainous region using weather research and forecast model (WRF) simulation and moderate resolution imaging spectroradiometer data (MODIS)-terra fractional snow cover products. Remote Sensing, 9(8), 1-19

246. Panidhapu, A., Li, Z., Aliashrafi, A., & Peleato, N. M. (2020). Integration of weather conditions for predicting microbial water quality using Bayesian Belief Networks. Water research, 170, 115349.

247. Park, Y., Kim, M., Pachepsky, Y., Choi, S. H., Cho, J. G., Jeon, J., & Cho, K. H. (2018). Development of a nowcasting system using machine learning approaches to predict fecal contamination levels at recreational beaches in Korea. Journal of environmental quality, 47(5), 1094-1102.

248. Park, Y., Pachepsky, Y. A., Cho, K. H., Jeon, D. J., & Kim, J. H. (2015). Stressor-response modeling using the 2D water quality model and regression trees to predict chlorophyll-a in a reservoir system. Journal of Hydrology, 529(P3), 805-815.

249. Park, Y., Pachepsky, Y., Hong, E. M., Shelton, D., & Coppock, C. (2017). Escherichia coli release from streambed to water column during baseflow periods: a modeling study. Journal of environmental quality, 46(1), 219-226.

250. Park, Y.S., Cereghino, R., Compin, A. & Lek, S. (2003). Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecological Modelling, 160, 3, 265-280, ISSN: 0304-3800

251. Partyka, M. L., Bond, R. F., Chase, J. A., and Atwill, E. R. (2018). Spatiotemporal variability in microbial quality of western US agricultural water supplies: a multistate study. J. Environ. Qual. 47, 939–948. doi: 10.2134/jeq2017.12.0501

252. Partyka, M. L., Bond, R. F., Chase, J. A., Kiger, L., & Atwill, E. R. (2016). Multistate evaluation of microbial water and sediment quality from agricultural recovery basins. Journal of environmental quality, 45(2), 657-665.

253. Pascual, M. & Dunne, J.A. (2006). From small to large ecological networks in a dynamic world, In Ecological Networks: Linking Structure to Dynamics in Food Webs, Pascual M. & Dunne A. (Editors), 3-24, Oxford University Press, ISBN-10: 0195188160, Oxford

254. Pavione, D. M. S., Bastos, R. K. X., & Bevilacqua, P. D. (2013). Quantitative microbial risk assessment applied to irrigation of salad crops with waste stabilization pond effluents. Water Science and Technology, 67(6), 1208-1215.

255. Pearson, H. W., Mara, D. D., Mills, S. W., & Smallman, D. J. (1987). Physico-chemical parameters influencing faecal bacterial survival in waste stabilization ponds. Water science and technology, 19(12), 145-152.

256. Pedrera-Parrilla, A., Pachepsky, Y.A., Taguas, E. V., Martos-Rosillo, S., Giráldez, J. V., Vanderlinden, K. (2017). Concurrent temporal stability of the apparent electrical conductivity and soil water content. Journal of Hydrology, 544, 319–326. https://doi.org/10.1016/j.jhydrol.2016.10.017

257. Perry, M. A., & Niemann, J. D. (2007). Analysis and estimation of soil moisture at the catchment scale using EOFs. Journal of Hydrology, 334(3-4), 388-404.

258. Piorkowski, G., Jamieson, R., Bezanson, G., Hansen, L. T., & Yost, C. (2014). Reach specificity in sediment *E. coli* population turnover and interaction with waterborne populations. Science of the Total Environment, 496, 402-413.

259. Polat, H., Topalcengiz, Z., & Danyluk, M. D. (2020). Prediction of Salmonella presence and absence in agricultural surface waters by artificial intelligence approaches. Journal of Food Safety, 40(1), e12733.

260. Poulsen, M. N., Pollak, J., Sills, D. L., Casey, J. A., Rasmussen, S. G., Nachman, K. E., & Schwartz, B. S. (2018). Residential proximity to high-density poultry operations associated with campylobacteriosis and infectious diarrhea. International journal of hygiene and environmental health, 221(2), 323-333.

261. Prakash, R., Tharun, V. P., & Devi, S. R. (2018). A comparative study of various classification techniques to determine water quality. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1501-1506). IEEE.

262. Pyo, J., Ha, S., Pachepsky, Y.A., Lee, H., Ha, R., Nam, G., Kim, M.S., Im, J., Cho, K.H. (2016). Chlorophyll-a concentration estimation using three difference bio-optical algorithms, including a correction for the low-concentration range: the case of the Yiam reservoir, Korea . Taylor Francis, 7, 407–416. https://doi.org/10.1080/2150704X.2016.1142680

263. Quetglas, A., Ordines, F., & Guijarro, B. (2011). The use of artificial neural networks (ANNs) in aquatic ecology. IntechOpen.

264. Quilliam, R. S., Clements, K., Duce, C., Cottrill, S. B., Malham, S. K., & Jones, D. L. (2011). Spatial variation of waterborne Escherichia coli–implications for routine water quality monitoring. Journal of Water and Health, 9, 734-737.

265. R Development Core Team. (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria.

266. Ransom, C. J., Kitchen, N. R., Camberato, J. J., Carter, P. R., Ferguson, R. B., Fernández, F. G., . . . Shanahan, J. F. (2019). Statistical and machine learning methods evaluated for incorporating soil and weather into corn nitrogen

recommendations. Computers and Electronics in Agriculture, 164 doi:10.1016/j.compag.2019.104872

267. Renwick, W. H., Smith, S. V., Bartley, J. D., & Buddemeier, R. W. (2005). The role of impoundments in the sediment budget of the conterminous United States. Geomorphology, 71(1-2), 99-111.

268. Ribera, L. A., Yamazaki, F., Paggi, M., & Seale, J. L. (2016). The Food Safety Modernization Act and production of specialty crops. Choices, 31(1), 1-7.

269. Rincon, A. G., & Pulgarin, C. (2004). Effect of pH, inorganic ions, organic matter and H2O2 on *E. coli* K12 photocatalytic inactivation by TiO2: implications in solar water disinfection. Applied Catalysis B: Environmental, 51(4), 283-302.

270. Rokitko, P. V., Romanovs' ka, V. O., IuR, M., & Chorna, N. A. (2004). Effect of UV-radiation and drying on bacterium diversity in soil. Mikrobiolohichnyi zhurnal (Kiev, Ukraine: 1993), 66(1), 68-77.

271. Rokitko, P. V., Romanovskaya, V. A., Malashenko, Y. R., Chernaya, N. A., Gushcha, N. I., & Mikheev, A. N. (2003). Soil drying as a model for the action of stress factors on natural bacterial populations. Microbiology, 72(6), 756-761.

272. Sah, L., Rousseau, D. P., & Hooijmans, C. M. (2012). Numerical modelling of waste stabilization ponds: where do we stand?. Water, Air, & Soil Pollution, 223(6), 3155-3171.

273. Sah, L., Rousseau, D. P., Hooijmans, C. M., & Lens, P. N. (2011). 3D model for a secondary facultative pond. Ecological Modelling, 222(9), 1592-1603.

274. Saldaña, Z., Sánchez, E., Xicohtencatl-Cortes, J., Puente, J. L., & Girón, J. A. (2011). Surface structures involved in plant stomata and leaf colonization by Shiga-toxigenic Escherichia coli O157: H7. Frontiers in microbiology, 2, 119.

275. Saravanan, R., Yang, J., Jun, M., Schumacher, C., Wang, J., & Wong, R. (2020). Towards data-driven approaches for simulating rainfall in climate models. Earth and Space Science Open Archive ESSOAr.

276. Savill MG, Hudson JA, Ball A, et al. (2001). Enumeration of Campylobacter in New Zealand recreational and drinking waters. J Appl Microbiol 91:38–46.

277. Scharff, R. L. (2010). Health-related costs from foodborne illness in the United States.
2010. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.172.4518 . Accessed January 2022.

278. Schliep, K., Hechenbichler, K., and Schliep, M.K. (2016). kknn: Weighted k-Nearest Neighbors. R package version 1.3.1. https://CRAN.R-project.org/package=kknn

279. Scott, E. E., Leh, M. D., & Haggard, B. E. (2017). Spatiotemporal variation of bacterial water quality and the relationship with pasture land cover. Journal of water and health, 15(6), 839-848.

280. Searcy, R. T., & Boehm, A. B. (2021). A Day at the Beach: Enabling Coastal Water Quality Prediction with High-Frequency Sampling and Data-Driven Models. Environmental Science & Technology, 55(3), 1908-1918.

281. Seo, D., Kim, M., & Ahn, J. H. (2012). Prediction of chlorophyll-a changes due to weir constructions in the Nakdong River using EFDC-WASP modelling. Environmental Engineering Research, 17(2), 95-102.

282. Shelton, D. R., Karns, J. S., Coppock, C., Patel, J., Sharma, M., & Pachepsky, Y. A. (2011). Relationship between eae and stx virulence genes and Escherichia coli in an agricultural watershed: implications for irrigation water standards and leafy green commodities. Journal of food protection, 74(1), 18-23.

283. Shelton, D. R., Pachepsky, Y. A., Kiefer, L. A., Blaustein, R. A., McCarty, G. W., & Dao, T. H. (2014). Response of coliform populations in streambed sediment and water column to changes in nutrient concentrations in water. water research, 59, 316-324.

284. Simpson, R. B., Zhou, B., & Naumova, E. N. (2020). Seasonal synchronization of foodborne outbreaks in the United States, 1996–2017. Scientific reports, 10(1), 1-16.

285. Sinton, Lester W., Carollyn H. Hall, Philippa A. Lynch, and Robert J. Davies-Colley. (2002). Sunlight Inactivation of Fecal Indicator Bacteria and Bacteriophages from Waste Stabilization Pond Effluent in Fresh and Saline Waters. Applied and Environmental Microbiology 68(3), 1122–31. https://doi.org/10.1128/AEM.68.3.1122-1131.2002.

286. Smallman, D. J. (1986). An ecological appraisal of waste stabilization pond performance (Doctoral dissertation, University of Leeds).

287. Smith, B. A., & Fazil, A. (2019). Climate change and infectious diseases: The challenges: How will climate change impact microbial foodborne disease in Canada? Canada Communicable Disease Report, 45(4), 108.

288. Smith, M. G. (1995). Survival of *E. coli* and Salmonella after chilling and freezing in liquid media. Journal of food science, 60(3), 509-512.

289. Smith, P., Carroll, C., Wilkins, B., Johnson, P., Gabhainn, S. N., & Smith, L. P. (1999). The effect of wind speed and direction on the distribution of sewage-associated bacteria. Letters in applied microbiology, 28(3), 184-188.

290. Smith, R. E., Allen, C. D., & Charlton, M. N. (2004). Dissolved organic matter and ultraviolet radiation penetration in the Laurentian Great Lakes and tributary waters. Journal of Great Lakes Research, 30(3), 367-380.

291. Sokolova, E., Ivarsson, O., Lilliesträm, A., Speicher, N. K., Rydberg, H., & Bondelind, M. (2022). Data-driven models for predicting microbial water

quality in the drinking water source using *E. coli* monitoring and hydrometeorological data. Science of The Total Environment, 802, 149798.

292. Sokolova, E., Pettersson, T. J., Bergstedt, O., & Hermansson, M. (2013). Hydrodynamic modelling of the microbial water quality in a drinking water source as input for risk reduction management. Journal of hydrology, 497, 15-23.

293. Solaiman, S., Allard, S. M., Callahan, M. T., Jiang, C., Handy, E., East, C., ... & Micallef, S. A. (2020). Longitudinal assessment of the dynamics of Escherichia coli, total coliforms, Enterococcus spp., and Aeromonas spp. in alternative irrigation water sources: a CONSERVE Study. Applied and environmental microbiology, 86(20), e00342-20.

294. Solic, Mladen, and Nada Krstulovic. (1992). Separate and Combined Effects of Solar Radiation, Temperature, Salinity, and pH on the Survival of Faecal Coliforms in Seawater. Marine Pollution Bulletin 24, 411–16. https://doi.org/10.1016/0025-326X(92)90503-X.

295. Solomon, E. B., Yaron, S., & Matthews, K. R. (2002). Transmission of Escherichia coli O157: H7 from contaminated manure and irrigation water to lettuce plant tissue and its subsequent internalization. Applied and environmental microbiology, 68(1), 397-400.

296. Steele, M., Mahdi, A., & Odumeru, J. (2005). Microbial assessment of irrigation water used for production of fruit and vegetables in Ontario, Canada. Journal of Food Protection, 68(7), 1388-1392.

297. Stidson, R.T., Gray, C.A., Mcphail, C.D., 2012. Development and use of modelling techniques for real-time bathing water quality predictions. Water Environ. J. 26, 7–18

298. Stocker, 2017. Release, Survival, and Removal of Bovine Manure-borne Indicator Bacteria Under Simulated Rainfall. Masters Thesis. University of Maryland.

299. Stocker, M. D., Pachepsky, Y. A., & Shelton, D. R. (2014). Performance of Weibull and Linear Semi-logarithmic Models in Simulating Escherichia coli Inactivation in Waters. Journal of environmental quality, 43(5), 1559-1565.

300. Stocker, M. D., Pachepsky, Y. A., Hill, R. L., & Shelton, D. R. (2015). Depth-dependent survival of Escherichia coli and enterococci in soil after manure application and simulated rainfall. Applied and environmental microbiology, 81(14), 4801-4808.

301. Stocker, M. D., Pachepsky, Y. A., Hill, R. L., Sellner, K. G., Macarisin, D., & Staver, K. W. (2019). Intraseasonal variation of *E. coli* and environmental covariates in two irrigation ponds in Maryland, USA. Science of the Total Environment, 670, 732-740.

302. Stocker, M. D., Penrose, M., & Pachepsky, Y. A. (2018). Spatial patterns of Escherichia coli concentrations in sediment before and after high-flow events in a first-order creek. Journal of environmental quality, 47(5), 958-966.

303. Stocker, M. D., Rodriguez-Valentin, J. G., Pachepsky, Y. A., & Shelton, D. R. (2016). Spatial and temporal variation of fecal indicator organisms in two creeks in Beltsville, Maryland. Water Quality Research Journal of Canada, 51(2), 167-179.

304. Stocker, M. D., Smith, J. E., Hernandez, C., Macarisin, D., & Pachepsky, Y. (2019). Seasonality of *E. coli* and enterococci concentrations in creek water, sediment, and periphyton. Water, Air, & Soil Pollution, 230(9), 1-12.

305. Stocker, M., Jeon, D., Sokolova, E., Lee, H., Kim, M., Pachepsky, Y.,(2020). Accounting for the Three-Dimensional Distribution of Escherichia coli Concentrations in Pond Water in Simulations of the Microbial Quality of Water Withdrawn for Irrigation. Water, 12, 1708. https://doi.org/10.3390/w12061708

306. Stocker, M.D, Pachepsky, Y.A., Smith, J.E., Morgan, B.J., Hill, R.L., Kim, M. (2021) Persistent patterns of *E. coli* concentrations in two irrigation ponds from 3 years of monitoring. Water, air, & soil pollution. Volume awaiting assignment. DOI : 10.1007/s11270-021-05438-z.

307. Strawn, L. K., Fortes, E. D., Bihn, E. A., Nightingale, K. K., Gröhn, Y. T., Worobo, R. W., & Bergholz, P. W. (2013). Landscape and meteorological factors affecting prevalence of three food-borne pathogens in fruit and vegetable farms. Applied and environmental microbiology, 79(2), 588-600.

308. Strogatz, S.H. (2001). Exploring complex networks. Nature, 410, 6825, 268-276, ISSN: 0028-0836

309. Struck, S. D., Selvakumar, A., & Borst, M. (2008). Prediction of effluent quality from retention ponds and constructed wetlands for managing bacterial stressors in storm-water runoff. Journal of irrigation and drainage engineering, 134(5), 567-578.

310. Sukias, James P. S., Chris C. Tanner, Robert J. Davies-Colley, John W. Nagels, and Ronald Wolters. (2001). Algal Abundance, Organic Matter, and Physico-chemical Characteristics of Dairy Farm Facultative Ponds: Implications for Treatment Performance. New Zealand Journal of Agricultural Research 44 (4), 279–96. https://doi.org/10.1080/00288233.2001.9513485.

311. Suslow, T. V., Oria, M. P., Beuchat, L. R., Garrett, E. H., Parish, M. E., Harris, L. J., ... & Busta, F. F. (2003). Production practices as risk factors in microbial food safety of fresh and fresh-cut produce. Comprehensive reviews in food science and food safety, 2, 38-77.

312. Tan, B., Ng, C. M., Nshimyimana, J. P., Loh, L. L., Gin, K. Y. H., & Thompson, J. R. (2015). Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges, and future opportunities. Frontiers in microbiology, 6, 1027.

313. Taylor, M. (2017). sinkr: Collection of functions with emphasis in multivariate data analysis. R Package Version 0.6. Available at https://rdrr.io/github/marchtaylor/sinkr/

314. Thoe, W., Gold, M., Griesbach, A., Grimmer, M., Taggart, M.L., Boehm, A.B., 2014. Predictingwater quality at Santa Monica Beach: evaluation of five different models for public notification of unsafe swimming conditions. Water Res. 67, 105–117.

315. Thoe, W., Wong, S.H.C., Choi, K.W., Lee, J.H.W., 2012. Daily prediction of marine beach water quality in Hong Kong. J. Hydro-Environ. Res. 6, 164–180.

316. Thomas, M. K., Fontana, S., Reyes, M., Kehoe, M., & Pomati, F. (2018). The predictability of a lake phytoplankton community, over time-scales of hours to years. Ecology Letters, 21(5), 619–628.

317. Thomas, M. K., Murray, R., Flockhart, L., Pintar, K., Fazil, A., Nesbitt, A., Marshall, B., Tataryn, J. & Pollari, F. (2015). Estimates of foodborne illness–related hospitalizations and deaths in Canada for 30 specified pathogens and unspecified agents. Foodborne Pathogens and Disease, 12(10), 820-827.

318. Topalcengiz, Z., Strawn, L. K., & Danyluk, M. D. (2017). Microbial quality of agricultural water in Central Florida. PLoS One, 12(4), e0174889.

319. Tousi, E. G., Duan, J. G., Gundy, P. M., Bright, K. R., & Gerba, C. P. (2021). Evaluation of *E. coli* in sediment for assessing irrigation water quality using machine learning. Science of The Total Environment, 149286.

320. Traister, E., & Anisfeld, S. C. (2006). Variability of indicator bacteria at different time scales in the upper Hoosic River watershed. Environmental science & technology, 40(16), 4990-4995.

321. Truchado, P., Hernandez, N., Gil, M. I., Ivanek, R., & Allende, A. (2018). Correlation between *E. coli* levels and the presence of foodborne pathogens in surface irrigation water: establishment of a sampling program. Water research, 128, 226-233.

322. Truitt, L. N., Vazquez, K. M., Pfuntner, R. C., Rideout, S. L., Havelaar, A. H., & Strawn, L. K. (2018). Microbial quality of agricultural water used in produce preharvest production on the Eastern Shore of Virginia. Journal of food protection, 81(10), 1661-1672.

323. Tyrrel, S. F., & Quinton, J. N. (2003). Overland flow transport of pathogens from agricultural land receiving faecal wastes. Journal of Applied Microbiology, 94, 87-93.

324. U.S. Environmental Protection Agency (U.S. EPA). 2015. Polluted Runoff: Nonpoint Source (NPS) Pollution. United States Environmental Protection Agency. https://www.epa.gov/nps/nonpoint-source-agriculture. Accessed January 2022.

325. U.S. Environmental Protection Agency (U.S. EPA). 2021. National Summary of State Information. United States Environmental Protection Agency. https://ofmpub.epa.gov/waters10/attains_nation_cy.control#prob_source . Accessed January 2022.

326. U.S. Environmental Protection Agency (U.S. EPA). (2002). Method 1603. Escherichia coli (*E. coli*) in water by membrane filtration using modified membrane thermotolerant Escherichia coli agar (modified mTEC). EPA 821-R-02-023. September 2002.

327. U.S. Environmental Protection Agency (U.S. EPA). (2005) Method 1603: Escherichia coli (*E. coli*) in Water by Membrane Filtration Using Modified membrane-Thermotolerant Escherichia coli Agar (Modified mTEC. EPA-821-R-04-025 U.S. Environmental Protection Agency, Office of Water Washington, D.C.

328. U.S. Environmental Protection Agency (U.S. EPA). (2009). Fact sheet: Introduction to Clean Water Act (CWA) Section 303 (d) impaired waters lists. < https://nepis.epa.gov/Exe/tiff2png.exe/P1008DPU.PNG?-r+75+-g+7+D%3A%5CZYFILES%5CINDEX%20DATA%5C06THRU10%5CTIFF%5C00000889%5CP1008DPU.TIF>. Accessed on November 1st , 2021.

329. U.S. Environmental Protection Agency (U.S. EPA). (2009). Statistical analysis of groundwater monitoring data at RCRA facilities: unified guidance. EPA 530/R-09-007. U.S. EPA Office of Resource Conservation and Recovery, Washington, DC.

330. U.S. Environmental Protection Agency (U.S. EPA). (2010). Sampling and Consideration of Variability (Temporal and Spatial) For Monitoring of Recreational Waters. EPA-823-R-10-005. U.S. EPA Office of Water.

331. U.S. Environmental Protection Agency (U.S. EPA). (2014). Method 1603: Escherichia coli (*E. coli*) in water by membrane filtration using modified membrane-thermotolerant Escherichia coli agar (modified mTEC). EPA 821-R-14-010. U.S. EPA Office of Water, Washington, DC.

332. U.S. Environmental Protection Agency (U.S. EPA). 2012. Recreational Water Quality Criteria and Methods. Environmental Protection Agency. https://www.epa.gov/wqc/recreational-water-quality-criteria-and-methods. Accessed January 2022.

333. U.S. Environmental Protection Agency (U.S. EPA). 2013. Revised Total Coliform Rule and Total Coliform Rule. Environmental Protection Agency. https://www.epa.gov/dwreginfo/revised-total-coliform-rule-and-total-coliform-rule. Accessed January 2022.

334. USDA – United States Department of Agriculture (2015). 2015–2020 Dietary Guidelines for Americans. 8th Edition. December 2015. Available from http://health.gov/dietaryguidelines/2015/guidelines/. Accessed January 2020.

335. USGS, 2019. Irrigation Water Use. Water Resources. United States Geological Survey. https://www.usgs.gov/mission-areas/water-resources/science/irrigation-water-use. Accessed January 2022.

336. Uyttendaele, M., Jaykus, L. A., Amoah, P., Chiodini, A., Cunliffe, D., Jacxsens, L., Holvoet, K., Korsten, L., Lau, M., McClure, P., Medema, G.,

Sampers, I., & Rao Jasti, P. (2015). Microbial hazards in irrigation water: standards, norms, and testing to manage use of water in fresh produce primary production. Comprehensive Reviews in Food Science and Food Safety, 14(4), 336-356.

337. Vachaud, G., Passerat De Silans, A., Balabanis, P., Vauclin, M. (1985). Temporal Stability of Spatially Measured Soil Water Probability Density Function. Soil Science Society of America Journal, 49, 822–828. https://doi.org/10.2136/sssaj1985.03615995004900040006x

338. Van Kessel, J. S., Pachepsky, Y. A., Shelton, D. R., & Karns, J. S. (2007). Survival of Escherichia coli in cowpats in pasture and in laboratory conditions. Journal of Applied Microbiology, 103(4), 1122-1127.

339. Vanderlinden, K., Vereecken, H., Hardelauf, H., Herbst, M., Martínez, G., Cosh, M.H., Pachepsky, Y.A. (2012). Temporal Stability of Soil Water Contents: A Review of Data and Analyses. Vadose Zone Journal, 11, vzj2011.0178. https://doi.org/10.2136/vzj2011.0178

340. Vapnik, V., Guyon, I., & Hastie, T. (1995). Support vector machines. Mach. Learn, 20(3), 273-297.

341. Vereecken, H., Pachepsky, Y., Simmer, C., Rihani, J., Kunoth, A., Korres, W., Graf, A., Franssen, H.H., Thiele-Eich, I. and Shao, Y., 2016. On the role of patterns in understanding the functioning of soil-vegetation-atmosphere systems. Journal of Hydrology 542 (2016) 63–86

342. Vereecken, H., Schnepf, A., Hopmans, J.W., Javaux, M., Or, D., Roose, T., Vanderborght, J., Young, M.H., Amelung, W., Aitkenhead, M., Allison, S.D., Assouline, S., Baveye, P., Berli, M., Brüggemann, N., Finke, P., Flury, M., Gaiser, T., Govers, G., Ghezzehei, T., Hallett, P., Hendricks Franssen, H.J., Heppell, J., Horn, R., Huisman, J.A., Jacques, D., Jonard, F., Kollet, S., Lafolie, F., Lamorski, K., Leitner, D., McBratney, A., Minasny, B., Montzka, C., Nowak, W., Pachepsky, Y., Padarian, J., Romano, N., Roth, K., Rothfuss, Y., Rowe, E.C., Schwen, A., Šimůnek, J., Tiktak, A., Van Dam, J., van der Zee, S.E.A.T.M., Vogel, H.J., Vrugt, J.A., Wöhling, T., Young, I.M. (2016). Modeling Soil Processes: Review, Key Challenges, and New Perspectives. Vadose Zone Journal, 15, vzj2015.09.0131. https://doi.org/10.2136/vzj2015.09.0131

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

344. Vivant, A. L., Garmyn, D., & Piveteau, P. (2013). Listeria monocytogenes, a down-to-earth pathogen. Frontiers in cellular and infection microbiology, 3, 87.

345. Vogeleer, P., Tremblay, Y. D., Mafu, A. A., Jacques, M., & Harel, J. (2014). Life on the outside: role of biofilms in environmental persistence of Shiga-toxin producing Escherichia coli. Frontiers in microbiology, 5, 317.

346. Wade, T. J., Pai, N., Eisenberg, J. N., & Colford Jr, J. M. (2003). Do US Environmental Protection Agency water quality guidelines for recreational waters prevent gastrointestinal illness? A systematic review and meta-analysis. Environmental health perspectives, 111(8), 1102-1109.

347. Wang, H. (2019). Irrigation efficiency and water withdrawal in US agriculture. Water Policy, 21(4), 768-786.

348. Wang, L., Zhu, Z., Sassoubre, L., Yu, G., Liao, C., Hu, Q., & Wang, Y. (2021). Improving the robustness of beach water quality modeling using an ensemble machine learning approach. Science of The Total Environment, 765, 142760.

349. Wang, R., Kim, J., & Li, M. -. (2021). Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. Science of the Total Environment, 761 doi:10.1016/j.scitotenv.2020.144057

350. Warriner, K., Huber, A., Namvar, A., Fan, W., & Dunfield, K. (2009). Recent advances in the microbial safety of fresh fruits and vegetables. Advances in food and nutrition research, 57, 155-208.

351. Weller, D. L. (2018). Pathogens, Precipitation, Poop and Produce: The Ecology and Control of Foodborne Pathogens in Produce Preharvest Environments (Doctoral dissertation, Cornell University).

352. Weller, D. L., Love, T. M., & Wiedmann, M. (2021). Interpretability versus accuracy: A comparison of machine learning models built using different algorithms, performance measures, and features to predict *E. coli* levels in agricultural water. Frontiers in artificial intelligence, 4.

353. Weller, D., Belias, A., Green, H., Roof, S., & Wiedmann, M. (2020a). Landscape, water quality, and weather factors associated with an increased likelihood of foodborne pathogen contamination of New York streams used to source water for produce production. Frontiers in sustainable food systems, 3, 124.

354. Weller, D., Brassill, N., Rock, C., Ivanek, R., Mudrak, E., Roof, S., ... & Wiedmann, M. (2020b). Complex interactions between weather, and microbial and physicochemical water quality impact the likelihood of detecting foodborne pathogens in agricultural water. Frontiers in microbiology, 11, 134.

355. Weller, D., Wiedmann, M., & Strawn, L. K. (2015). Irrigation is significantly associated with an increased prevalence of Listeria monocytogenes in produce production environments in New York State. Journal of food protection, 78(6), 1132-1141.

356. Wen, X., Chen, F., Lin, Y., Zhu, H., Yuan, F., Kuang, D., Jia, Z & Yuan, Z. (2020). Microbial Indicators and Their Use for Monitoring Drinking Water Quality—A Review. Sustainability, 12(6), 2249.

357. Whipps, J. M., Hand, P., Pink, D. A., & Bending, G. D. (2008). Human pathogens and the phyllosphere. *Advances in applied microbiology*, *64*, 183-221.

358. White, K., Dickson-Anderson, S., Majury, A., McDermott, K., Hynds, P., Brown, R. S., & Schuster-Wallace, C. (2021). Exploration of *E. coli* contamination drivers in private drinking water wells: An application of machine learning to a large, multivariable, geo-spatio-temporal dataset. Water Research, 197, 117089.

359. Whitman, R. L., Nevers, M. B., Korinek, G. C., & Byappanahalli, M. N. (2004). Solar and temporal effects on Escherichia coli concentration at a Lake Michigan swimming beach. Applied and environmental microbiology, 70(7), 4276-4285.

360. Whitman, R.L., Shively, D.A., Pawlik, H., Nevers, M.B., Byappanahalli, M.N. (2003). Occurrence of Escherichia coli and enterococci in Cladophora (Chlorophyta) in nearshore water and beach sand of Lake Michigan. Applied and Environmental Microbiology, 69, 4714–4719. https://doi.org/10.1128/AEM.69.8.4714-4719.2003

361. Wilkinson, J., Jenkins, A., Wyer, M., & Kay, D. (1995). Modelling faecal coliform dynamics in streams and rivers. Water Research, 29(3), 847-855.

362. Winfield, M. D., & Groisman, E. A. (2003). Role of nonhost environments in the lifestyles of Salmonella and Escherichia coli. Applied and environmental microbiology, 69(7), 3687-3694.

363. Won, G., Kline, T. R., & LeJeune, J. T. (2013). Spatial-temporal variations of microbial water quality in surface reservoirs and canals used for irrigation. *Agricultural water management*, *116*, 73-78.

364. Woo, S. Y., Jung, C. G., Lee, J. W., & Kim, S. J. (2019). Evaluation of watershed scale aquatic ecosystem health by SWAT modeling and random forest technique. Sustainability, 11(12), 3397.

365. Wool, T. A., Davie, S. R., & Rodriguez, H. N. (2003). Development of three-dimensional hydrodynamic and water quality models to support total maximum daily load decision process for the Neuse River Estuary, North Carolina. Journal of Water Resources Planning and Management, 129(4), 295-306.

366. World Health Organization (WHO), 2015. WHO's first ever global estimates of foodborne diseases find children under 5 account for almost one third of deaths. Geneva: World Health Organization.

367. World Health Organization (WHO), 2020. Food Safety: Key Facts. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/food-safety. Accessed October 2021.

368. Wright, K. M., Crozier, L., Marshall, J., Merget, B., Holmes, A., & Holden, N. J. (2017). Differences in internalization and growth of E scherichia coli O157: H7 within the apoplast of edible plants, spinach and lettuce, compared

with the model species N icotiana benthamiana. Microbial biotechnology, 10(3), 555-569.

369. Wright, M. N., and Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1),1 – 17

370. Wu J, Long SC, Das D, Dorner SM. (2011). Are microbial indicators and pathogens correlated? A statistical analysis of 40 years of research. J Water Health 9:265–78

371. Wu, J., Rees, P., Storrer, S., Alderisio, K., & Dorner, S. (2009). Fate and Transport Modeling of Potential Pathogens: The Contribution From Sediments 1. JAWRA Journal of the American Water Resources Association, 45(1), 35-44.

372. Wyer, Mark D., David Kay, Huw Morgan, Sam Naylor, Simon Clark, John Watkins, Cheryl M. Davies, Carol Francis, Hamish Osborn, and Sarah Bennett. (2018). Within-Day Variability in Microbial Concentrations at a UK Designated Bathing Water: Implications for Regulatory Monitoring and the Application of Predictive Modelling Based on Historical Compliance Data." Water Research X 1, 100006. https://doi.org/10.1016/j.wroa.2018.10.003.

373. Yang, Y., Yin, X., & Yang, Z. (2016). Environmental flow management strategies based on the integration of water quantity and quality, a case study of the Baiyangdian Wetland, China. Ecological Engineering, 96, 150-161.

374. Zang, C., Huang, S., Wu, M., Du, S., Scholz, M., Gao, F., Lin, C., Guo, Y., Dong, Y. (2011). Comparison of relationships between pH, dissolved oxygen and chlorophyll a for aquaculture and non-aquaculture waters. Water, Air, and Soil Pollution, https://doi.org/10.1007/s11270-010-0695-3

375. Zhang, C., Gao, X., Wang, L., & Chen, Y. (2012). Analysis of agricultural pollution by flood flow impact on water quality in a reservoir using a three-dimensional water quality model. Journal of Hydroinformatics, 15(4), 1061-1072.

376. Zhang, G., L. Ma, L.R. Beuchat, M.C. Erickson, V.H. Phelan, M.P. Doyle. (2009). Lack of internalization of Escherichia coli O157:H7 in lettuce (Lactuca sativa L.) after leaf surface and soil inoculation. J. Food Prot., 72 (2009), pp. 2028-2037

377. Zhao, L., Zhang, X., Liu, Y., He, B., Zhu, X., Zou, R., & Zhu, Y. (2012). Three-dimensional hydrodynamic and water quality model for TMDL development of Lake Fuxian, China. Journal of Environmental Sciences, 24(8), 1355-1363.

378. Zheng, J., Allard, S., Reynolds, S., Millner, P., Arce, G., Blodgett, R. J., & Brown, E. W. (2013). Colonization and internalization of Salmonella enterica in tomato plants. Applied and environmental microbiology, 79(8), 2494-2502.