A continuous, in-situ, near-time fluorescence sensor coupled with a machine learning model for highly accurate detection of fecal contamination in drinking water: Design, characterization, and field validation

by

Emily Bedell

B.S., University of Nevada, Reno, 2012

M.S., Portland State University, 2016

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Civil and Environmental Engineering

2022

Committee Members:

Evan Thomas, Chair

Joe Brown

Julie Korak

Karl Linden

Claire Monteleoni

Bedell, Emily (Ph.D., Environmental Engineering)

A continuous, in-situ, near-time fluorescence sensor coupled with a machine learning model for highly accurate detection of fecal contamination in drinking water: Design, characterization, and field validation

Thesis directed by Prof. Evan Thomas

Equitable access to reliable, affordable, and safe drinking water is essential to human health and livelihood. Globally, two billion people use a drinking water source that is contaminated with feces. Low-cost, field-deployable, near-time methods for assessing water quality are not available when and where waterborne infection risks are greatest. In this dissertation, I describe the development and testing of a novel device for the measurement of online, in-situ, and remotely reporting tryptophan-like fluorescence (TLF), making use of recent advances in deep-ultraviolet light emitting diodes (UV-LEDs) and sensitive silicon photomultipliers. TLF is an emerging indicator of microbial water quality that is associated with members of the coliform group of bacteria and therefore potential fecal contamination. After optimizing the sensor's sensitivity to 0.05 ppb tryptophan, I demonstrated the close correlation between TLF and *E. coli* in model waters and proof of principle with sensitivity of 33 CFU/100mL for lab grown *E. coli* and 10 CFU/100mL for *E. coli* in wastewater.

I characterized the sensor's behavior to multiple fluorescence quenching parameters through benchtop analysis. Fluorescence response declined with water temperature and a correction factor was calculated. Inner filter effects were shown to have negligible impact in an operational context. Biofouling was demonstrated to increase the fluorescence signal by approximately 82%, while mineral scaling reduced the sensitivity of the sensor by approximately 5%. A training and validation data set for a machine learning model was built by installing four sensors on Boulder Creek, Colorado for 88 days and enumerating 298 grab samples for *E. coli* with membrane filtration. The machine learning model incorporated a proxy feature for fouling (time since last cleaning) which improved model performance. The model was able to predict high risk fecal contamination with 83% accuracy

(95% CI: 78% - 87%), sensitivity of 80%, and specificity of 86%. A model distinguishing between all World Health Organization established risk categories performed with an overall accuracy of 64%. The sensor design combined with the highly skilled model has the ability to provide water service providers as well as individual consumers more reliable and informative data about fecal contamination risk in their drinking water. Findings to date suggest that this device represents a scalable solution for remote monitoring of drinking water supplies to identify high-risk fecal contamination in drinking water in near-time. Such information can be immediately actionable to reduce risks and would reduce cost of microbial testing greatly, improving health and wellness of consumers and enabling water service providers more access to funds that can be used to increase access to clean water.

Dedication

For my big brother, Casey. ”If you’re not having fun, you’re not doing it right.”

Acknowledgements

First and foremost I'd like to thank my advisor, Evan Thomas for his continued and unrelenting support. Thank you for pushing me when I didn't want to be pushed. Thank you for listening to me constantly complain about the pitfalls of our institutions and society. You've empowered me to stand up for what I believe is right, even when it's hard.

Thank you to my mom and dad for their unwavering and continual support during the hardest time of our lives. I wouldn't have been able to make it through this without your nurturing love.

Thank you to my partner, Marcello for making sure I am fed, correcting me when I question myself and my abilities, providing me with endless laughs, making sure I get outside, and supporting me to support myself. I could not have done this without you, even though you would disagree with that.

Thank you to Taylor Sharpe for teaching me multiple design techniques and being so patient with me every step of the way. Your input and feedback on nearly every part of the design was vital to the project's success. Our work breaks in the lab to talk about politics or social justice issues was always a treat that kept me inspired through long days.

Thank you to Olivia Harmon for supporting me in the lab work and being my right-hand person all the way through to the end. We had many long days in the lab and not one experiment went right the first time, without your perseverance and determination, we wouldn't have been able to get all the results that we did. You're also a rockstar dancer that I'll go dancing with any day or night.

Thank you to Katie Fankhauser for your input and work on the machine learning models.

Being able to work on this section of the research with an amazing data scientist like yourself was a true honor. I look forward to learning much more from you in the future, both in work and in life.

Finally, I could not have gotten through this work without the support of my lab group and friends especially Laura MacDonald, Abby Bradshaw, Chantal Iribagiza, and Katie Fankhauser. The community we worked hard to build helped me get through the hardest of days. No matter where I go, I know I'll have you all there to support me and there's nothing else I could really ask for in this life. Thank you.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

Equitable access to clean drinking water is a human right essential for human health and livelihood. Globally, the extent and impact of lack of access to clean drinking water, free of pathogens and chemicals, is unclear. As many as two billion people worldwide use a drinking water source that is contaminated with feces (Bain et al., 2014a; WHO, 2019; UNICEF/WHO, 2021). Drinking fecally contaminated water causes adverse health effects including diarrhea and stunting, particularly among children under 5 in low- and middle-income countries (Goddard et al., 2020; Clasen et al., 2007; Pickering et al., 2019). Diarrheal diseases are the fifth leading cause of morbidity for people of all ages and the third leading cause for children ages 0-9 (Abbafati, 2020). Approximately 60% of all diarrheal deaths are related to water quality globally (Prüss-Ustün et al., 2019).

Although there is a much higher incidence of diarrhea associated with an unsafe drinking water source in low- and middle-income countries, outbreaks are common in high-income countries as well (DeFlorio-Barker et al., 2021). At least 40 million people in the United States rely on domestic wells for their drinking water supply, which are not regulated by the U.S. Environmental Protection Agency (EPA) (Johnson et al., 2019). Between 1971 and 2008, 30% of all waterborne outbreaks in the U.S. were associated with the consumption of untreated groundwater (Craun et al., 2010). The EPA estimates that approximately 16.4 million cases of acute gastrointestinal illness are due to unsafe drinking water in the United States annually (Messner, 2006).

Presently, fecal contamination in drinking water poses large threats to public health, globally. Moving forward, these threats may become more severe as extreme whether events increase in intensity and frequency with climate change (Coumou and Rahmstorf, 2012). Both flooding and drought have the potential to cause disease outbreaks related to drinking water quality (Nichols et al., 2018). Power outages resulting from extreme weather events leading to failure in existing treatment or lack of treatment is a leading cause of waterborne outbreaks in Canada and the United States (Pons et al., 2015).

1.1 Microbial Water Quality Monitoring

Microbial contamination is an enormous challenge for water service providers for which rapid detection is critical to prevent microbial outbreaks. A majority of findings and studies on the extent of fecal contamination in drinking water and it's impact on human health share a common conclusion: there is an enormous lack of data. In both academic research and professional practice, estimates of contamination are mostly based on infrequent point measurements with small sample sizes (Bain et al., 2014a). Current methods to detect microbial contamination fall into two categories. The first is counting the number of unspecific bacteria in the water, referred to as heterotrophic plate counts (HPC). The second principle is to use indicator microorganisms. Fecal indicator organisms are microorganisms that exist together with pathogens, originate in the mammalian intestinal tract, are specific to this environment, and are proportional to the amount of fecal contamination (Skovhus and Højris, 2018). The indicator organisms for drinking water quality preferred by the WHO are thermotolerant coliforms (TTCs). One thermotolerant species, *Escherichia coli (E. coli)*, has been identified as the best indicator for fecal matter because of its nearly exclusive residence in the mammalian intestinal system and consistent presence in fecally contaminated sewage, natural waters, and soils (WHO, 2017). In freshwater, at least 80% of the total TTCs enumerated are *E. coli* and the two are sometimes used interchangeably in water quality monitoring (Hachich et al., 2012). However, indicator organisms are not necessarily pathogenic and, instead, are usually only a gauge of the likelihood of fecal contamination, and thereby, the presence of other enteric pathogens.

The products currently available for *E. coli* enumeration fall under three categories: Presence–Absence, Most Probable Number (MPN), and Colony Counting (Bain et al., 2012). All of these tests require retrieving a sample, typically from a point of water collection or household consumption, along with follow-up laboratory analyses. Microbial enumeration tests, on average, cost approximately $21 to conduct including consumables, equipment, lab, and logistics (Delaire et al., 2017). All of these tests also require 18–48 hours of incubation before the results can be analyzed. This limitation is potentially dangerous: by the time contamination is detected, consumers may already be exposed (Besner et al., 2011). Other important barriers of traditional testing methods include high initial costs, extensive training requirements, inability to provide information on the source of contamination, and a high probability of missed contamination events (Sorensen et al., 2018a). These constraints often prevent service providers from maintaining accurate and precise measures of microbial contamination, further harming their ability to validate their treatment processes, assess the quality of source waters, perform operational and routine monitoring, and provide verification of the quality of their end product (WHO/OECD, 2002). Since fecal contamination of water sources is highly variable temporally, seasonally, and associated with extreme weather events, studies and water service providers often underestimate contamination and exposure rates (Kostyla et al., 2015). These limitations not only cause unknown exposure to fecal contamination, but also erroneous predictions of contaminated water's effect on diarreal incidence globally (Bain et al., 2014b).

The desire to detect fecal contamination in drinking water should have one primary goal: to reduce risk of outbreaks that are harmful to human health. The World Health Organization (WHO) has grouped *E. coli* contamination concentrations into five risk categories: low (1-9 CFU/100 mL), intermediate (10–99 CFU/100 mL), high (100–999 CFU/100 mL), and very high (>1000 CFU/100 mL) (WHO, 2017). These risk categories are not a measurement of health risk, but instead a risk that the water actually contains fecal contamination. *E. coli* is not a perfect indicator of fecal contamination and its use to predict pathogenic contamination has been debated for decades. Not all *E. coli* strains are pathogenic or harmful to human health. *E. coli* has low specificity (ability to detect only pathogenic contamination) and sensitivity (ability to detect any amount of pathogenic contamination) and does not assess whether pathogenic contaminates that are present are viable Krewski et al. (2004). The association between *E. coli* counts in drinking water and diarreal disease is often weak and variable Gruber et al. (2014); Brown et al. (2008). Nevertheless, there is evidence that exposure to *E. coli* is associated with negative health outcomes (Gruber et al., 2014). Quantitative microbial risk assessment (QMRA) has been used widely as a tool for estimating the risks associated with exposures to pathogens in the environment. QMRA generally uses a dose-response model that predicts the probability of infection given a dose exposure magnitude (Haas et al., 2014). Studies on the association between TTCs in drinking water and diarrhea show a dose-response effect consistent with the WHO risk categories (Hodge et al., 2016). There is very little evidence of increased odds of diarrhea with contamination levels between 1 and 10 CFU/100mL *E. coli*, the category designated as ”low-risk” Hodge et al. (2016). This fact, combined with the fact that concentrations of indicator organisms can be highly variable in surface waters, makes establishing a fecal contamination risk level, in terms of health, extremely complex. Large spikes of contamination can be due to rainfall, defecation directly into source waters, or washing of contaminated items like diapers (Levy et al., 2009). The size and frequency of these spikes in contamination can be highly variable and will often be missed by traditional microbial water testing (Enger et al., 2013). Multiple studies show that compared to consistent low risk levels of fecal contamination, highly concentrated spikes of fecal contamination are of highest risk to human health, yet there is currently no way to detect them in real-time (Daly and Harris, 2021; Brown and Clasen, 2012; Haas and Betz, 1996).

Instrumentation used to identify other potential indicators of fecal contamination, including residual chlorine or turbidity, also have significant limitations. Residual chlorine presence in drinking water is used to indicate an absence of most disease-causing organisms. For detecting residual chlorine, amperometric and colorimetric sensors have been widely used in water treatment and distribution systems. Amperometric sensors use electrodes to measure a change in current caused by the chemical reduction of hypochlorous acid. These sensors are very sensitive to changes in the water’s pH and thus require frequent recalibration, which has an impact on the sensor’s accuracy over time (Malkov, 2009). Colorimetric sensors rely on chemical reagents to react with the residual chlorine in the water, and quantify the amount of residual chlorine using a spectrophotometer.

The requirement of constant dosing with reagents makes these sensors difficult to employ remotely and autonomously Mohtasebi et al. (2017). Turbidity is the measurement of cloudiness of water. Multiple studies have shown a correlation between turbidity and *E. coli*" measurements (Vidon et al., 2008; Dorner et al.; Money et al., 2009). There are various methods and instrumentation for measuring turbidity and they do not often produce comparable results. Depending on the size and make-up of the particles in the water, different measurement methods may produce conflicting results (**?**). The turbidity sensors that are available for in-line, continuous monitoring are expensive and require site- and device-specific calibration, limiting their potential for low-cost, remote use (Gillett and Marchiori, 2019).

1.2 Tryptophan-like Fluorescence

Using fluorescence spectroscopy to measure tryptophan-like fluorescence (TLF) in drinking water has shown potential to address many of the challenges presented by traditional fecal contamination monitoring methods. Fluorescence occurs when a molecule absorbs a photon from light at a specific wavelength, known as the excitation wavelength. The absorbed photon causes a fluorophore to jump to a higher energy state (S_2), then rapidly relaxes to the lowest vibrational state (S_1) (Fig. 1.1) . Once the fluorophore drops back down to its ground electronic state (S_0), it emits a photon at another, higher, specific wavelength, known as the emission wavelength (Lakowicz). TLF has an excitation wavelength (λ_{ex}) of 275 nm and an emission wavelength (λ_{em}) around 260 nm (Coble et al., 1991). TLF measurements rely on the intrinsic fluorescence of the aromatic amino acid, tryptophan, whose presence is often related to microbial activity (Fox et al., 2017). Compared to the other two aromatic amino acids (phenylalanine and tyrosine), tryptophan is the dominant intrinsic fluorophore, absorbs light at the longest wavelength, and displays the largest extinction coefficient. Microbial activity also converts tryptophan to indole which enhances its fluorescence output because indole exhibits a similar fluorescent signature to tryptophan and fluoresces at approximately 33% greater intensity (Lakowicz; Sorensen et al., 2018b). E. coli produces indole from lactose and tryptophan causing it to have the highest TLF per occurrence compared to other bacteria (Cumberland et al., 2012). A major challenge with associating TLF with bacterial concentrations in drinking water is the presence of multiple compounds emitting TLF. TLF can be contained within bacterial cells as well as be associated with particles or entirely freely dissolved (Baker et al., 2007; Sorensen et al., 2016). This leaves uncertainty over what is actually being measured.

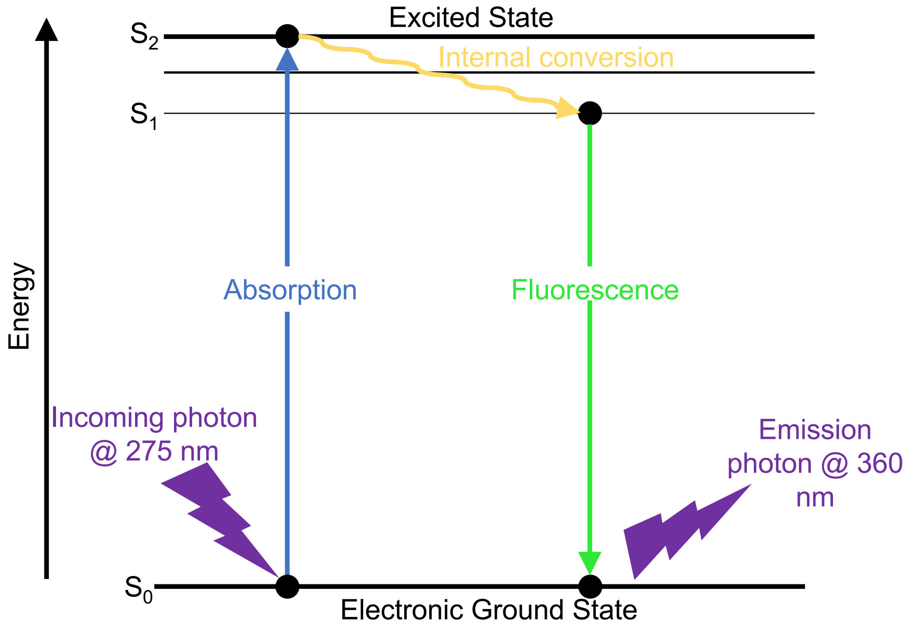


Figure 1.1: Jablonski Diagram diagram demonstrating tryptophan-like fluorescence with an excitation wavelength of 275 nm and an emission wavelength of 360 nm

The presence of TLF in water can indicate that tryptophan is present on its own as ’free’ molecules or bound in proteins, peptides, or humic structures where microbial activity is occurring(Hudson et al., 2008). Tryptophan and its derivatives may be intracellular or extracellular, both of which will display TLF. If intracellular, they are expected in bacteria as structural and functional proteins used in metabolic pathways, metabolic byproducts, and endospore formation. As extracellular molecules, they would be found as secreted signaling molecules, metabolic byproducts, exotoxins, and cellular debris (Fox et al., 2017). Tryptophan contains an amine group, a carboxylic acid group, and a side chain indole group. The indole moiety allows tryptophan to fluoresce, making it one of three known amino acids (in addition to tyrosine and phenylalanine) with this property(Hudson et al., 2007). While indole is widespread in nature, high concentrations are found in mammalian intestinal tracts and wastes (feces). Extracellular indole concentrations have been measured as high as 0.5 mM in suspended cultures of *E. coli*(Lee et al., 2015). The intracellular and/or extracellular nature of TLF seems to depend on the type and source of the water. Sorensen et al. (2020) found that TLF in groundwater was predominately extracellular (96%) after measuring TLF levels before and after filtration at 0.22 μm Sorensen et al. (2020). Fox et al. (2017) found that the majority of TLF signal (75%) is intracellular in origin from lab grown bacteria inoculated in media (Fox et al., 2017). Since the yield of indole depends on the amount of exogenous tryptophan, it's logical that in natural waters where extracellular tryptophan is more available, TLF would be more extracellular than in lab grown settings Li and Young (2013).

1.3 Technology Development Opportunities

In 2017, the United Nations Children's Fund (UNICEF) identified real-time, in situ E. coli detection as a "target product" for research, development and ultimately UNICEF procurement. The Target Product Profile (TPP) presents minimum performance requirements for such a product which include being battery-based, minimal processing requirements, no need for reagent mixing or incubation, qualitative output based on quantifiable ranges of fecal contamination, the ability to sample a variety of water sources, and sensitivity and specificity goals equating to the ability to detect 10 colony forming units (CFUs) per 100 mL, with false positives and negatives below 10%. The product is required to have a detection time of less than 6 hours, a two-year minimum lifespan, and must be portable (UNICEF, 2019). More specific requirements are shown in Table 1.1.

Table 1.1: UNICEF Target Product Profile Requirements

Attribute	Acceptable Performance	Ideal Performance
Key Function	Detection of fecal contamination equivalent to *E. coli* in water	
Power Requirements	Portable power source or no power requirement	
Performance	FPR <10%, FNR <10%, over range of concentrations	
Life Span	2 years minimum, no cold chain required	

Continued on next page

Table 1.1: UNICEF Target Product Profile – continued from previous page

Attribute	Acceptable Performance	Ideal Performance
Performance Requirements		
Level of detection	Difference between presence/absence as well as low/moderate (1-100 CFU/100mL) and high levels (>100 per 100 mL)	Differentiation across four risk levels (0, 1-10, 11-100, >100 CFU per 100 mL)
User Requirements		
Testing methodology	Minimum number of process steps, rapid incubation allowed, preferred at room/body temperature	Minimum number of process steps, no reagent mixing required, no incubation required
Materials Used	Waterproof as a packaged product and durable for transportation	
Validation Requirements		
Time to Result	Less than 6 hours	Less than 30 minutes
Core Requirements		
Target Unit Price	Between $1,001 and $6,000	Up to or below $1,000
Field Test Requirements		
Number of samples in a day	5-10	More than 10
Presentation of results	Qualitative through clear visual cues or text based on quantifiable ranges	Quantified results as number/text or allow for simple quantification

The UNICEF TPP offers stringent design constraints, but they are generally for a portable detection device or kit, not a device that is continuously monitoring. An online sensor offers much higher sampling frequency, possibly continuous monitoring, provides an opportunity to detect short term or pulse contamination. Grab samples often miss these contamination events because they only last a short period, but they can still cause disease in people drinking the contaminated water (US EPA, 2017). Sensors can provide a more rapid response to contamination which will allow water service providers to be more proactive and less reactive in response to contamination. Another benefit is the strategic placement of sensors for locations that are critical or indicative of problems within a distribution system or treatment plant. Placing sensors in the distribution system near sensitive consumers like hospitals or food production plants could also be beneficial (Skovhus and

Højris, 2018).

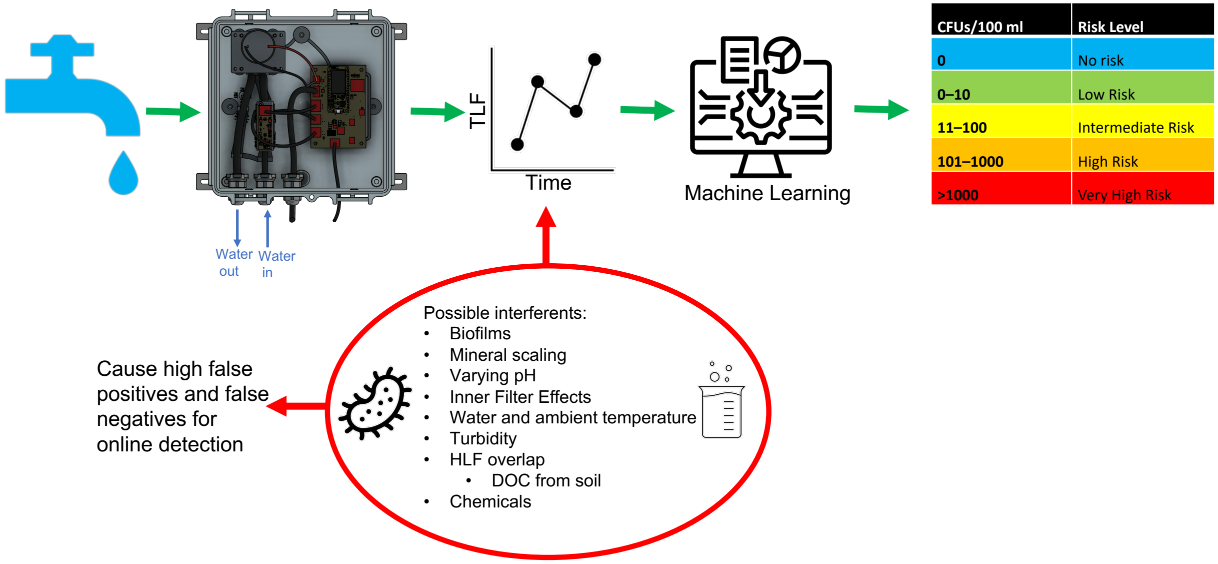
Several highly sensitive, portable TLF sensors are available, but they are not designed or marketed for autonomous, in-situ, continuous operation. Their output is either in arbitrary units (AU) or ppb of tryptophan values. One challenge facing these in-situ TLF sensors is their sensitivity to parameters outside of those they are attempting to measure, which can impact their output. Temperature, pH, and turbidity can all influence a TLF signal, although these impacts are most significant outside of typical target detection ranges. Increasing water temperature will reduce fluorescence intensity, signal quenching can be up to 15% in waters with low pH (below 4.5), and turbidity may have varying impacts on the signal depending on the contents of the water (Baker et al., 2007; Guilbault, 1973; Reynolds, 2003). Another limit to reading absolute value outputs from a TLF sensor is background noise in aquatic environments, which is highly variable, potentially causing a high number of false-positives. In addition, proteins in organic waste, xenobiotic compounds, and diesel pollution have all been shown to fluoresce in the TLF region (Sorensen et al., 2018b). In drinking water, there are multiple peaks of fluorescent dissolved organic matter (DOM) that overlap with TLF and could give rise to an apparent TLF signal. Humic-like fluorescence (HLF) can raise the baseline TLF for samples with low TLF, also increasing false positives (Ward et al., 2020).

Other challenges to continuous monitoring with a TLF sensor primarily include biofouling, mineral scaling, and baseline drift that attenuate and ambiguate the signal. The extent of biofouling varies as a function of the type and contents of water. Typically, frequent cleaning and calibration must be performed at each sample site to mitigate any buildup of biofouling or scaling on sensor lenses or windows. Common anti-biofouling mechanisms include the use of shutters, plates, and wipers which all require frequent maintenance (Coble et al., 2014).

Another approach to reduce signal drift, false positive rates, and false negative rates is through machine learning (ML) based synthetic calibration to enact noise-reduction and anomaly detection that enables alarm-threshold detection (Joslyn and Lipor, 2018). Through a combination of low-cost, robust hardware design and ML, it may be possible to address the signal drift limitations as well as improve upon the sensitivity (true positive rate) and specificity (true negative rate) of a sensor system by leveraging a combination of in-situ and remotely sensed data, trained and validated with manually collected ground-truth data. ML would allow for the integration of site and sensor-specific baseline calibration and the detection of sudden and large changes in TLF rather than relying on a universal threshold for all sensors, as well as input from other important data sources like seasonality, nearby sensor data, and data features serving as proxies for fouling.

In this dissertation, I design and validate an in-situ, near-time, remotely reporting TLF sensor system coupled with an ML model for the detection of fecal contamination risk in drinking water. Following suggestions and thresholds established in the literature, the sensor system was designed to differentiate among WHO designated risk levels correlated to FIB concentration. The risk categories are low (1-9 CFU/100 mL), intermediate (10–99 CFU/100 mL), high (100–999 CFU/100 mL), and very high (>1000 CFU/100 mL) (WHO, 2017). TLF measurements are best utilized to distinguish between these risk categories rather than providing a direct enumeration of microbial contamination (Ward et al., 2020). Sorensen et al. 2017 showed that a dissolved tryptophan concentration of 1.3 ppb could predict the presence of FIBs with a false-positive error rate of 18% and a false-negative error rate of 15%, thus the design limit of detection was set to 1 ppb dissolved tryptophan (Sorensen et al., 2018b). A schematic is shown in Fig. 1.2 to demonstrate the work needed to complete this dissertation including the design of the sensor, characterization of possible signal interference, and development of a machine learning model.

Along with the design and validation of the sensor system, I report on the characterization of multiple potential fluorescence quenching and interference parameters that may influence the operational limits of a continuous TLF sensor system. We also examine the impact of chlorine on the sensor signal when dosed in both contaminated and uncontaminated waters to provide some insight into whether the sensor is detecting live or dead cells. This study also investigates how the the sensor was integrated with a ML data system to compensate for biofouling, scaling, and background fluorescence noise. Finally, the sensor and ML system were deployed and validated in a field implementation.



CFUs/100 ml	Risk Level
0	No risk
0–10	Low Risk
11–100	Intermediate Risk
101–1000	High Risk
>1000	Very High Risk

Figure 1.2: Flow chart displaying the overall goals of this research. The design of a sensor coupled with a machine learning model that can detect anomalies and reduce noise caused by interferents in an online, continuously reporting TLF sensor to detect fecal contamination risk in drinking water

1.4 Dissertation Synopsis

In Chapter 1, I provide an overview of the motivation and background for the research, technology development, and field deployment. In Chapter 2, I describe the technology development and prototype iterations for optimizing sensitivity of the sensor along with sensitivity validation in a lab setting. In Chapter 3, I present results of lab characterization of varying water quality's impact on the sensor's signal as well as how the sensor reacts to biofouling and mineral scaling. In Chapter 4, I present the results of a field validation of the sensor's functionality along with development of a machine learning model. In Chapter 5, I summarize the study insights and conclusions.

Figure 1.3 presents a summary and flow chart of the research. The figure shows the design aspects, design requirements, experiments, output, publications, and timeline for each research question.

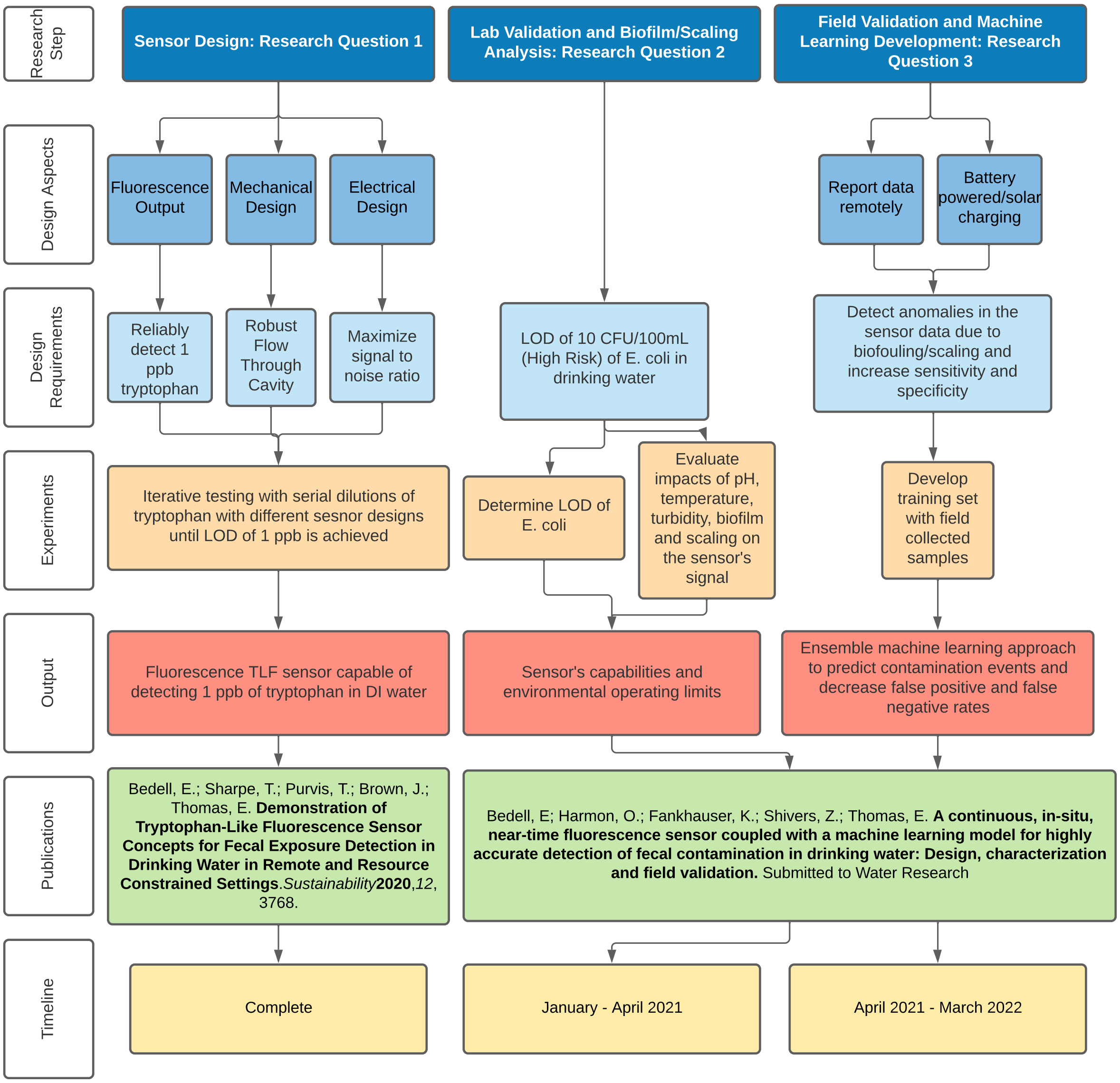

Research Step
Sensor Design: Research Question 1
Lab Validation and Biofilm/Scaling Analysis: Research Question 2
Field Validation and Machine Learning Development: Research Question 3
Design Aspects
Fluorescence Output
Mechanical Design
Electrical Design
Report data remotely
Battery powered/solar charging
Design Requirements
Reliably detect 1 ppb tryptophan
Robust Flow Through Cavity
Maximize signal to noise ratio
LOD of 10 CFU/100mL (High Risk) of E. coli in drinking water
Detect anomalies in the sensor data due to biofouling/scaling and increase sensitivity and specificity
Experiments
Iterative testing with serial dilutions of tryptophan with different sesnor designs until LOD of 1 ppb is achieved
Determine LOD of E. coli
Evaluate impacts of pH, temperature, turbidity, biofilm and scaling on the sensor's signal
Develop training set with field collected samples
Output
Fluorescence TLF sensor capable of detecting 1 ppb of tryptophan in DI water
Sensor's capabilities and environmental operating limits
Ensemble machine learning approach to predict contamination events and decrease false positive and false negative rates
Publications
Bedell, E.; Sharpe, T.; Purvis, T.; Brown, J.; Thomas, E. Demonstration of Tryptophan-Like Fluorescence Sensor Concepts for Fecal Exposure Detection in Drinking Water in Remote and Resource Constrained Settings.Sustainability2020,12, 3768.
Bedell, E; Harmon, O.; Fankhauser, K.; Shivers, Z.; Thomas, E. A continuous, in-situ, near-time fluorescence sensor coupled with a machine learning model for highly accurate detection of fecal contamination in drinking water: Design, characterization and field validation. Submitted to Water Research
Timeline
Complete
January - April 2021
April 2021 - March 2022

Figure 1.3: Dissertation work flowchart: Design, Characterization, and Field Validation of a continuous, in-situ, near-time fluorescence sensor coupled with a machine learning model for highly accurate detection of fecal contamination in drinking water.

1.5 Research Goals and Methods

This work evaluates the functionality and ability of an in-situ, continuously reporting fluorescence sensor to monitor fecal contamination in water. The design of the sensor, impacts to the sensor's signal from quenchers and interferents, limit of detection of the sensor, field validation, and development of a machine learning model is described.

Research Objective 1: Design and test a fluorescence sensor capable of detecting 1 ppb tryptophan

To investigate this question, fluorescence optimization was performed in order to achieve a measurable emission of 1 ppb of tryptophan. This output was optimized by sourcing high power LEDs and highly sensitive light detectors as well as optimizing optical components like lenses, bandpass filters, and windows. A robust and reliable electrical design was built to maximize the signal to noise ratio (SNR). The design aspects, fluorescence output, electrical, and mechanical, were considered individually and together as a whole in order to achieve optimization of sensitivity along with minimization of cost. The detection limits of the sensor in terms of lab grown *E. coli* and *E.coli* present in wastewater were determined.

RQ1.1: Can an optical TLF sensor be designed at much lower cost than currently available on the market to achieve a minimum detection limit (MDL) sufficient to reliably detect high risk fecal contamination in drinking water?

Hypothesis: Yes, with the appropriate combination of high-powered UV LEDs pulsing at a high frequency, a high sensitivity photomultiplier, and a voltage amplifier

RQ1.2: What is the detection limit of the sensor in terms of lab grown E. coli and wastewater dilutions?

Hypothesis: The detection limit for E. coli will be around 10 CFU/100 ml in wastewater dilutions, and around 10^3 CFU/100mL of lab grown *E. coli.*

Research Objective 2: Determine the sensor's feasibility in regard to variability in water quality, and determine how biofilms and mineral scaling impact the sensor's signal

To investigate this research objective, the sensor was subjected to various changes in water quality including temperature, pH, turbidity, and chlorine. A study was conducted on biofilm formation, mineral scaling, background noise variability and the subsequent impact on sensitivity and detection limits.

RQ2.1: How will pH, temperature, and turbidity affect the fluorescence output? What correction factors or operational limits will need to be set?

Hypothesis: Increasing temperature will reduce the fluorescence signal, low pH (below 4.5) will reduce the fluorescence signal, turbidity will have varying impacts depending on the content of the water. A temperature sensor will be added to the system. The sensor will not perform adequately for water with turbidity ¿ 10 NTUs or pH outside of the range 5-9.

RQ2.2: How does the formation of biofilms and mineral scaling impact the sensor's signal?

Hypothesis: Biofilms will interfere with both excitation and emission intensity causing the signal to increase, mineral scaling will cause the signal to decrease.

Research Objective 3: Develop a machine learning model to mitigate signal drift as well as improve on the sensor's sensitivity and specificity

A training set was built by monitoring four sensors on Boulder Creek for 13 weeks. This data was used to building a machine learning model to predict fecal contamination risk level with the sensor output as the primary informer.

RQ3.1: Can ML be implemented to mitigate the effects biofilms and increase the sensitivity and specificity of the sensor?

Hypothesis: Ensemble ML will be developed to predict contamination events and improve sensor's overall ability to detect contamination.

Chapter 2

Sensor Design and Limit of Detection Validation

Part of this chapter appears in *Sustainability*, with the following authorship attribution: Emily Bedell, Taylor Sharpe, Timothy Pruvis, Joe Brown, and Evan Thomas

2.1 Introduction

The Joint Monitoring Program for Water Supply and Sanitation (JMP) estimates that globally, at least 2 billion people use a drinking water source that is contaminated with fecal matter (WHO, 2019). Drinking water containing fecal contamination is a leading cause of preventable diseases and higher mortality, particularly through diarrheal infections, which overwhelmingly affect children under five in low- and middle-income countries (WHO, 2005). Monitoring fecal contamination in drinking water supplies is a critical function of water service providers. In low- and middle-income countries, service providers often cannot afford the monitoring technologies proven to provide robust data. While the World Health Organization's (WHO) Guidelines for Drinking Water Quality have been adopted by most water service providers globally, microbial water quality testing needed to analyze the fecal contamination risk is conducted infrequently (Delaire et al., 2017). This is because testing routines are time-consuming, expensive, require trained personnel and consumables, and compete for resources.

Multiple recent studies have shown TLF measurements can be used as an alternative or additive risk assessment tool to traditional microbial testing methods. Strong correlations are shown between the presence of TLF and that of heterotrophic bacteria, E. coli, and total coliforms in drinking water (Sorensen et al., 2015; Baker et al., 2015; Nowicki et al., 2019). Nowicki et al. (2019) showed that TLF is “precise, rapid, and practical for groundwater sampling”, but that it should not act as a proxy for E. coli measurements (Nowicki et al., 2019). The ability of the instruments tested to produce exact correlation to traditionally used plate counts is limited because of noise in the output signal as well as the detection limit of the sensor. TLF is shown to have greater success distinguishing between the established WHO microbial risk levels (Sorensen et al., 2018b). The WHO has established decimal categories of potential health risk related to E. coli or thermotolerant coliform (TTCs) concentrations. These risk categories are low (1-10 CFU/100 mL), intermediate (10–100 CFU/100 mL), high (100–1000 CFU/100 mL), and very high (>1000 CFU/100 mL) WHO (2017). This approach has the potential to improve risk assessment of microbial contamination in drinking water, especially when coupled with traditional methods.

Most established literature only examines and demonstrates these theories with submersible or cuvette-based, portable fluorimeters that are currently available on the market. Sorensen et al. (2018) compiled data from recent studies analyzing the ability of both submersible (the UviLux from Chelsea Technologies Group Ltd., UK) and cuvette-based (the SMF4 from STS Instrument Ltd., UK) fluorimeters to indicate fecal contamination risk in drinking water. The study found that TLF had the ability to classify high-risk sources containing ¿10 CFUs/100 mL. This coliform threshold was correlated with a tryptophan concentration of 1.3 parts per billion (ppb) (Sorensen et al., 2018b). The sensors analyzed were not intended for long-term, autonomous operation and were expensive, ranging from $4000–$10,000. Recent advances in the semiconductor industry are quickly driving down the costs of high sensitivity and high-power components included within these sensors, namely deep-UV light emitting diodes (LEDs) and sensitive semiconductor photodiodes and photomultipliers. The only attempt at the design of a low-cost, flow-through fluorimeter sensor for fecal contamination detection in drinking water was conducted by Simões and Dong (2018) (Simões et al., 2021). In that study, a sensor design was presented with a limit of detection of 1.4 x 103 CFU/mL E. coli.

All fluorescence signal can be subjected to inner filter effects (IFE) at high organic matter (OM) concentrations. IFE occurs when a portion of the excitation light is absorbed before it reaches the point in the sample where fluorescence occurs, this is known as primary IFE. Re-absorption can also occur with a portion of the light emitted from the fluorophore before it reaches a detector, known as secondary IFE (Kubista et al., 1994). IFEs result in a attenuation of fluorescence signal. If quantitative results of concentration are desired, corrections must be made for IFEs, if a qualitative result of risk level of concentration is desired, corrections may not be necessary.

In this chapter, I describe the design and validation process of a flow-through, in-situ, continuously monitoring TLF sensor. In order to achieve high sensitivity from the sensor, careful attention had to be paid to each design aspect including: electrical, mechanical, and optical component optimization. The sensor also needed the capability to operate remotely and send data through cellular networks to an online database.

2.2 Methods

2.2.1 Initial Designs

An iterative design approach was conducted in order to initially meet the design criteria of significant sensor signal sensitivity to 1 ppb tryptophan. The sensitivity of the instrument was optimized by completing successive experiments with differing components and placement. First, a static system was designed and optimized using a UV quartz cuvette (FireflySci) containing the sample water. Components were held in place by 3D printed parts designed in Autodesk's Fusion360 and printed on a Formlabs Form2 stereolithography printer. The initial prototype components include a UV-LED centered around 280 nm (Marktech Optoelectronics) with no secondary optical bandpass filter; a UV Photodiode centered around 350–360nm (Marktech Optoelectronics), a double converging lens with UV transmission coating (Edmund Optics), and an emission bandpass filter (Edmund Optics). This initial design is shown in Fig 2.1.

The sensitivity was initially tested by mixing solutions of 1, 10, 100, 10000, 100000 ppb

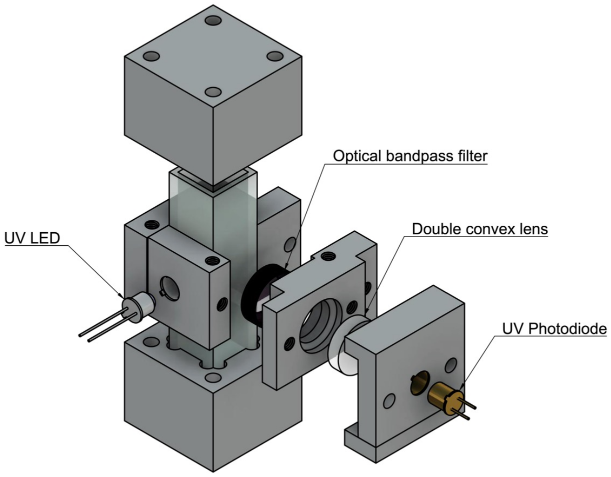


Figure 2.1: Initial prototype design assembly. A stand alone, cuvette-based unit was designed to compare to the bench top unit. Components were held in place by 3D printed parts.

L-tryptophan in deionized (DI) water. The concentrations were mixed using serial dilutions and measured on the same day of mixing. Each concentration was pipetted into the quartz cuvette, placed in the prototype, the LED was powered using a regulated power supply and 100 readings were taken from the photodiode by a Keithley 6485 Picoammeter. The magnitude of outputs from the picoammeter were on the order of 10^{-12} amps. Readings were averaged and the noise level (reading at DI water) was subtracted from each measurement. Signal optimization was attempted by successive experiments with differing components and placement.

The first component and configuration tests included four different iterations. First, the original prototype, second, moving the emitter back to its focal length, third, placing a concave mirror across from the emitter, and fourth, placing a concave mirror across from the detector.

The prototype did not consistently produce a significant change in signal at the desired sensitivity in the 1–10 ppb range. A hypothesis was formed that increasing the number of emitters would increase the sensitivity by increasing the signal-to-noise ratio (SNR). Experiments were then run using 1, 2, and 4 LEDs of the same type and the SNR was compared at 100 ppb L-tryptophan measurements. These results were also compared to the SNR produced using two high-powered (100mA) LED's along with a higher sensitivity photodiode (Table 2.1).

Table 2.1: Signal to Noise Ratio (SNR) for different LED configurations and combined with a higher sensitivity photodiode

	One LED	**Two LEDs**	**Four LEDs**	**2 High Power SMD LEDs**	**2 High Power SMD LEDs, Higher Sensitivity Photodiode**
SNR for 100 ppb tryptophan	2.24	3.87	6.48	8.04	10.04

Using two high-powered LEDs consistently showed the sensitivity needed at lower concentrations - 1, 3, and 10 ppb of tryptophan – sufficient to move forward with the flow-through design.

Once the desired sensitivity was achieved on a static, cuvette-utilized setup, a flow-through unit was designed (Fig. 2.2). The preliminary flow-through design used three 12.5 mm diameter quartz windows for the LEDs and photodiode to fluoresce and detect the sample. The design incorporates O-rings to seal the sample flowing through a 12 mm x 12 mm cavity. The flow-through prototype was further developed by adding a femtoampere input bias current electrometer amplifier (Analog Devices) to convert the output signal from the photodiode from picoamps to millivolts. Water was pumped through the prototype using a peristaltic pump at a flow-rate of 10 mL/s and the sensitivity of the design was tested using 1, 3, 10, 100 ppb L-tryptophan solution.

The device configuration with 2 UV LEDs and a high sensitivity photodiode combined with an amplifier greatly increased the sensitivity, giving a significant signal change between DI water and 1 ppb tryptophan (Fig. 2.3. There are significant differences ($p<0.01$) between DI and 1 ppb and DI and 3 ppb, but not between 1 ppb and 3 ppb. This lab setup is modeled in Fig 2.4.

At this stage, the sensor's output was compared to a Horiba FluoroMax-4 benchtop fluorimeter (Fig. 2.5). An emission scan was performed on the FluoroMax-4 with an excitation wavelength of 275 nm and emission wavelengths from 300-400. The emission counts per second (CPS) at 340 nm were compared to the prototype output voltage from the prototype.

Multiple challenges arose with the flow through design. The windows weren't able to apply enough pressure to the o-rings to create a good seal, causing multiple leaks. The electromagnetic

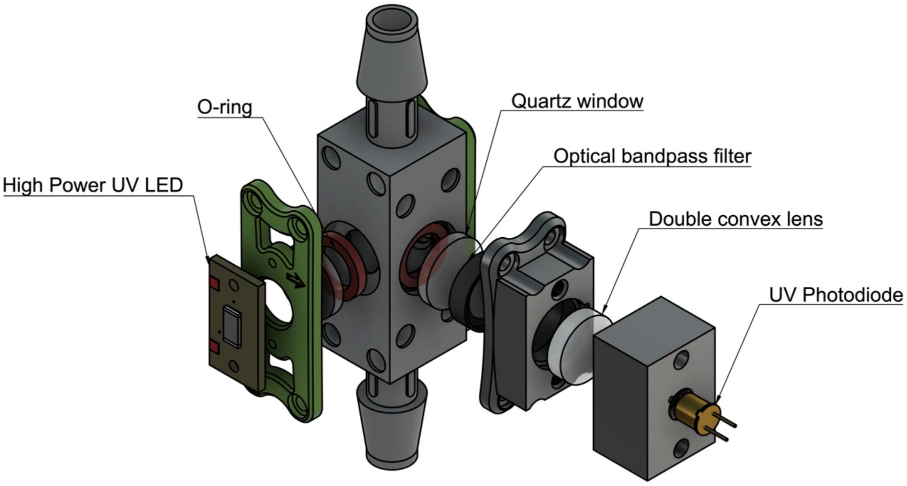


Figure 2.2: Flow-through prototype design. A flow-through design was constructed to test the potential of continuous and autonomous monitoring. Components were assembled using 3D printed parts and fasteners. A water-tight seal was provided by applying pressure on O-rings with a UV coated quartz window.

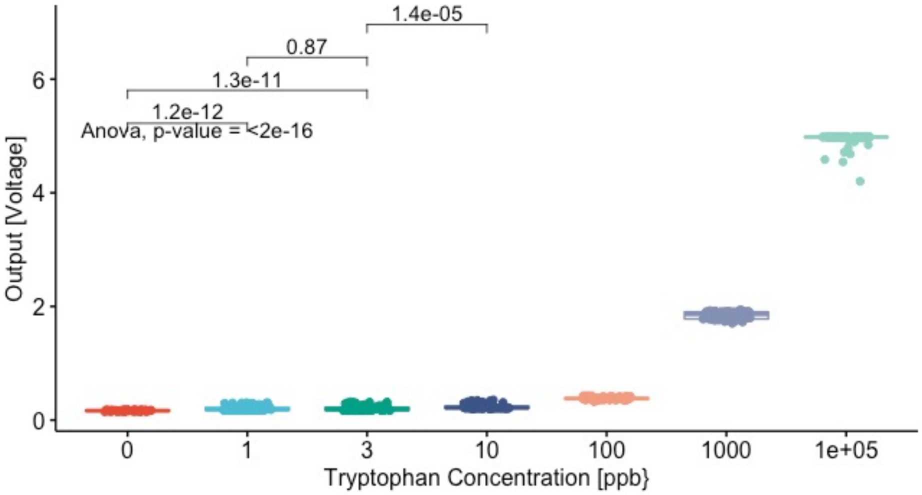


Figure 2.3: First flow through prototype response from tryptophan dissolved in deionized water. Boxes indicate the interquartile range and median, whiskers indicate maximum and minimum values except where outliers are indicated. Bars indicate the significant differences between indicated concentrations calculated by a Students t-test. Analysis of Variance (ANOVA) p-value shows the difference from DI to 1 ppb, DI to 3 ppb, 1 ppb to 3 ppb, and 3 ppb to 10 ppb.

sensitivity of the photodiode also caused very high noise in the signal. For these two reasons, the next prototype design consisted of a flow through cuvette and a silicone photomultiplier. A Broadcom silicon photomultiplier (SiPM) (www.broadcom.com) was used to detect the fluorescence

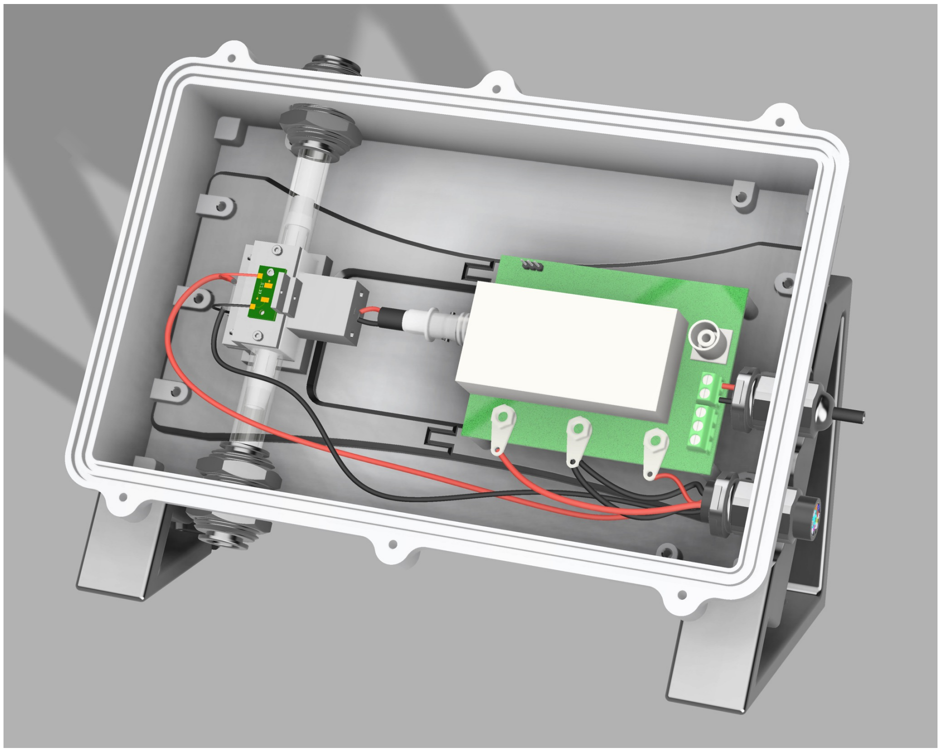


Figure 2.4: Lab setup flow through design. Piping shown where water flows in and out of the sensor. Two high powered UV LEDs excite the fluorophore and its emission is read by a high sensitivity photodiode. The signal is converted from current to voltage using an op-amp with high gain.

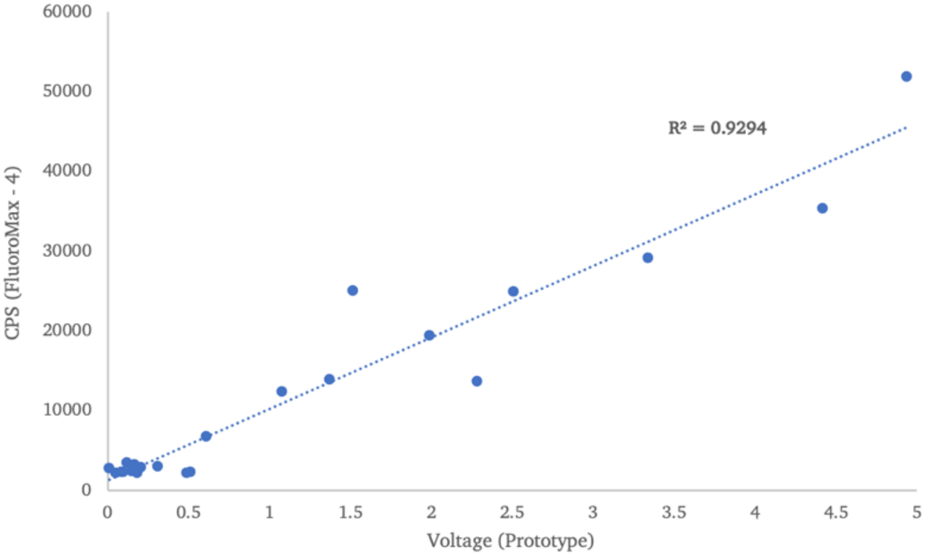


Figure 2.5: Prototype comparison to Horiba FluoroMax 4 benchtop fluorimeter

emissions. An SiPM is a highly sensitive, solid state photodetector made of an array of single-photon avalanche diodes (SPADs) connected in parallel. Compared to photomultiplier tubes or photodiodes, which are used in most high sensitive TLF sensors, SiPMs are less vulnerable to magnetic fields, more robust and compact, and require low bias voltage and power (Broadcom, 2019).SiPMs are capable of very high ($10^5 - 10^7$) internal gain (μ).

Once lab testing was done on the flow through cuvette, SiPM configuration, printed circuit board (PCB) design began to move the sensor from a lab based unit powered by a bench-top powersupply to a field based unit that could operate autonomously. The following design goals were set for the lab based unit:

(1) Autonomous, in-situ functionality

(2) High sensitivity to tryptophan concentrations

(3) Simple maintenance and cleaning of optics

(4) Range and gain levels applicable to different natural and treated waters

(5) Detect and measure proxies for fouling as features for the machine learning model

2.2.1.1 Electrical Design of Field Unit

An amplifier board was built in order to read and report the fluorescence signal from the SiPM. The high gain of an SiPM is directly related to the input bias voltage or, more specifically, the overvoltage, which is the voltage in excess of the breakdown voltage. Small changes in the bias voltage can potentially create significant changes in measurements, thus in the electrical design, the bias voltage was measured by the analog to digital converter (ADC).

The original Broadcom amplifier test kit contains GHz bandwidth op-amps designed for sensing very low light levels. The pulsing and sensing requirements needed to achieve TLF sensitivity design goals were significantly relaxed, thus the SiPM detector was operated in a mode that can be modeled as a very sensitive photodiode device. A transimpedance amplifier (TIA) was used to amplify the measured current output into usable voltage.

The two high-powered UV LEDs (Seoul Viosys) with outputs centered around 275 (+/- 10) nm were set perpendicular to the SiPM. These LEDs were chosen for their high power and wavelength precision around 275 nm (Seoul Viosys, 2018). The precision around 275 nm ensures the measured fluorescence is in the TLF region.

A Particle Boron from Particle Industries, Inc. was used as the main microcontroller. Boron has an integrated LTE modem, which allows for upgrades in the field and real-time data feedback, without physical access. The Boron controlled signal measurement from a water temperature sensor (SparkFun Electronics), ambient enclosure temperature (via 1-wire sensor located on the main PCB), SiPM bias voltage, SiPM output, and DAC reference voltage. Also controlled by the Boron was pulsing the two UV LEDs, a 12V peristaltic pump to pull water through the sensor, charging of a 3.7V lithium ion battery via 10W solar panel (Voltaic Systems), and data storage to a MicroSD card for backup and buffering if the LTE connection got lost. Fig. 2.6 shows system diagram for the PCB boards, including inputs and outputs into the Particle Boron microcontroller.

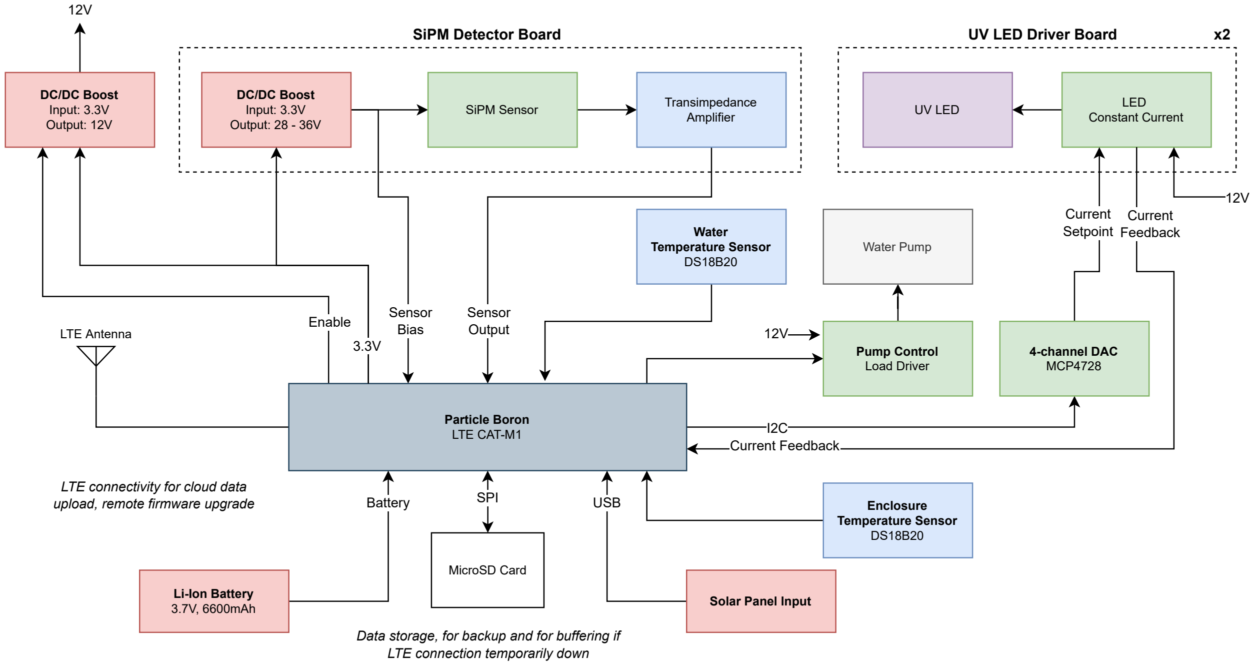


Figure 2.6: Electrical System Diagram for main PCB board on the field unit prototype. The Particle Boron microcontroller controls when and how measurements are taken and transmits the data via LTE cellular connection

The fluorescent light levels present in even low concentration tryptophan solutions were higher than what the Broadcom SiPM is intended for, thus it is functionally acting like a highly sensitive, highly efficient photodiode. A transimpedance circuit is used to convert the SiPM current to a voltage that the ADC in the Boron can sample. This circuit also has an inverting gain, set by a feedback resistor. A filtered DC bias of 3.2V (just below the ADC maximum voltage of 3.3V) is applied since a positive current causes the TIA output to go down.

A bias of approximately 32V is required for the SiPM sensor to be used. This bias voltage influences the sensor output. The actual bias voltage is sampled by the ADC during each measurement. This signal is buffered through a resistor divider and op-amp to reduce loading on the main measurement. The circuit and circuit block diagram for the SiPM sensor and amplifier are shown in Fig. 2.7.

The UV LED driver is designed to drive the LED to a constant current, pulsed in the millisecond (ms) range. It is important for the LED current to be very stable when on, since any variation in the current will couple into the output measurement. Furthermore, since different water samples may have very different levels of fluorescent response and noise, the LED current setpoint is variable instead of fixed. The setpoint voltage given to the LED driver circuit to follow is provided by a DAC. The UV LED driver circuit is composed of an op-amp constant current source using a MOSFET and a sense resistor for current feedback. Gain between the input voltage and the output current is set by this feedback resistor. The LED circuit diagram and circuit block diagram are shown in Fig. 2.8.

Firmware was developed in order to control the number of measurements taken, the length of LED pulses, the number of samples averaged for one measurement, and the current levels provided to the LEDs. Measurements from the sensor were sent through the LTE gateway to an online database. The parameters collected from the sensor are shown in Table 2.2.

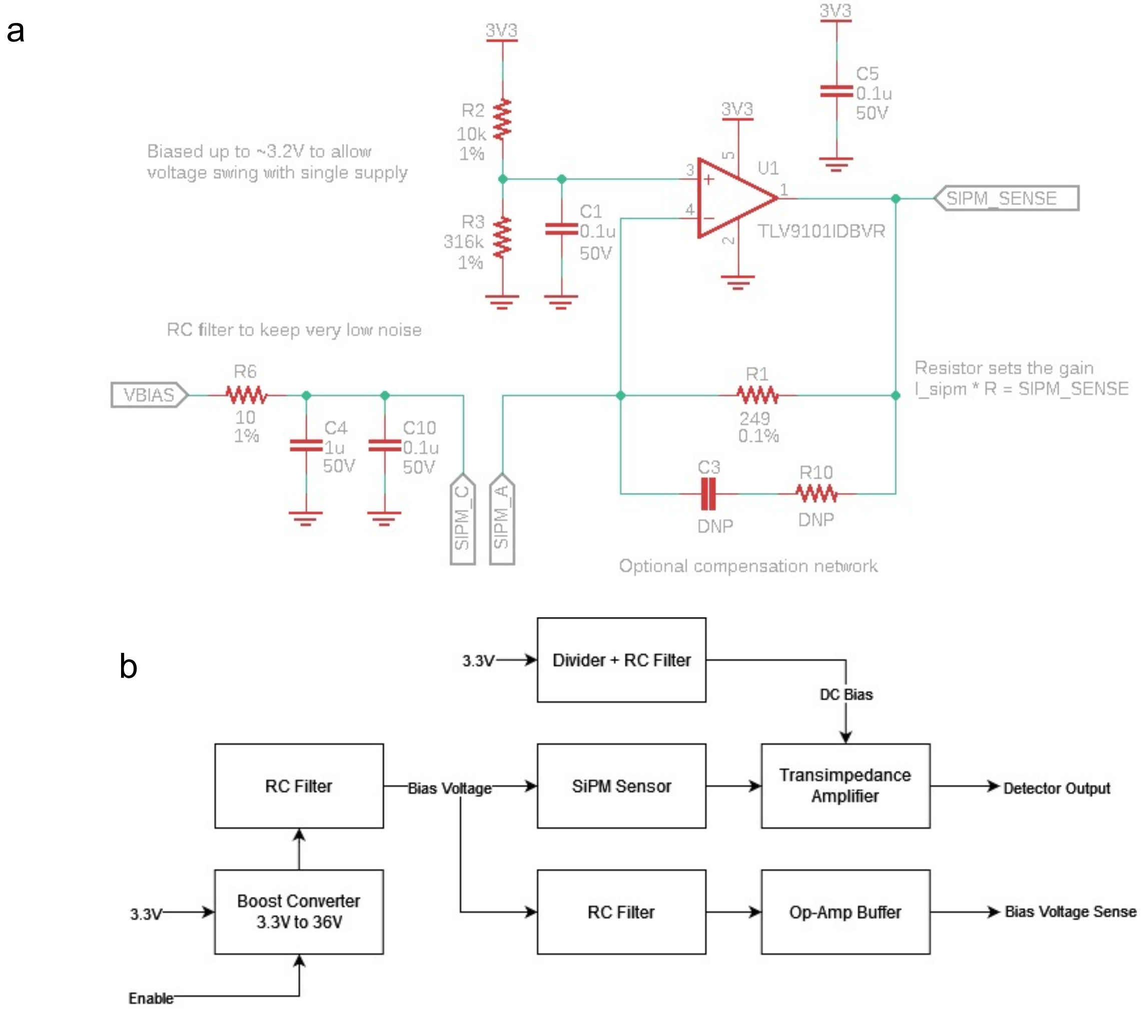


Figure 2.7: SiPM circuit diagrams; a. Circuit diagram illustrating the SiPM sensor, amplifier, resistors, and capacitors; b. Block diagram illustrating inputs and outputs to the for the SiPM sensor

2.2.1.2 Mechanical design of field unit

The sensor was enclosed in a water proof IP66 rated Polycase enclosure measuring 7.71 x 7.71 x 5.90 inches. Inputs into the enclosure were the 10W solar panel, water input, water output, and water temperature sensor. The solar panel charged a 3.7V, 6600mAh lithium ion battery. They system diagram for the mechanical design is shown in Fig. 2.9.

The sensor uses a UV quartz flow through cuvette from FireFlySci (www.fireflysci.com). The pathlength of the cuvette is 1 cm. The LEDs and SiPM PCB boards are mounted orthogonal onto

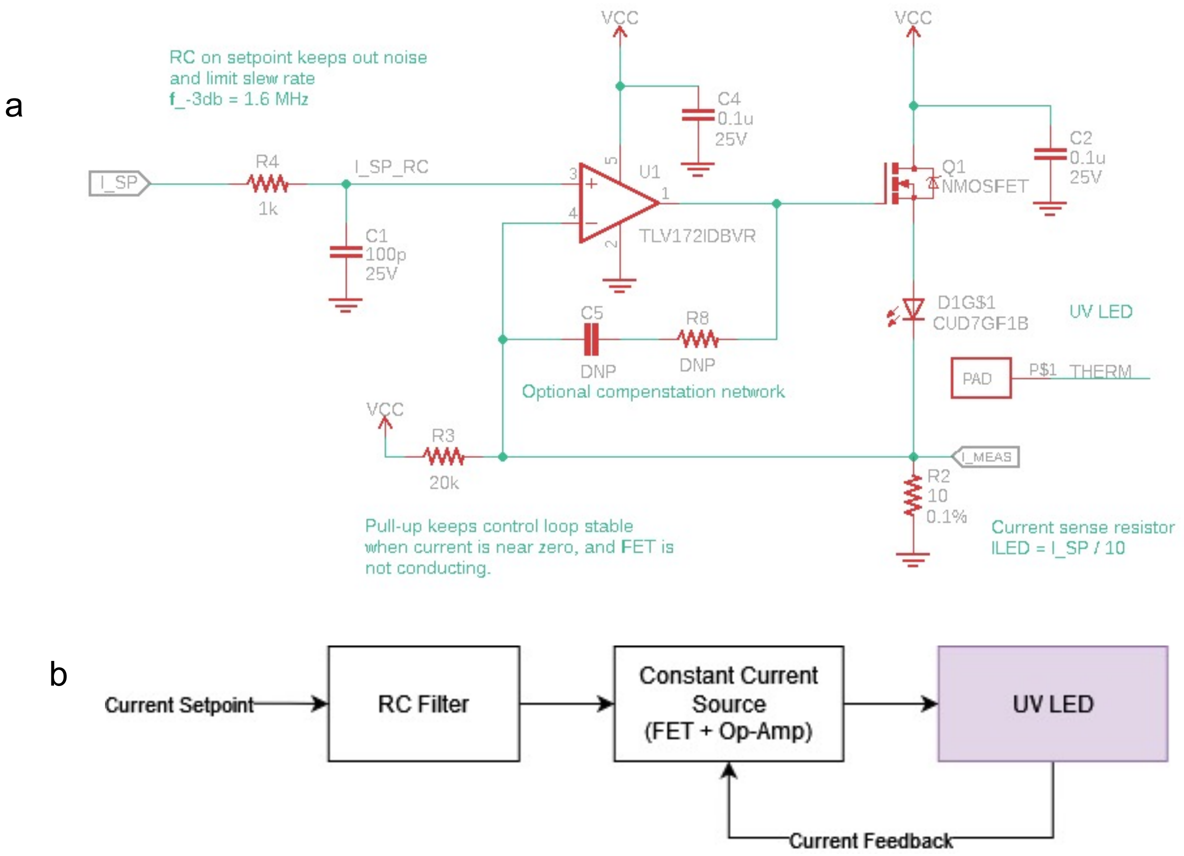


Figure 2.8: LED circuit diagrams; a. Circuit diagram illustrating input current and control; b. Block diagram illustrating current feedback for control

a high precision 3D printed sleeve that slides onto the cuvette. Mounted on the sleeve between the SiPM and the cuvette is a bandpass filter centered around 357 +/-22 nm (Edmund Optics). Abrasion-resistant rubber tubing is used to connect the cuvette to the peristaltic pump. The main PCB board, cuvette sleeve, and pump are mounted in a Polycase waterproof enclosure with a custom 3D printed mount. The enclosure is mounted on a universal camera stake with tamper resistant screws so it can be easily and securely staked into the ground. Sensor and mounting design is shown in Fig. 2.10. The bill of materials (BOM) for the sensor is shown in Table 2.3, the total material cost came to $1,150.61.

Table 2.2: Parameters measured by the sensor and sent to an online database

Parameter	Units	Description
Timestamp	Seconds since epoch	Date and Time when data was sent (seconds precision)
Group ID		Group designation by sample (all measurements for a cycle share a Group ID)
Time Start	Seconds since epoch	Time the measurement started
Pulse Time	Milliseconds	LED pulse time (ms)
LED Current Setpoint	Milliamps	LED current setpoint requested for this measurement
LED 1 & 2 Raw Current	Counts (oversampled)	Raw ADC measurement of LED currents
SiPM Raw Output	Counts (oversampled)	Raw ADC measurement of SiPM detector circuit output voltage while LEDS are on
SiPM Raw Offset	Counts (oversampled)	Raw ADC measurement of SiPM detector circuit output voltage while LEDs are off
Raw Bias Voltage	Counts (oversampled)	Raw ADC measurement of the bias voltage feedback signal
Raw Reference Voltage	Counts (oversampled)	Raw ADC measurement of DAC reference voltage
SiPM Voltage	Volts	SiPM detector signal Offset - Signal
Bias Voltage	Volts	Bias voltage input to the SiPM, corrected for resistor divider
LED Current 1 & 2	Milliamps	Currents received by the LEDs
Reference Voltage	Volts	Measured voltage at ADC of the DAC's internal reference voltage
Internal Temperature	Degrees C	Temperature measured on the PCB inside the enclosure
Water Temperature	Degrees C	Temperature of the water at time of measurement

2.2.2 Lab Validation

Lab validation was conducted to measure the signal sensitivity of the sensor to parameter changes in water. All results from lab validation experiments (Tryptophan sensitivity, *E. coli* sensitivity, wastewater sensitivity) were analyzed in R version 4.0.5. Analysis of Variance (ANOVA) as well as t-tests were conducted on the outputs to examine the statistical significance of experimental effects.

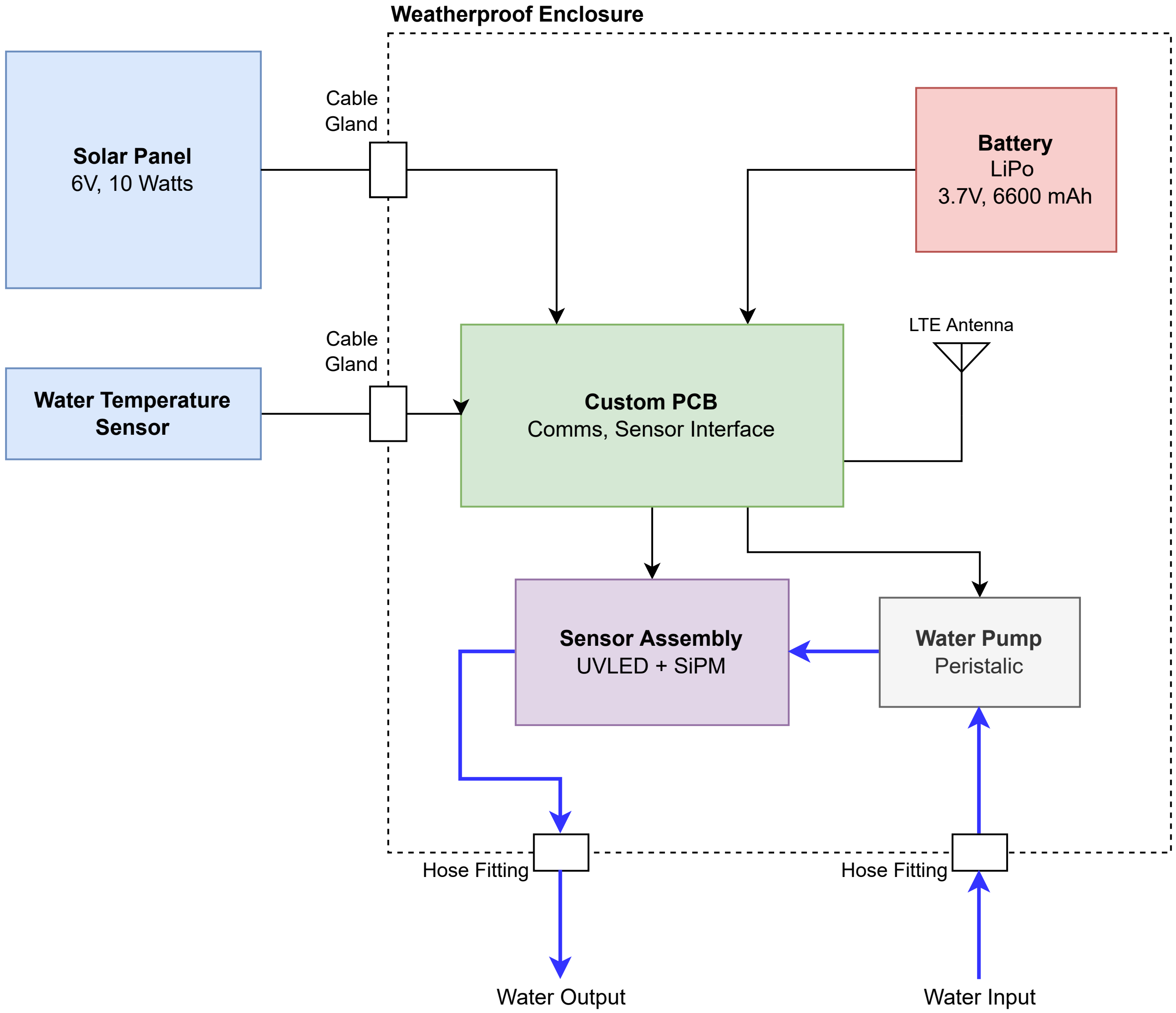


Figure 2.9: System Diagram showing inputs and outputs to the waterproof enclosure

2.2.2.1 Tryptophan Sensitivity

Standard L-tryptophan solutions (made with Sigma-Aldrich reagent grade L-tryptophan) were made by mixing 1000 mL of deionized (DI) water and 0.1 g powdered tryptophan for 30 min to create a solution of 100 ppm tryptophan. This stock solution was used to prepare serial dilutions of 0.05, 0.1, 0.5, 1, 3, 10, 30, 70 ppb L-tryptophan standards. Each solution was kept for a maximum of 72 hours. Before testing, the sensor was rinsed by pumping DI water through for 60 seconds. To collect data for each solution, starting with the lowest concentration (DI water) and ending with

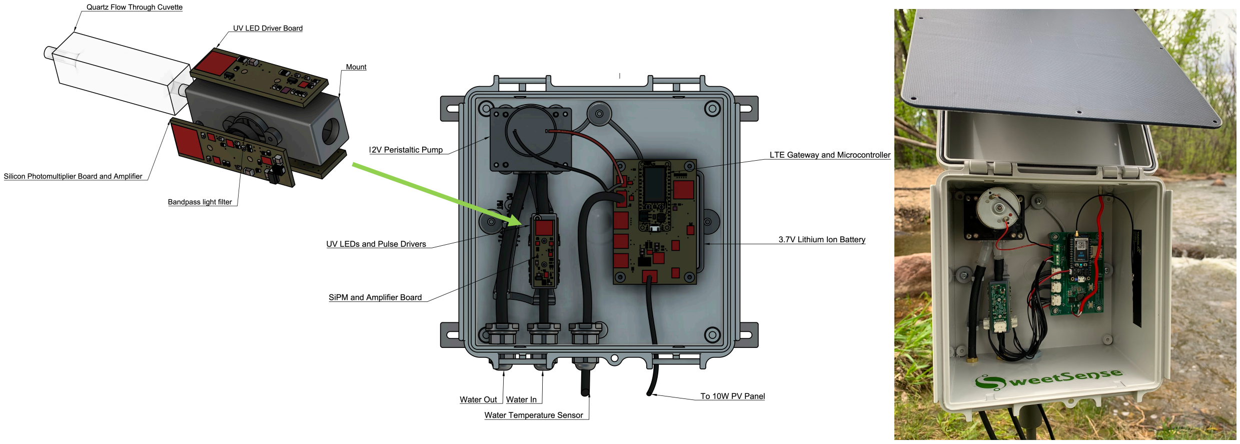


Figure 2.10: The sensor model shows the configuration of the peristaltic pump which pulls water through a flow through cuvette. UV LED and SiPM driver boards are mounted around the cuvette and connected to a microcontroller that controls measurements taken by the SiPM, water temperature sensor, and board temperature sensor. The Particle Boron board then transmits the data via LTE to an online platform. The sensor is also shown fully set-up with the solar panel mounted, next to the Boulder Creek.

the highest concentration (100 ppb), the inlet tube was placed in the solution and the outlet tube placed in a waste container. Ten TLF measurements from the sensor, with 80 samples averaged per measurement were taken for each solution. The sensor was rinsed between each solution by running DI water through for 30 seconds.

2.2.2.2 Lab Grown *E. coli* Sensitivity

Lab grown *E. coli* dilutions were prepared using *E. coli* (K-12 strain) stored in individual tubes in a freezer at -80°C. The cell culture was prepared overnight in standard nutrient broth (Difco) at 121 rpm and 37°C. The culture was then centrifuged at 3000 rpm for 10 minutes and washed in a phosphate buffered (PBS) solution to remove the growth medium and dead cell material. This step was repeated two times. The growth medium will fluoresce, thus it must be removed. PBS was tested for it's fluoresce properties and they were found to be null at TLF wavelengths. The working culture was used to make 8 dilutions, reducing each one by a power of 10. The *E. coli* concentrations present in the dilutions were verified by membrane filtration following EPA Approved Hach Co.: 10029 method. m-ColiBlue24 broth indicates *E. coli* colonies by blue coloration resulting

Table 2.3: Bill of materials for the final design of the field Sensor

Qty	**Item**	**Unit Cost**	**Subtotal**
1	Water pump	$44.79	$44.79
1	Main PCB	$15.50	$15.50
1	Main PCB components	$89.55	$89.55
1	Detector PCB	$12.07	$12.07
1	Detector PCB components	$69.35	$69.35
2	UV LED PCB	$8.55	$17.10
2	UV LED PCB components	$26.83	$53.66
1	LiPoly battery, 6600mAh	$29.50	$29.50
1	Solar panel, USB output, 10W	$65.00	$65.00
1	Solar panel cable extension with flying leads	$4.00	$4.00
1	Enclosure, IP67	$46.39	$46.39
1	Cable gland - temp sensor	$4.03	$4.03
1	Cable gland - PV panel	$11.25	$11.25
2	Water fittings	$7.37	$14.74
1	Micro SD card, 8 GB	$5.90	$5.90
1	Temperature probe	$12.00	$12.00
13	Clikmate pre-crimped wire, 150mm	$0.86	$11.18
4	Clikmate 4 pos housing	$0.25	$1.00
2	Clikmate 5 pos housing	$0.30	$0.60
1	Cuvette	$387.00	$387.00
1	Bandpass filter	$215.00	$215.00
1	Tubing	$1.00	$1.00
1	Stand	$40.00	$40.00
		Total	$1,150.61

from specific activity of β -glucuronidase and TC by red coloration resulting from specific activity of β -galactosidase (Tallon et al.). Samples were plated in triplicates and incubated at 35°C for 20-24 hours. The E. coli solutions were made right before each test and kept for a maximum of 48 hrs. Sensor data for each solution, starting with the lowest concentration (DI) and ending with the highest concentration was collected and analyzed using the same method that was used to collect the tryptophan data.

2.2.2.3 Wastewater Sensitivity

Wastewater effluent was collected from the Boulder Wastewater Treatment Facility and stored at 4°C until it was used for testing for a maximum of 5 days. Standard dilutions were made by mixing DI water and wastewater effluent to prepare 10%, 12.5%, 25%, 50%, and 100% dilutions of wastewater effluent. Each solution was made right before testing and kept for a maximum of 48 hours. Sensor data for each dilution, starting with the lowest concentration (DI) and ending with the highest concentration (100% wastewater effluent dilutions), was collected and analyzed using the same method that was used to collect the tryptophan data. Membrane filtration was used to enumerate *E. coli* and total coliforms (TC) present in each dilution by plating a filter with m-ColiBlue24 broth (EPA Approved Hach Co.: 10029 method). Samples were plated in triplicates and incubated at 35°C for 20-24 hours.

2.2.2.4 Inner Filter Effects

Samples from the wastewater sensitivity experiment were collected from the dilutions made and set aside. Each dilution was run on a UV-visible spectrophotometer (Agilent Cary 400) to collect absorbance data for 200 to 800 nm at 1 nm increments using a 1 cm path length cuvette. The spectrophotometer used had a maximum absorbance of 1 and a scan rate of 600 nm/min. The absorbance data was used to calculate the impact of possible inner filter effects (IFE) on the sensor's output (Kubista et al., 1994). IFE corrected measurements were then calculated using Eq. (2.1).

$$F_{corr} = F_{obs} * 10^{(A_{\lambda ex}+A_{\lambda em})/2} \tag{2.1}$$

Where F_{corr} is the sensor's corrected fluorescence signal, F_{obs} is the sensor's observed fluorescence signal, $A_{\lambda ex}$ is the absorbance at the excitation wavelength, and $A_{\lambda em}$ is the fluorescence observed at the emission wavelength.

2.3 Results

2.3.1 Laboratory Characterization

2.3.1.1 Varying Gain Levels

Varying the gain levels by varying the current levels to the LEDs showed that the sensor could achieve both high sensitivity and high range in a single measurement if needed. With low current to the LEDs, 5 or 10mA, the sensor could read levels up to 100 ppm tryptophan concentration. The highest current supplied to the LEDs, 200mA, showed extremely high sensitivity at low tryptophan levels, but saturated at 70 ppb tryptophan (Fig. 2.11).

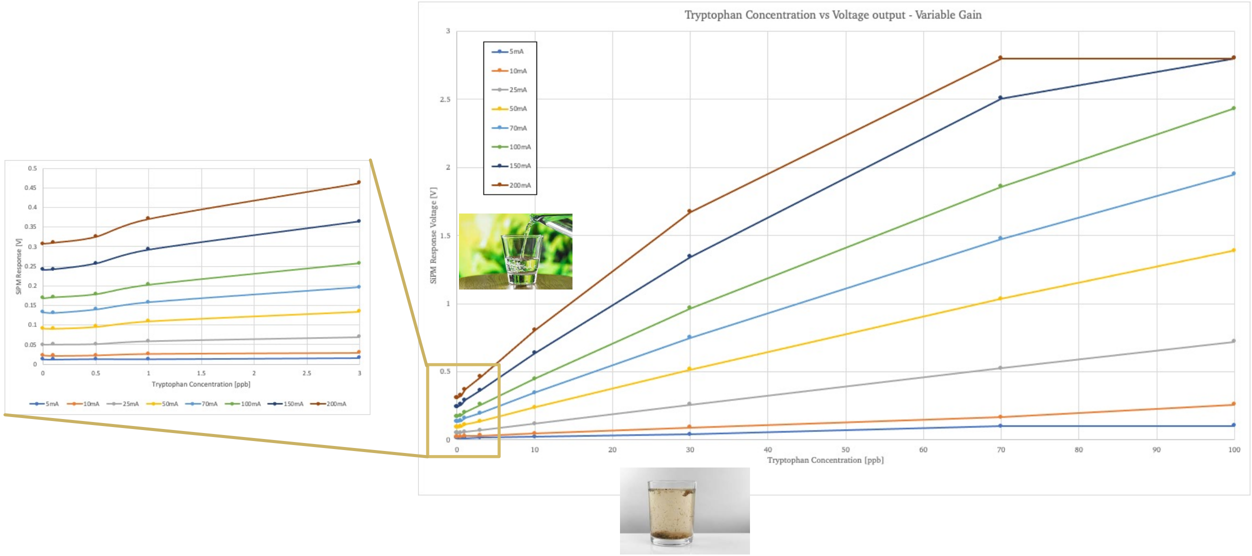


Figure 2.11: Sensor's output from varying both current levels to the LEDs and tryptophan concentrations in DI water. A zoomed in portion of the graph is shown for the lower levels of tryptophan. Water glasses are shown to signify that higher sensitivity would be used for water with less background noise, and lower sensitivity would be used for water with more background noise, where a higher range is needed

2.3.1.2 Tryptophan sensitivity

The sensor was able to significantly detect a difference between DI water and 0.05 ppb tryptophan (p <0.01 according to EPA Method Detection Limit Procedure) at a range of current inputs to the LEDs (EPA, 2016) (Fig. 2.12). Thus, the design goal of 1 ppb tryptophan (signifying high risk contamination) was met and exceeded. As the current to the LEDs increases, the sensitivity at low concentrations of tryptophan increases; as the current decreases, a higher range of tryptophan concentrations is detectable.

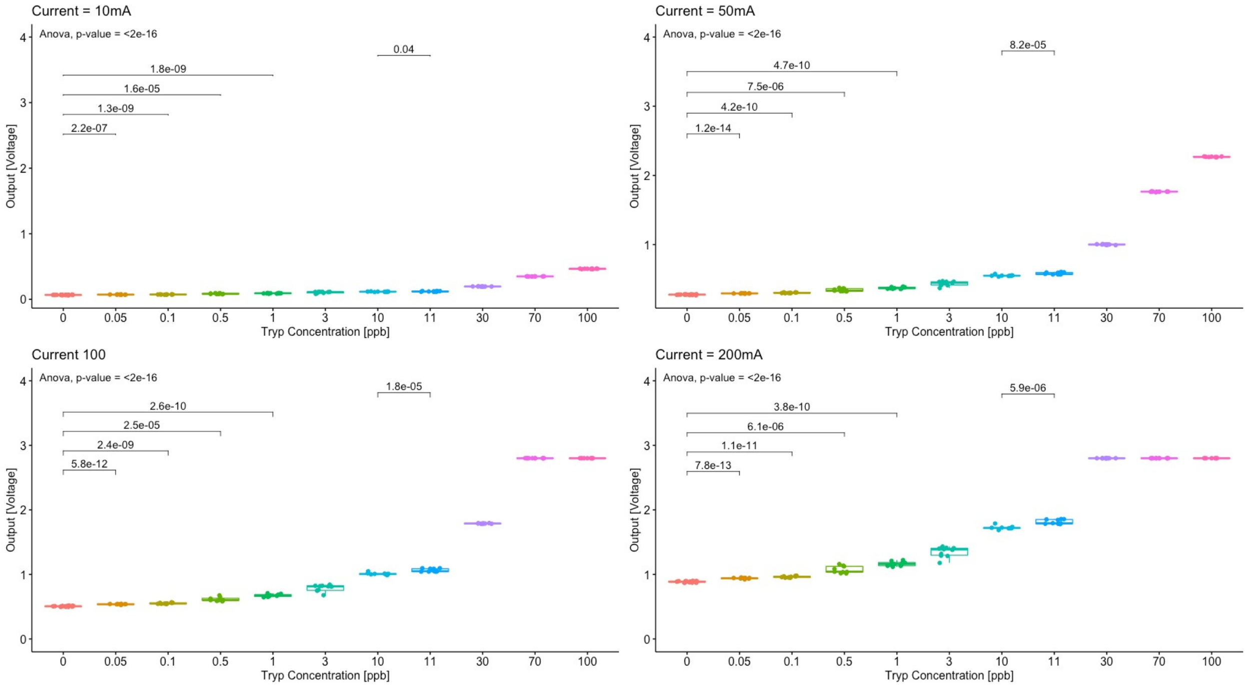


Figure 2.12: Sensor response from tryptophan dissolved in deionized water at four different current levels powering the LEDs, indicated at the top of each graph. Boxes indicate the interquartile range and median, whiskers indicate maximum and minimum values except where outliers are indicated. Bars indicate the significant differences between indicated concentrations calculated by a Students t-test. Analysis of Variance (ANOVA) p-value shows the difference across all concentrations.

2.3.1.3 Lab Grown *E. coli* Sensitivity

Lab grown *E. coli* K-12 increased the sensor response with a significant sensitivity to 33 CFU/100mL (Fig. 2.13). Above 33 CFU/100mL, the higher concentrations are also significantly different from the signal output at DI water, but the mean of the measurements fluctuates up and down until $1.5x10^3$ CFU/100 mL. This could be either due to IFE or inconsistencies in the concentration of cells in each sensor measurement. Since water is flowing through the sensor, it is possible that with each reading, the concentration present in the cuvette varies. Nonetheless, a significant change from DI water to each concentration is demonstrated, so even if the sensor output varies above DI water, the difference in contamination vs no contamination is shown by the sensor.

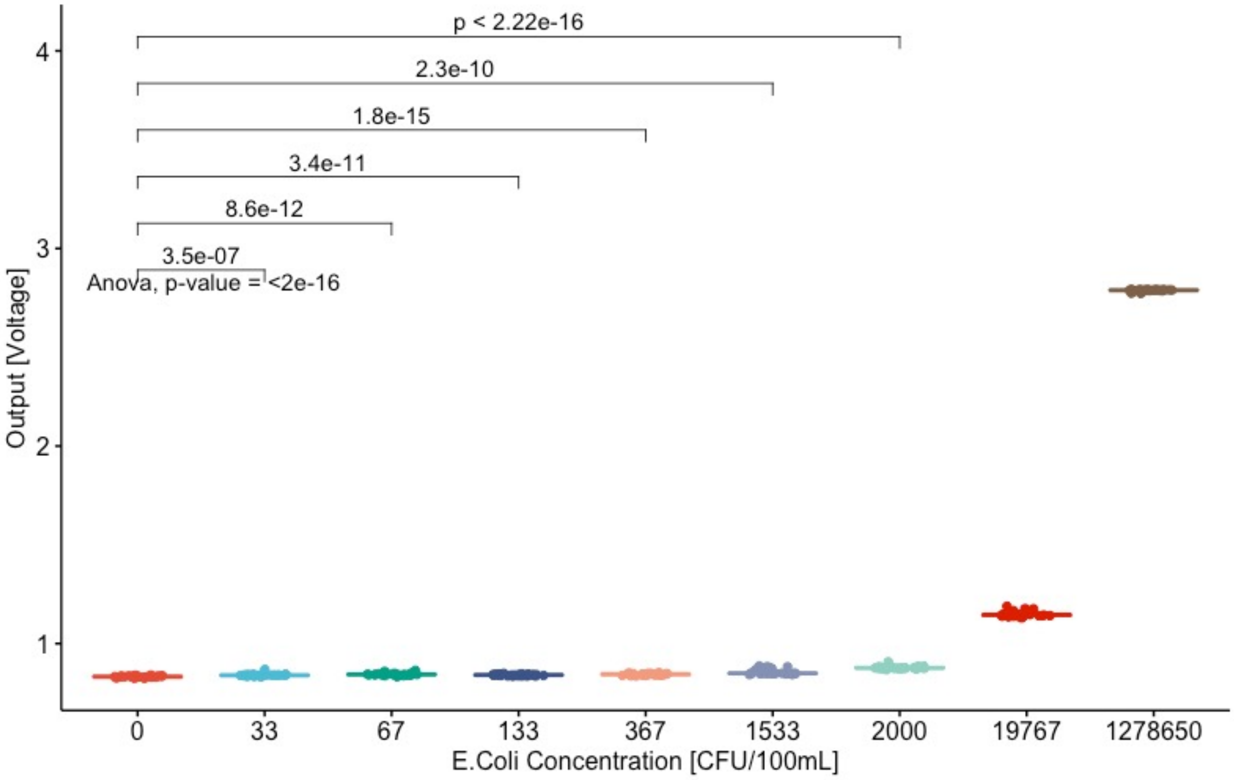


Figure 2.13: Sensor response to lab grown *E. coli.* Boxes indicate the interquartile range and median, whiskers indicate maximum and minimum values except where outliers are indicated. The bar between 0 and 10 CFU/100mL indicate the significant differences between indicated concentrations calculated by a Students t-test.

Table 2.4: Mean of sensor measurements at each lab grown *E. coli* concentration

E. coli Concentration [CFU/100mL]	**Mean Output Voltage**
0	0.833
33	0.841
67	0.845
133	0.841
367	0.845
1533	0.855
2000	0.878
19767	1.15
1278650	2.79

2.3.1.4 *E. coli* Sensitivity in Wastewater

The sensor was able to significantly detect ($p < 0.01$) *E. coli* concentrations in wastewater effluent above 10 CFU/100mL, which signifies intermediate risk contamination (Fig. 2.14). The R^2 between *E. coli* present in the wastewater and sensor output was 0.93. There was a drop in sensor output in the range of 1000 CFU/100mL. The drop in sensor output could be attributed partly to IFE, but could also be a result of light scatter from particles or inconsistancies in concentrations flowing through the sensor. The absorbance data collected on the benchtop spectrophotometer shows increasing absorbance as the concentrations increase. The calculated corrected fluorescence due to IFE increases the R^2 between *E. coli* and sensor output to 0.95 (Table 2.5).

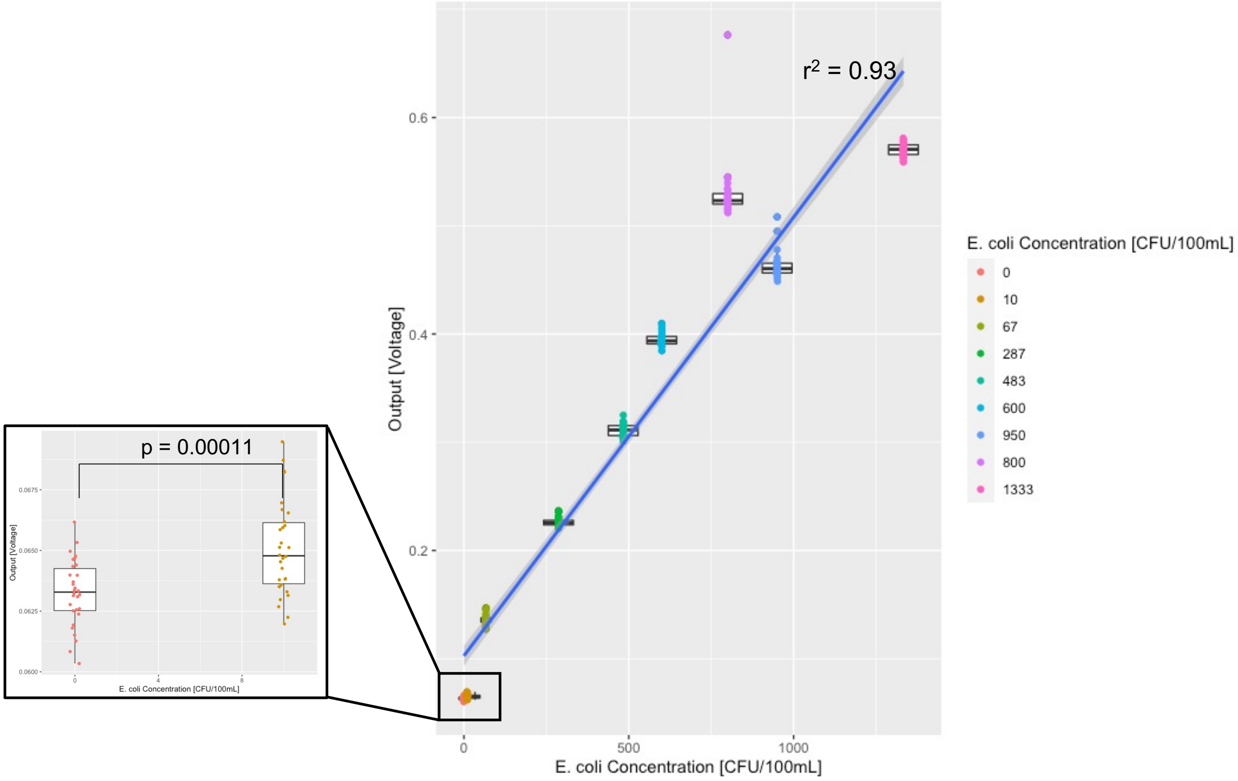


Figure 2.14: Sensor response from wastewater effluent dilutions graphed continuously. Boxes indicate the interquartile range and median, whiskers indicate maximum and minimum values except where outliers are indicated. The bar between 0 and 10 CFU/100mL indicate the significant differences between indicated concentrations calculated by a Students t-test.

Table 2.5: Corrected fluorescence based on Inner Filter Effects

E.coli **Concentration [CFU/100mL]**	**Measured Fluorescence [V]**	$A_{\lambda ex}$	$A_{\lambda em}$	**Corrected Fluorescence [V]**
0	0.06	0.00	0.00	0.06
10	0.07	0.01	0.00	0.07
67	0.14	0.01	0.00	0.14
287	0.23	0.03	0.01	0.24
483	0.31	0.05	0.02	0.34
600	0.40	0.07	0.03	0.44
800	0.53	0.10	0.04	0.62
950	0.46	0.11	0.04	0.55
1333	0.57	0.13	0.05	0.70

2.4 Conclusion

In this work, I explored the initial development of a low-cost, continuously monitoring TLF sensor to remotely report fecal contamination risk in drinking water. Through many iterations, I was able to design a highly sensitive, in-situ, autonomous, remotely reporting TLF sensor for a relatively low cost. The overall bill of materials for the sensor is approximately $1,000. This price reflects ordering supplies in small batches, if we were to produce these sensors at scale, the BOM would reduce significantly. Similar, probe style sensors designed for in-situ, but not autonomous use, can be purchased at prices ranging from $5,000 - $6,000 (Sorensen et al., 2018b).

Using high powered LEDs and a SiPM showed a significant sensitivity to tryptophan of 0.05 ppb tryptophan. This sensitivity is currently better than high sensitivity sensors on the market, which range from 0.1 ppb to 0.17 ppb (Simões and Dong, 2018; Khamis et al., 2015).

These design approaches also successfully demonstrated a correlation between TLF and *E. coli*. The sensor showed a sensitivity to lab grown *E. coli* of 33 CFU/100mL and to *E. coli* in WWE of 10 CFU/100mL. A lower sensitivity is possible in WWE compared to lab grown *E. coli* because there is more extracellular material that contains fluorophores in WWE, where most of the TLF signal is coming from. This presents evidence of the current potential to quantify TLF instantaneously at a sensitivity that is meaningful for monitoring drinking water quality. This sensitivity limit will not allow for presence/absence detection of contamination or detection of low or intermediate risk, but can provide information and data on sources containing high risk contamination. Since *E. coli* is an indicator of fecal contamination, this prototype design proved the potential for TLF to detect potentially harmful pathogens in drinking water, even with the cheapest available technology. The flow through design proves that a TLF product capable of continuous detection is a viable option for real or near-time detection of fecal contaminated in drinking water, which is currently being consumed by approximately 2 billion people globally.

The variable gain option integrated into the firmware of the sensor allows for the user to decide between higher sensitivity or higher range of output. These options allow the sensor to be useful in a variety of different environments from surface water to groundwater to tap water.

Laboratory experiments showed that IFE and possible light scattering impact the sensor's signal significantly. The sensor's signal may be attenuated from absorption of either the excitation or emission wavelengths and/or light scattering from particles. Since IFE becomes noticeable at high concentrations of contamination, it's impacts on the signal may not present a problem for differentiating between high risk and not high risk contamination. Light scattering occurring in the sensor's cuvette may increase it's signal along side TLF.

The design presented in this paper has the potential to respond the majority of UNICEF's needs for a novel, real-time, in-situ *E. coli* detection device. The device is battery based, has a processing time requirement of milliseconds, eliminates the need for a reagent or incubation, and presents qualitative output based on fecal contamination ranges greater than 10 CFUs/100 mL. Further testing will need to be conducted to determine the sensor's false positive and false negative error rates.

The novelty of this work lies in the sensor cost and its ability to monitor water contamination risk continuously. The prototype sensor is envisioned to be an order of magnitude more affordable than currently available on the market by employing newly available semiconductor technologies. Current sensors available for TLF measurements are portable and handheld, the sensor presented in this paper will have the ability to be installed in-line in drinking water distribution systems and report remotely in near-time. This has the potential to provide water service providers with actionable fecal contamination risk data that will positively impact important decision making, leading to improved health and livelihood of consumers.

Moving forward, interference in the current design's signal from physicochemical parameters will need to be determined. Temperature, pH, and turbidity, though typically outside of typical natural ranges, can all impact TLF signal. The current configuration has a water temperature sensor integrated, but corrections to the signal for water temperature need to be determined and limits may need to be set on specific use cases of the sensor. Some of these challenges can be addressed by advanced signal processing utilizing data from large sensor networks deployed in one geographical area.

Further, I will display an attempt to address signal drift through machine-learning based synthetic calibration to enable alarm-threshold detection. Through a combination of low-cost, robust hardware design and machine learning, I attempt to address the signal drift limitations through long-term characterization of in-situ water quality, and identify potential microbial contamination through alarm-based event detection. This will allow for site-sensor specific baseline calibration and the detection of sudden and large changes in TLF rather than relying on a universal threshold for all sensors.

Chapter 3

Evaluating signal interference on an in-situ TLF sensor including variation in water quality, temperature, biofilm and scaling formation

Part of this chapter appears in a paper submitted for publication in *Water Research*, with the following authorship attribution: Emily Bedell, Olivia Harmon, Katie Fankhauser, Zachery Shivers, and Evan Thomas

3.1 Introduction

The need for a rapid and reliable detection method of microbial contamination in water treatment and monitoring is well known. In order to achieve real time detection, online monitoring is required. Sensors designed for online monitoring with autonomous operation must take into account multiple factors that will influence their signal and output. Most probe type fluorescence sensors that have been proven to be able to predict microbial water quality health risk have not taken varying water quality into account. Since these sensors do not perform online monitoring, they also don't have to consider biofouling or mineral scaling impacts.

The main inhibitor of accuracy and longevity of online fluorescence measurements has long been considered to be biofouling. Biofouling occurs when biofilms form on a sensor's lenses. Biofilm formation happens in four main stages:

(1) Bacterial attachment to a surface

(2) Microcolony formation

(3) Biofilm maturation

(4) Detachment and dispersal of bacteria that may colonize new areas

All aquatic sensors will biofoul if given enough time deployed in natural water settings. The extent of biofouling varies as a function of the environment and contents of the water (Coble et al., 2014). Biofouling can impact fluorescence data by either decreasing or increasing the signal. The fouling material can physically block both excitation or emission light to decrease the signal. If the fouling material fluoresces at the same wavelength measured by the sensor it can increase the signal (Delauney et al., 2010). In order to utilize a fluorescence sensor that will be impacted by biofouling, some form of mitigation needs to take place. Current anti-biofouling techniques include shutters, plates, and whipers for open faced sensors. Copper tubing and tubing covered with foil or black tape to block light has been used for flow through instruments (Manov et al.). Other techniques include pressured air to clean surfaces, nanocoating technologies, or nano-treated plastics to prevent biofilm adhesion. All of these techniques contain limitations. Mechanical systems can easily brake or cause obstruction in light paths and coatings can interfere with excitation or emission intensity.

Hard water with excess calcium and magnesium may cause mineral scaling on a sensor's optical windows. The taste threshold set by WHO for the calcium ion is in the range of 100-300 mg/l and lower for magnesium. Hardness of water is the combined mineral concentratioin. Hardness above 200 mg/l may cause scale deposition over time on any hardware which interacts with the water. When hardness falls below 100mg/l it tends to have a low buffering capacity and can be more corrosive to pipes (WHO, 2017). Calcium carbonate ($CaCO^3$) has a high refractive index. Transmitted light decreases as a function of increasing refractive index, leading sensor signals to decrease as minerals begin to scale. Minerals also have high reflectivity which can cause scattering or higher signal output, but a lower overall sensitivity of a sensor (Okazaki et al., 2017).

Other varying water quality parameters may impact the sensor's signal. Water temperature, pH, and turbidity have all been shown to impact TLF signals. Rising water temperature can cause TLF to quench up to 35%. Rising temperature increases the likelihood that electrons fall back to their ground state without emitting a photon. The impact of thermal quenching is related to the amount of exposure the fluorophore gets from the energy source. This exposure can vary between free amino acids, tryptophan within a protein, and tryptophan or other tryptophan-like fluorophores present in molecules (Baker, 2005).

pH is known to interfere with fluorescence measurements outside of natural ranges. This interference is due to influence of either deprotonation or protonation of acidic or basic functional groups bound directly to fluorophores. Depending on contents in water, pH outside of neutral ranges can either increase or decrease fluorescence (Coble et al., 2014). Spencer et al. 2007 studied the impact of pH on fluorescence over a range of 2-10 and found that within natural pH levels typically observed in freshwaters, the response of fluorescence signals was limited, but outside of natural pH levels fluorescence properties were quite sensitive to pH (Spencer et al., 2007). Literature shows varying impacts on DOM fluorescence with variation of pH, this can be attributed to the unknown structures associated with DOM and humic substances. There is general agreement throughout the literature that for pH between 6 and 8 there is no adjustment needed for flourescence measurements.

Turbidity is the cloudiness of water caused by a large number of individual particles. Inorganic or organic particles in natural waters can be problematic for any in-situ optical measurement. Particles can increase light scattering within a sample volume and depending on their make up, either increase or decrease the signal output. Measuring a sensor's response to turbidity is complex in that particles may absorb excitation or emission light or scatter light directly back into the sensor's detector. Particles in natural systems will not be homogeneous and will not interact with the sensor in a predictable manner (Coble et al., 2014).

It is unclear whether TLF sensors are detecting live, inactivated, or dead cell material that fluoresces in the region of interest. The main interest in microbial contamination is living mircroorganisms because of the risk of disease. Ideally a sensor can distinguish between live, dead, and inactivated cells. Varying treatment processes impact cells differently, so it is important to establish the impact of each treatment process on a sensor's response. Live, viable cells have metabolic activity and are capable of multiplying. Inactivated cells are still alive and have some metabolic activity, but are not capable of multiplying. Cells are typically inactivated by UV treatment. Dead cells are incapable of multiplying and they eventually degrade as their cell membranes lyse open. Disinfection by chlorination will lyse open and kill cells. Sorensen et al. 2020 showed TLF is measuring predominately extracellular material that fluoresces in groundwater, not bacterial cells. This study monitored TLF before and after filtration (Sorensen et al., 2020). Related to the extracellular manner of TLF is the impact chlorine has on TLF signals. Multiple studies have evaluated the effect of chlorine addition in wastewater, all showing reduction of fluorescence intensities after disinfection (Hambly et al., 2010; Murphy et al., 2011). Hambly et al. 2010 shows fluorescence change throughout different steps of a wastewater treatment plant. Chlorination was shown to reduce fluorescence intensities the most out of the different treatment methods. Li et al. (2019) measured fluorescence intensity before and after UV disinfection and found that for the first minute of UV dose, the intensity increased, but for any dose longer, the intensity decreased. This can be attributed to initial protein unfolding when the cells are first exposed to UV radiation, exposing more amino acids to direct UV. Tryptophan and its derivatives are then denatured by continuous UV radiation, and the fluorescence intensity decreases (Li et al., 2019). From this literature, it seems that TLF is not sensing live cells, but microbial activity present in water.

In this chapter, I will characterize the sensor's response to interference from these various environmental parameters. It is key to understand how the TLF sensor designed in chapter 1 will be impacted by different contents and makeup of treated and untreated waters. In order to develop a sensor system that can operate autonomously, it is important to characterize how different factors will impact the signal in order to optimize response time and decision making around a possible contamination event.

3.2 Methodology

3.2.1 pH Variation

A single tryptophan solution of 50ppb was mixed following the methods described above. A 0.1 M solution of HNO3 and a 0.1 M solution of NaOH was created to vary the pH of DI, 50 ppb tryptophan with DI water, tap water, and 50 ppb tryptophan with tap water from pH 3 to 11 in 0.5 increments. Increments of 1 mL of HNO3 or NaOH were added to each dilution and mixed on a stir plate for 4 min until a steady state of the pH-adjusted solution was reached. After 4 min, a pH meter (Vernier Software Technology) was used to record the pH before water was pumped through the sensor for 10 seconds then a measurement was taken.

3.2.2 Turbidity Variation

In order to characterize the sensor's response to one type of particle, different amounts of sediment were added to DI water and their turbidity monitored along with the fluoresce measurements. The sediment chosen for this experiment was a clay called Fuller's Earth (D50 = 11.9 µm). All sediment used in this experiment was treated with hydrogen peroxide to remove organic matter then rinsed with DI water and dried in an oven at 65 °C for 24 hrs. Turbidity was evaluated for DI and 50 ppb tryptophan. The treated sediment was weighted and added incrementally to each standard to collect data for 0, 20, 50, 100, 150, 200 NTUs. Before testing each dilution, the inlet tube was placed in a separate DI bottle used for flushing out the tubes and cuvette and the outlet tube in a separate waste bottle then water was pumped through for 60 seconds. Measured treated sediment was added to the dilution and mixed for 5 min. Water with varying turbidity levels was pumped through the sensor for 10s and a measurement was taken. 50 mL of the dilution was collected to measure turbidity on the turbidimeter. The sensor data collected was evaluated in RStudio to create a line graph that was used to show the difference between each tryptophan concentration as turbidity varied.

3.2.3 Water Temperature Variation and Correction

Tryptophan solutions were made following the methods described above. Each solution (DI water, 1, 3, 10, 30, 70, 100, 200 ppb) was refrigerated overnight at 4°C. Solutions were removed from the refrigerator one at a time and tested in ascending concentration. Solutions were placed on a hotplate stirrer and stirred during the entirety of the test, heat was applied when the solution began to reach room temperature. The temperature of the water was monitored using a thermocouple from 7°C to 35°C. Measurements were taken by the sensor at least three times per degree Celsius.

3.2.4 Chlorination Impacts

To determine the impact of chlorine on the sensor's output and whether the sensor is detecting live, inactive, or dead cells, 0.24 mL of bleach was added to 1000 mL wastewater effluent, a 50 ppb tryptophan solution, and DI water and mixed for 30 minutes. Total chlorine present was measured at minute 1 and 30. Free chlorine was measured at minute 30. 10 measurements with the sensor were taken for each solution (wastewater effluent, wastewater effluent + bleach, 50 ppb tryptophan, 50 ppb tryptophan + bleach, DI water, DI + bleach). The concentration of *E. coli* and TC was enumerated by membrane filtration using the method described above.

3.2.5 Biofouling Estimation

To establish biofilm growth and estimate its impact on sensor signal, two sensors sampled tap water that was occasionally spiked with wastewater effluent continuously for four weeks. Grab samples were taken three times a day and enumerated for E. coli and total coliforms by membrane filtration using the methods described above. Biofilm growth was inferred through proxy markers that estimated their existence or absence. The general assumption was that the proxy substance being measured was directly related to biofilm existence on the cuvette walls (Azeredo et al., 2017). In each sensor, the cuvette was cleaned periodically. Before cleaning, the biofilm growth was qualified with spectral analysis to create excitation-emission matrices (EEMs) and quantified with membrane filtration. Fluorescence of the biofilm growth was collected using a fluorescence spectrofluorimeter

(Fluoromax-4). Emission wavelengths were measured from 300 to 400 nm in 2 nm increments, excitation wavelengths from 300 to 240 nm in 10 nm increments, and the excitation and emission bandpass was set to 5 nm and a 0.25 sec integration time.

Absorbance data were collected using an ultraviolet-visible spectrophotometer (Agilent Cary 4000) which has a maximum absorbance of 1 and a scan rate of 600 nm/min. Absorbance spectra were measured from a wavelength of 200 to 800 nm at 1 nm intervals.

After fluorescence and absorbance data was collected, the EEMs were corrected using staRdom (R package version 1.1.14) (Pucher et al., 2019). The R package staRdom corrects inner filtering, Rayleigh and Raman scattering, and then subtracts background and baseline signal using data from a cuvette containing only DI water.

The biofilm that formed in the cuvette was enumerated through membrane filtration using the methods described above. The cuvette was filled with a solution of PBS and water, shaken for 30 secs to detach biofilm. This was performed 15 times. 10 mL of this solution was passed through a membrane filter.

3.2.6 Mineral Scaling

Mineral scaling experiments were conducted using a scaling model solution mimicking the composition of the Colorado River consisting of calcium chloride dihydrate (0.0167 M), magnesium sulfate (0.0105 M), and sodium sulfate (0.0145 M) (Rahardianto et al., 2006).

Before each experiment, DI water was cycled through the cuvette in the sensor for 2 hours. Then the scaling model solution was added and cycled through the cuvette for 24 hours to form scaling on the cuvette walls.

To analyze the impact of mineral scaling on the sensor's signal, the mean sensitivity of the sensor's signal to 50 ppb tryptophan was compared before and after the scaling experiment.

3.3 Results

3.3.1 pH Impact

The effects of pH on the sensor's signal in varying tryptophan solutions were greater at high and low pH (Fig. 3.1). For low concentrations of tryptophan, the sensor response increased at both low and high pH, causing lower sensitivity at low pH. For 10 and 30 ppb tryptophan, the sensor response increased significantly at pH greater than 9. For 100 ppb tryptophan, the sensor's signal increased slightly below a pH of 6 and greatly above a pH of 9, causing the sensor to saturate at a pH of 9.5. This experiment showed that for neutral pH levels, between 6 and 8, the sensor's response is not significantly impacted at varying levels of tryptophan concentrations.

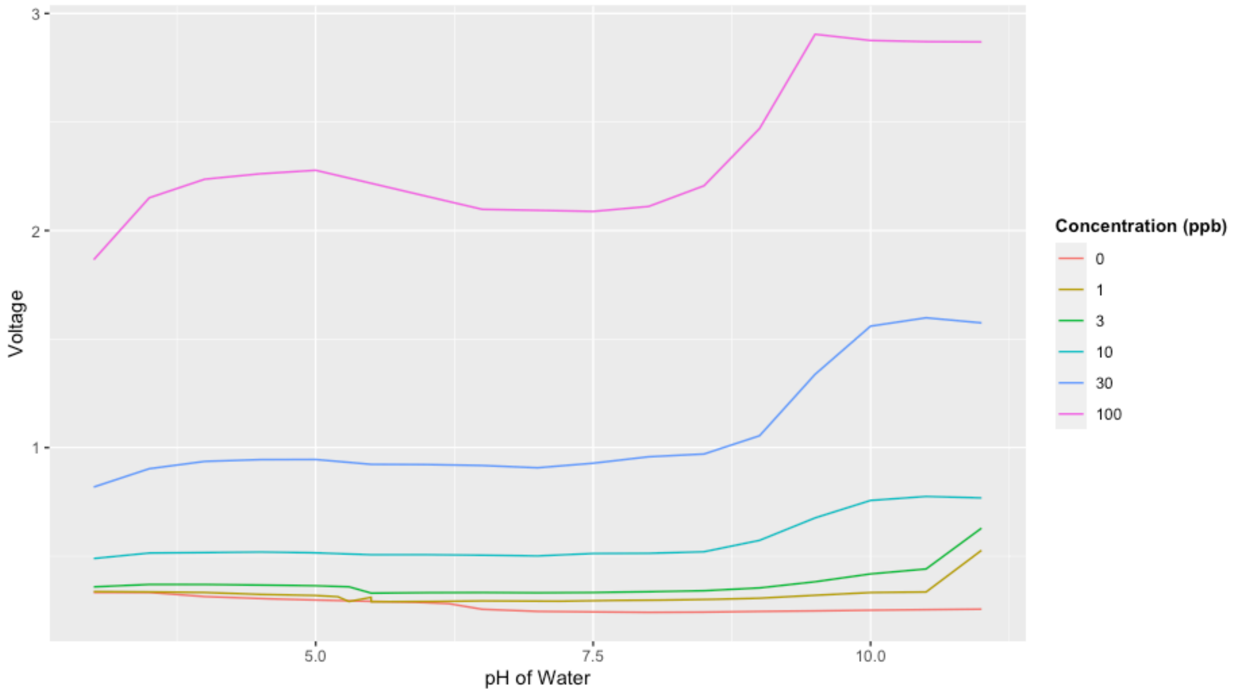


Figure 3.1: Sensor response to varying pH levels in varying tryptophan solutions

3.3.2 Turbidity Impact

The sensor's signal increased with increasing Fuller's Earth turbidity (Fig. 3.2). Multiple tests were run and each time, even with strict disinfection procedures of the clay, turbidity increased the sensor's response.

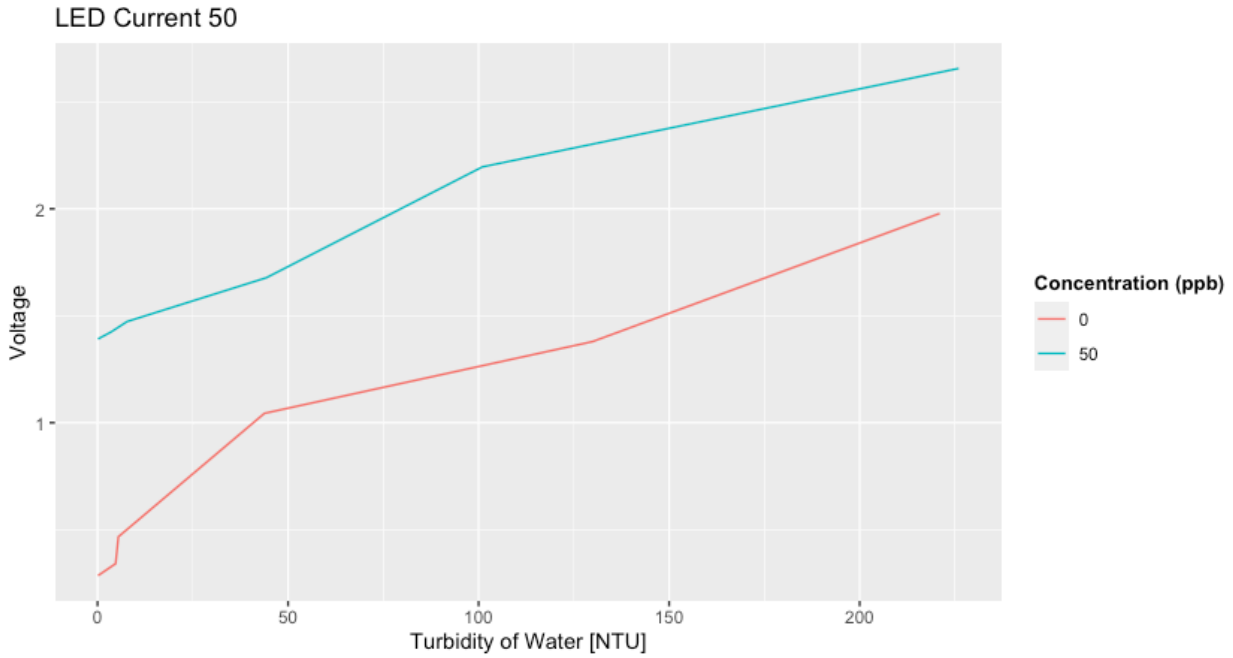


Figure 3.2: Sensor response to varying turbidity levels in DI and 50 ppb tryptophan

3.3.3 Chlorination Impact

Adding bleach to solutions of DI water, 50 ppb of tryptophan and to wastewater effluent significantly lowered the signal from the sensor (Fig. 3.3). The free chlorine present after 30 minutes in each solution was 5.9, 5.5, and 5.6 mg/L for the DI, 50 ppb tryptophan, and wastewater effluent, respectively.

3.3.4 Temperature Sensitivity

The sensor output was negatively correlated to water temperature (Fig. 3.4). The higher the concentration of tryptophan in the water, the more significant impact water temperature has on the signal. Similarly to Watras et al. 2011, fluorescence declined exponentially with water temperature at all concentrations tested, thus the methods for temperature correction of a fluorescence sensor described in that study were followed (Watras et al., 2011). A linear fit was first attempted with

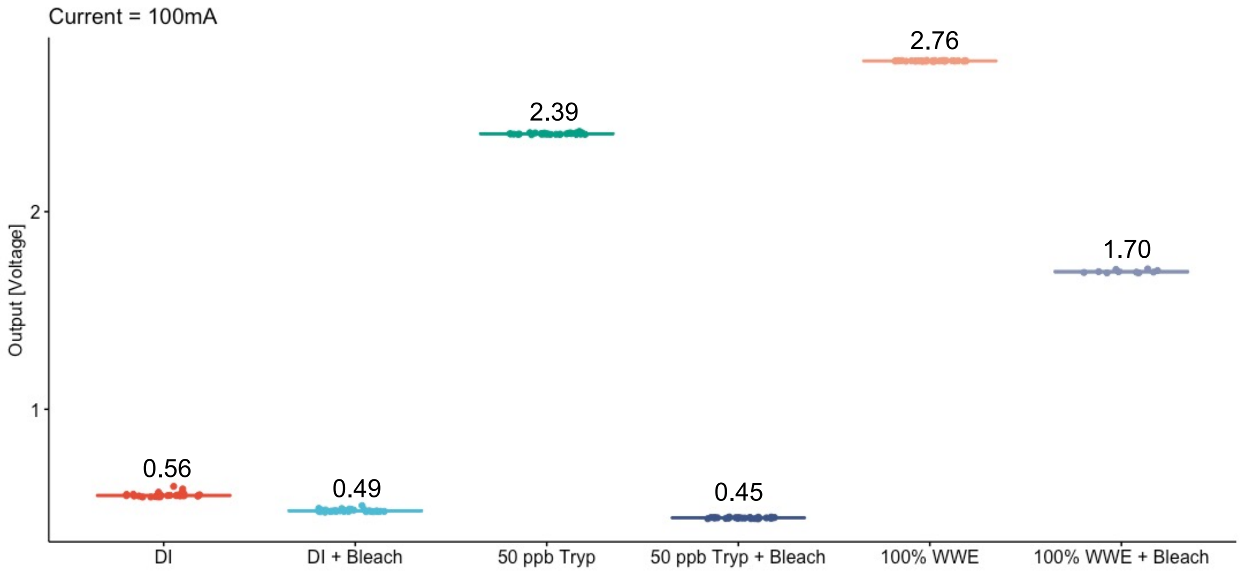


Figure 3.3: Sensor response adding bleach to a DI, a tryptophan solution, and wastewater effluent. Boxes indicate the interquartile range and median, whiskers indicate maximum and minimum values except where outliers are indicated. Means are displayed above each box for comparison.

the data, but less error was found in a negative exponential relationship. The data was fitted to the functional relationship:

$$TLF_m = TLF_r e^{\rho(T_m - T_r)} \tag{3.1}$$

Where T is temperature (°C), ρ is the temperature coefficient ($°C^{-1}$), the subscripts r and m stand for the reference and measured values. Eq. 3.1 was fit to each concentration for each LED current level. The values for ρ are shown in Table 3.1.

Table 3.1: ρ values calculated for each current level to correct measured fluorescence values for temperature of the water

ρ	Current [mA]
-0.03	10
-0.025	50
-0.02	100
-0.015	200

Using this value for ρ in Eq. 3.1, the effect of temperature can mostly be removed from the raw data (Fig. 3.4b) For the lower concentrations, correcting the data causes a small increase of the sensor output with temperature.

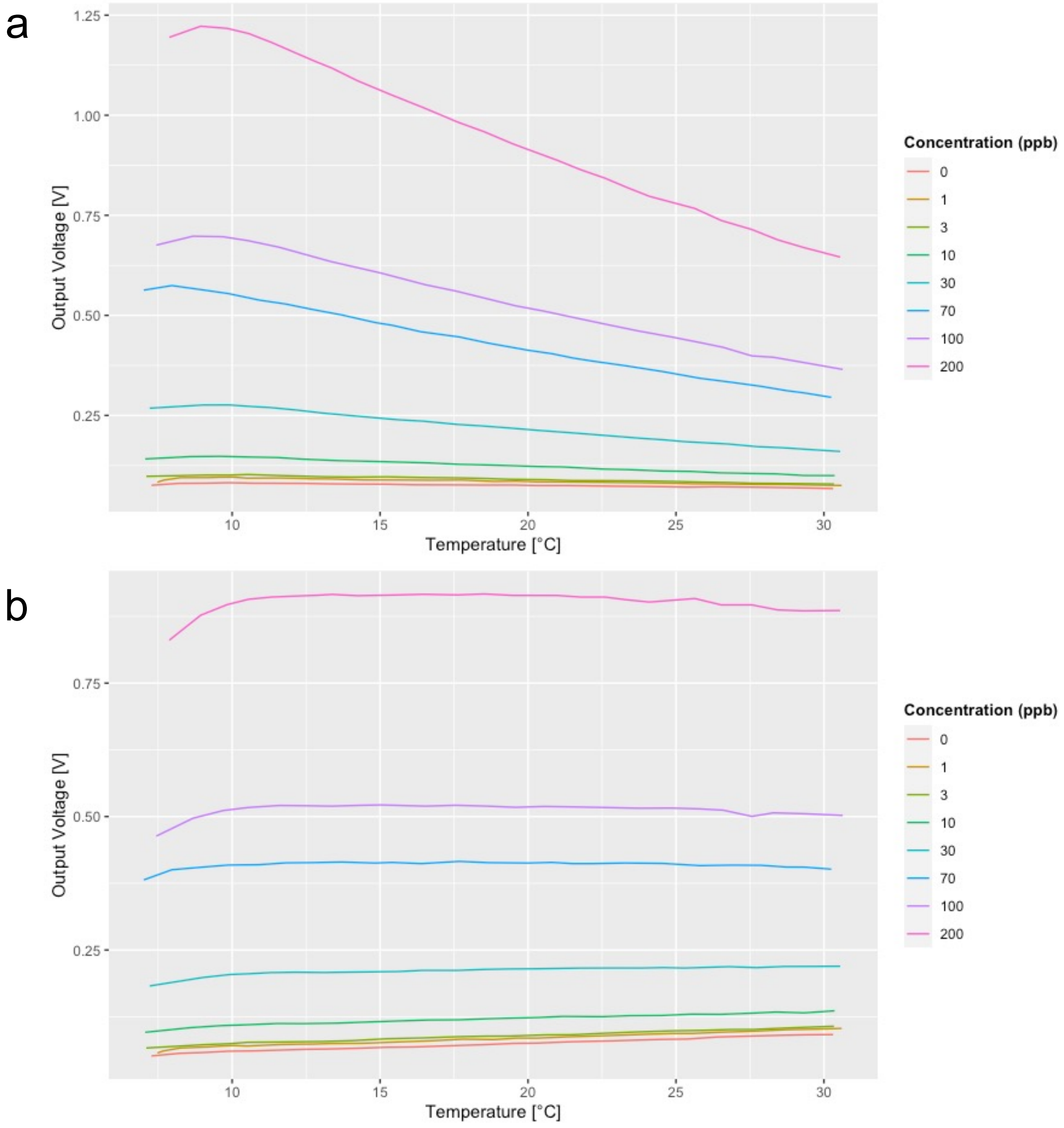


Figure 3.4: Temperature impact and correction at 10mA supplied to the LEDs a. Sensor response to increasing temperature at increasing tryptophan concentrations b. Data corrected to temperature at 20°C

3.3.5 Biofouling Sensitivity

As wastewater effluent was introduced to the sensor for extended periods of time, an increase in the sensors signal was observed even when only sampling DI water (Fig. 3.5a). This signal increase is assumed to be attributable to biofilm growth on the inside walls of the cuvette. Spectral EEMs for a biofouled cuvette using the clean cuvette as a blank show an increase in signal in the TLF range (Fig. 3.5b).

Monitoring absorbance in a cuvette with biofilms present will impact the outcome of the Beer-Lambert law, as it relies on path length and the concentration of a solution inside the cuvette. Varying the absorbance path length from 0.01 to 5 cm, impacted scale of the results, but since only the presence/absence of increased fluorescence was of concern, these impacts were disregarded. Since the cuvettes were filled only with DI water, the increase in fluorescence in Fig. 3.5b is inferred to be a proxy for presence of biofilms. *E. coli* enumeration showed 10 CFU/100mL were present inside the cuvette after this EEM was taken. The average amount of signal growth between wastewater spikes was calculated to be 82%.

Spikes in the sensor signal from contamination can be observed through biofouling induced signal increase. As little as 17 CFU/100mL show a significant increase in signal, even with a heavily fouled lens. There is an instance of sensor signal spike with no *E. coli* present in the wastewater effluent. An instance like this may be characterized as a false positive, but could also indicate that there was a contamination event at a previous time, but all live FIB had died off.

3.3.6 Mineral Scaling Sensitivity

After exposing the sensor to mineral scaling solution for 24 hours, there was an increase in the sensor's signal, but a decrease in the sensor's sensitivity (Fig. 3.6). Mineral scaling showed a 5% reduction in sensitivity of the sensor's ability to measure 50 ppb tryptophan.

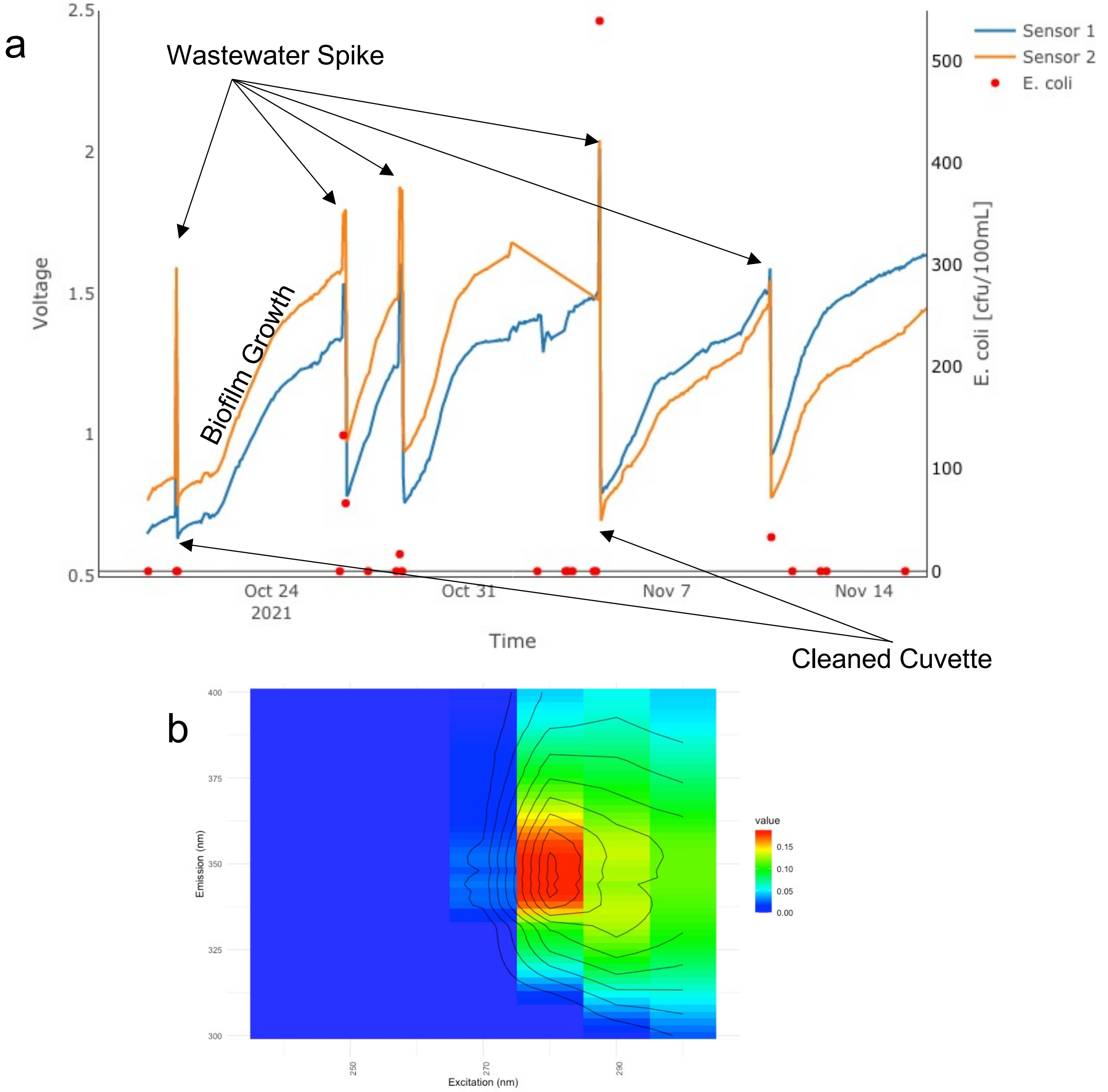


Figure 3.5: Biofilm test results: a. Data from two sensors plotted over one month of sampling tap water spiked with wastewater effluent five times throughout, combined with *E. coli* data from grab samples taken before and after wastewater effluent spikes. b. EEM of a cuvette that had been in a sensor after two weeks of sampling with wastewater effluent spikes in tap water.

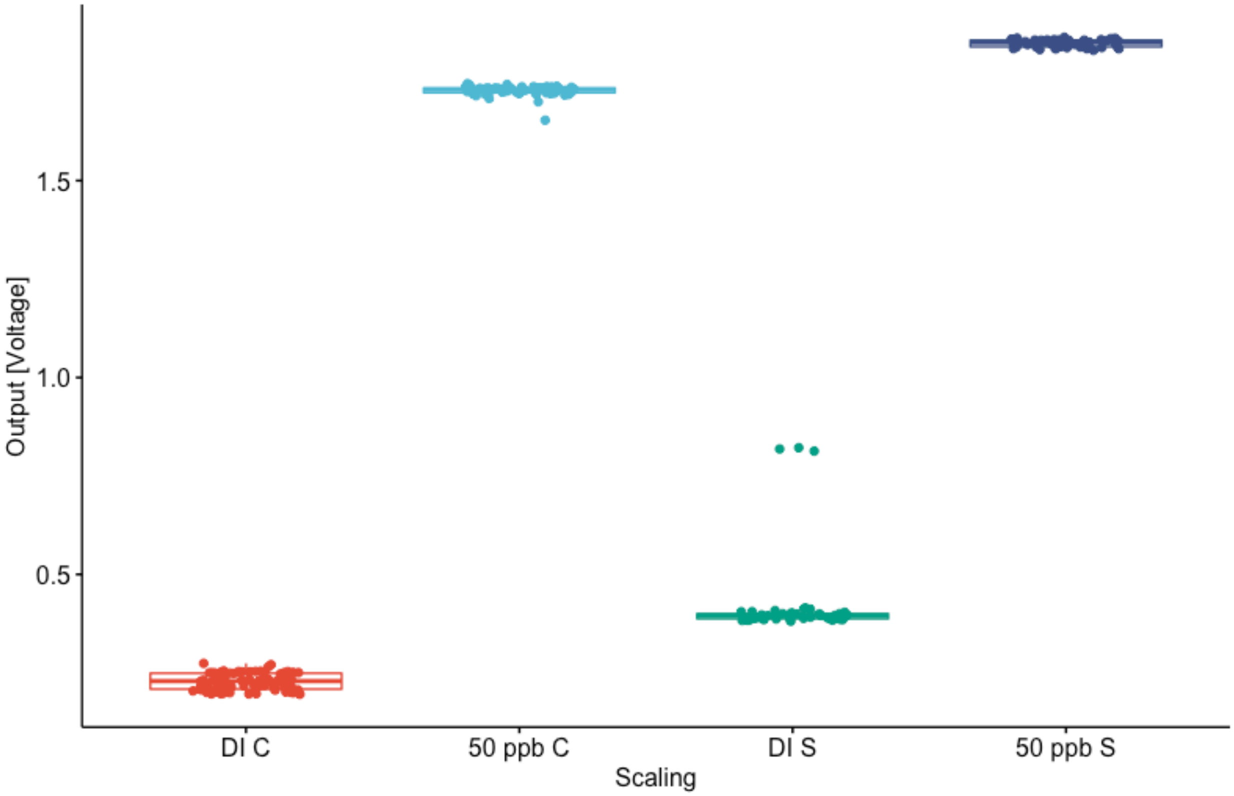


Figure 3.6: Sensor's response to exposure to mineral scaling solution. "C" represents a clean cuvette and "S" represents after the cuvette was exposed to scaling solution.

3.4 Conclusion

3.4.1 pH Impact

pH had an impact on the sensor's signal both above and below a pH range of 6 to 8. Since most natural water samples have a pH between 6 and 8, it is sufficient to measure fluorescence without monitoring pH. The sensor should not be used for water samples that have a pH outside of this range. The impact of pH on fluorescence signal can vary based on contents in the water, so these results may not be replicable when tested with natural waters, but multiple other studies show similar results with varying waters (Spencer et al., 2007).

3.4.2 Turbidity Impact

The increase in sensor output for all levels of turbidity up to 225 NTU does not conform to the findings of Khamis et al. 2015. In that study they found a rapid increase in readings to a maxiumum between 25-100 NTU and then a rapid decrease to 600 NTU. They found signal attenuation at turbidity greater than 200 NTU (Khamis et al., 2015). A possible explanation for our increased sensor output is possible organic coating left on the particles and not removed completely prior to running the experiment. Another possible explanation is that a short excitation wavelength is scattered very efficiently, causing more light scattering. The bandpass filter used in the sensor is centered around 357 nm with an optical window of +/- 44nm. If some of the clay happens to fluoresce in that region, it would be picked up by the SiPM. It is also possible that the removal of organic material using hydrogen peroxide reduced absorption, causing more scattering than absorption in the sample.

3.4.3 Chlorination Impact

Through adding bleach to DI water, DI with 50 ppb of tryptophan, and wastewater effluent, this study attempted to examine how chlorine impacts the sensor's signal as well as exactly what is fluorescing in these solutions. The experiment showed attenuation of the signal from bleach addition to DI water. In the tryptophan solution, all of the tryptophan was completely decomposed and the signal attenuated below that of DI. In a study by Alimova et. al, bleach completely decomposed tryptophan in lab grown *E. coli* including the destruction of the indole ring and possibly destroyed most of the cell's proteins and amino acids (Alimova et al., 2005). If this is the case, the significant signal that remains in the wastewater after the addition of bleach must be extracellular material other than tryptophan and its derivatives. These results support Sorenson et al. in suggesting that TLF fluorophores are predominately extracellular in groundwater, but more investigation should be done to determine what is fluorescing once tryptophan and it's derivatives have been decomposed (Sorensen et al., 2020). There could be overlap with another fluorophore that is continuing to persist despite the bleach.

3.4.4 Temperature correction factor

Increasing water temperature attenuates the sensor's signal. These impacts are due to an increase in temperature resulting in an increase in collisional quenching, which increases the likelihood of an excited electron to return to the ground state energy through a radiationless pathway. In other words, collisional quenching is when the excited-state fluorophore deactivates through contact with another molecule, leading to a decrease in fluorescence activity without a chemical reaction occurring (Watras et al., 2011). A correction factor was established through an exponential fit to the experimental data, this correction factor was input into the ML model to correct data with water temperature values to 20°C.

3.4.5 Fouling

Spiking the sensor with wastewater effluent caused an increase in the sensor's signal over time, even when there was no contamination present. This signal is predicted to be biofilm formation on the lenses of the cuvette, demonstrated by an increase in fluorescence of a cuvette filled with DI water by a benchtop flourimeter. This suggests that the fluorescence present is because of organic matter present on the lenses, not in the solution inside the cuvette. Exposing the sensor to a scaling solution showed a decrease in senstivity of the sensor to tryptophan solutions. Both of these fouling events show that data analysis

3.4.6 Sensor characterization

In this chapter varying water quality was tested in the sensor and its response was characterized. A correction factor for water temperature was calculated. That correction factor will be implemented when analyzing field data from the sensor. It was discovered that increasing turbidity with clay increased the sensor's signal. Further testing should be done with differing size and make up of particles to find the sensor's true limits with respect to turbidity. The sensor's response to varying pH showed that the operational limits of the sensor are between pH 6-8. Biofouling and scaling characterization showed that for this TLF sensor to be operated in-situ and monitor real time, multivariate analysis methods may need to be used for anamoly detection of fecal contamination risk levels.

Chapter 4

Development and validation of a machine learning model to characterize fecal contamination risk with a tryptophan-like fluorescence sensor

Part of this chapter appears in a paper submitted for publication in *Water Research*, with the following authorship attribution: Emily Bedell, Olivia Harmon, Katie Fankhauser, Zachery Shivers, and Evan Thomas

4.1 Introduction

As shown in the previous chapter, biofouling and mineral scaling can impact data by decreasing or increasing fluorescence signal or decreasing sensitivity. Other environmental parameters like temperature of the water, ambient temperature of the sensor, pH, and turbidity will introduce noise into the signal. For an in-situ sensor to be able to report data in real or near-time, multivariate data analysis must be conducted in order to detect anomalies or predict contamination risk levels. Anomaly detection is a process of discovering patterns in a dataset that do not conform to expected notions of normal behavior or fit in a dataset (Chandola et al., 2009). Machine learning (ML) tools have been shown to provide opportunities to overcome many limitations of in-situ, real-time monitoring sensors (Saboe et al., 2021). With the support of past data, known as training data, ML can analyze future data trends to offer insights that would not otherwise be available. ML is able to optimize data collection processes and prediction of parameters by leveraging additional datasets to decipher trends and signal patterns (Syafrudin et al., 2018). The potential to use ML to interpret

TLF data from an online sensor that will be subjected to fouling and other interferents has not yet been investigated despite successful applications for other types of water quality data (Hou et al., 2013; Jin et al., 2019; Liu et al., 2019).

Although online TLF measurements combined with ML to predict microbial water quality have not been analyzed, studies have been conducted using online sensors measuring other parameters combined with ML to predict water quality. Safi el al (2018) showed that water quality could be estimated using classical machine learning algorithms, Support Vector Machines (SVM), Neural Networks (NN), Deep Neural Networks (Deep NN), and k Nearest Neighbors (kNN). The highest accuracy prediction was 93% with Deep NN. In this study, they used sensors monitoring pH, turbidity, and temperature and classify water sources as safe if the pH is between 6.5 to 8.5 and the turbidity is below 5 NTU. Their ML models were trained with 667 samples gathered from 11 different water sources in Pakistan (Shafi et al., 2018). Sakizadeh (2016) was able to predict the water quality index (WQI) using 16 water quality parameters with artificial neural networks (ANN) with Bayesian regularization. This study showed correlation coefficients between observed and predicted values of 0.94 and 0.77, respectively (Sakizadeh, 2016). Many studies compare multiple ML methods to determine the one that predicts their desired outcome the best. One way to combine the advantages of multiple ML methods is to use Super Learner, a supervised ensemble ML tool. Super learner uses an ensemble of robust machine learning classification techniques, employing cross-validation methods to tune model parameters and protect from over-fitting data (van der Laan et al., 2007). Cross validation masks random sections of the training set and tests the performance of a learner trained on the remaining data against the masked data (Arlot and Celisse, 2010).

For our use case, it was hypothesized that a binary classification model may be appropriate to predict fecal contamination levels with the TLF sensor. Binary classification refers to a classification tast that has two class labels. In our case, each sensor location had two primary real-world conditions – they were classified as either “Below High Risk Contamination” or ”High Risk Contamination” where greater than 100CFU/100mL *E. coli* is present. As a first-order approximation to distinguish “Not High Risk Contamination” versus “High Risk Contamination” conditions, the TLF sensors indicate relative levels of TLF from when the sensor was first installed or cleaned. However, by itself this approximation is insufficient to reflect the true contamination level of the water. Fluorophores other than fecal contamination or biofilms could be causing the signal to increase (Fischer et al., 2012). Therefore, a more sophisticated anomaly classification system is required to distinguish between "High Risk Contamination", a true-positive condition, and "Not High Risk Contamination", a true-negative condition. In order to build and then validate the ML algorithms, a training set was built that reflects a ground-truth of contamination conditions.

Supervised ML can be used for a binary classification model to attempt to generate predictions that closely match outcomes of the training set. This model can then be used to generate predictions for new observations. In this case, the outcome would be high risk contamination status – "Below High Risk Contamination" vs "High Risk Contamination". These outcomes can be predicted with a number of "features" (covariates). Features are other measured variables that were somewhat predictive of the outcome. Part of the algorithm will contain "feature selection" where features that are not predictive of the outcome are ignored (Friedman, 2001). The flow chart in Fig. 4.1 describes the possible operating conditions, features, and hybrid approach used to predict "High risk contamination" vs "Below High Risk Contamination" with ML.

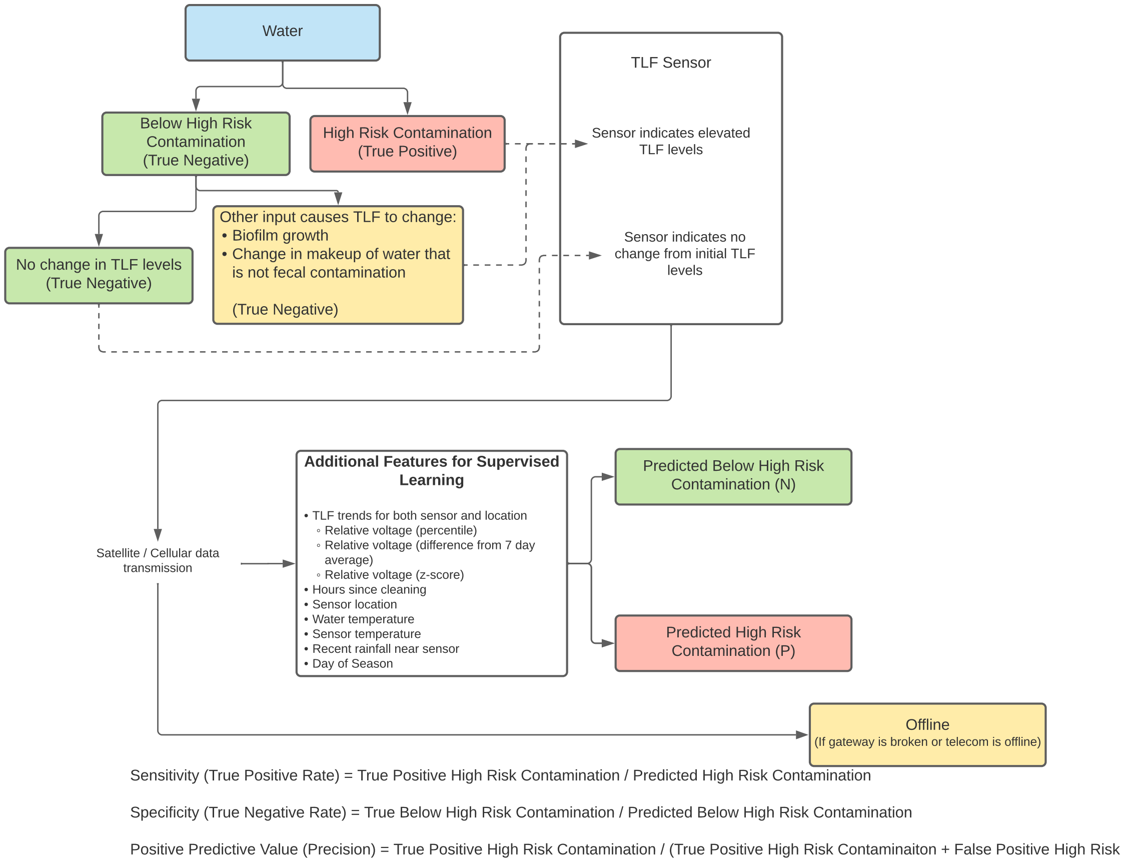
4.2 Methodology

4.2.1 Field Validation

In order to validate the sensor's functionality in the field, four sensors were placed on Boulder Creek in Boulder, Colorado in the summer of 2021 for 88 days.

4.2.1.1 Study Area

Boulder Creek flows out of the foothills and through Boulder, Colorado. The creek is a tributary of the South Platte River and its flow is primarily derived from snow melt and minor springs west of the city. Boulder has a semi-arid climate with an mean rainfall of 21 inches annually


Water
Below High Risk Contamination (True Negative)
High Risk Contamination (True Positive)
No change in TLF levels (True Negative)
Other input causes TLF to change:
• Biofilm growth
• Change in makeup of water that is not fecal contamination
(True Negative)
TLF Sensor
Sensor indicates elevated TLF levels
Sensor indicates no change from initial TLF levels
Satellite / Cellular data transmission
Additional Features for Supervised Learning
• TLF trends for both sensor and location
◦ Relative voltage (percentile)
◦ Relative voltage (difference from 7 day average)
◦ Relative voltage (z-score)
• Hours since cleaning
• Sensor location
• Water temperature
• Sensor temperature
• Recent rainfall near sensor
• Day of Season
Predicted Below High Risk Contamination (N)
Predicted High Risk Contamination (P)
Offline
(If gateway is broken or telecom is offline)

Sensitivity (True Positive Rate) = True Positive High Risk Contamination / Predicted High Risk Contamination

Specificity (True Negative Rate) = True Below High Risk Contamination / Predicted Below High Risk Contamination

Positive Predictive Value (Precision) = True Positive High Risk Contamination / (True Positive High Risk Contaminaiton + False Positive High Risk

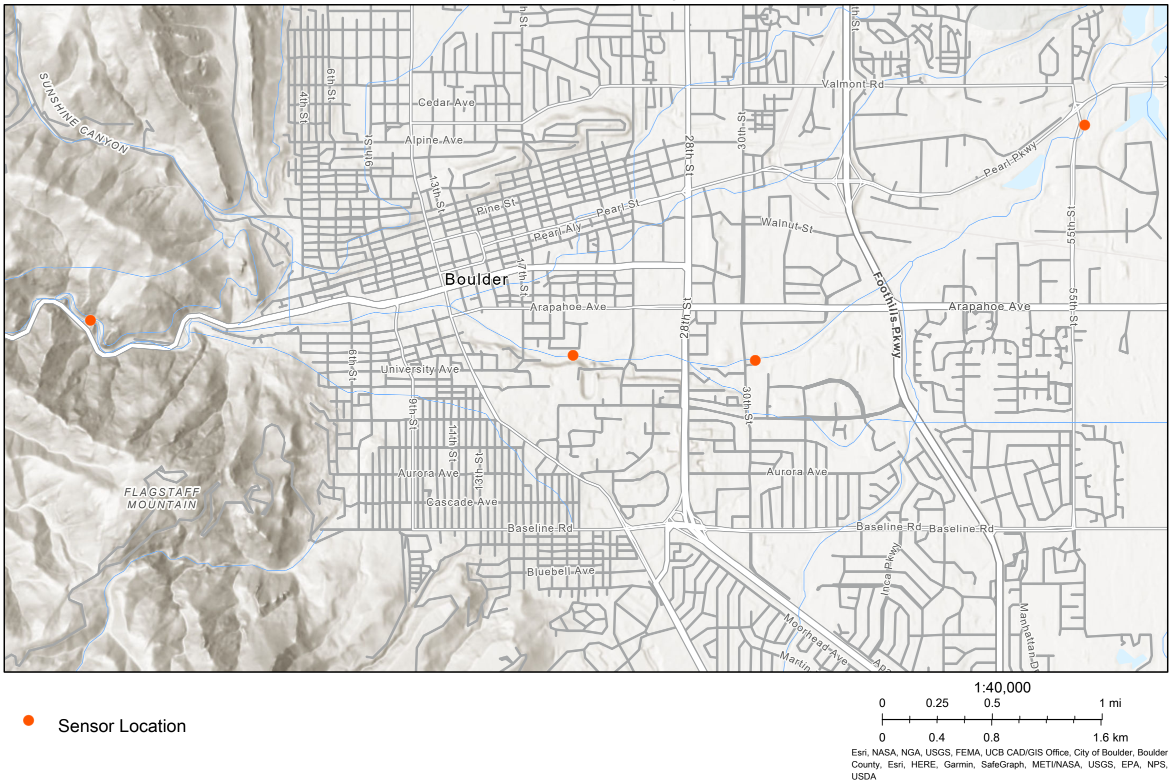
Figure 4.1: Data system flowchart showing the contamination conditions including the binary classification by the TLF sensor. Sensor data is transmitted through satellite or cellular gateway to be processed with additional features in an ML model that outputs predicted high risk contamination or not high risk contamination of the water.

(Murphy, 2006).

Four sensors were placed on Boulder Creek to monitor the fecal contamination at sites upstream, within, and downstream of the city (Fig. 4.2). The sensors were strategically placed where the City of Boulder conducts their monthly monitoring in order to compare results.

4.2.1.2 Sensor Sample Collection

Sensors were set to sample the water in the creek every 10 minutes. The sensor inlet and outlet tubes were submerged under the water, but above the creek floor by securing them to holes in a ceramic brick. The sampling sequence of the sensors was as follows:



Esri, NASA, NGA, USGS, FEMA, UCB CAD/GIS Office, City of Boulder, Boulder County, Esri, HERE, Garmin, SafeGraph, METI/NASA, USGS, EPA, NPS, USDA

Figure 4.2: Sensor and sampling locations along Boulder Creek in Boulder, Colorado, United States

Water was pumped through the sensor for 20 seconds to flush. A wait time was set to 2 seconds after pumping was complete so air bubbles were able to dissipate. To take a measurement, the LEDs were pulsed on for approximately 1000 ms and 80 readings taken from the SiPM were averaged. Water was then pumped through again for 5 seconds, and another measurement taken. This was repeated for a total of three measurements every 10 minutes.

In order to provide a diverse range of gain outputs and decrease saturation events, the current provided to the LEDs was set to 10, 50, 100, and 200mA. One data point was recorded at each current level for each measurement. The data was transmitted over cellular networks to an online database.

4.2.1.3 Ground Truth Enumeration

A training and validation data set of laboratory enumerated microbial contamination was developed in order to build and validate the ML model with the sensor data. Water samples were collected at each sensor site approximately 13 times per week for 13 weeks. Samples were collected from sensor sites within 1 minute of the sensor taking a fluorescence reading in order to match ground truth to sensor measurements. A phone based survey tool (www.mWater.co) was used to log and organize sampling data. Samples were collected in 50 mL sample bottles, put into a cooler, and transported to the lab for processing within two hours of collection. Membrane filtration was used to enumerate *E. coli* and TC. 10 mL of sample water was filtered through a 0.45 micrometer filter then incubated at 35°for 18-24 hours.

Plates were enumerated by counting the number of colonies present after incubation. The data for number of *E. coli* coliforms and TC was recorded in mWater.

The date and time of sensor installation, removal, replacement, and cleaning were also recorded in mWater.

4.2.1.4 Machine Learning Model

Two ML models were attempted. First, a binary classification of above and below high risk contamination and second, detection of contamination was semi-quantified into the five WHO risk categories.

Supervised ML models empirically find the best model fit by reducing the difference between the observed and predicted outcome. Ensemble ML (also known as Super Learning) applies multiple models (”learners”) to the same data and selects the optimal combination of them through cross-validation (van der Laan et al., 2007). For binary classification of contamination risk, there were 8 candidate learners: logistic regression, LASSO Regression, Random Forest (Pavlov, 2019), gradient boosted decision tree (XGBoost) (Chen and Guestrin, 2016), three k-Nearest Neighbors (Zhang, 2016) with k equal to 5, 10, or 15, and a null model. Logistic regression fits data to an ”S” shaped logistic function in order to classify data, providing the likelihood that a data point fits in one class or another. LASSO Regression, short for Lease Absolute Shrinkage and Selection Operator, is a regularization technique that is utilized to reduce complexity of the model. LASSO Regression uses shrinkage to shrink values towards a central point, like a mean. Random Forest classifier consists of a large number of individual decision trees that operate as an ensemble. Each tree provides a class prediction and the class with the most votes becomes the model's prediction. Gradient boost decision tree combines a series of weak decision trees that build and learn from the tree before them. K-Nearest Neighbor uses the idea of similarity, grouping parameters together that exhibit similar behavior. A null model satisfies a collection of constraints. The null model verifies whether the object in question displays some non-trivial features. (T Akinsola et al., 2017).

The multinomial classification considered up to 5 learners: LASSO Regression, Random Forest, XGBoost, independent binomials, and a null model. Binomial learners can be converted into multinomial learners by using a series of independent binomials. Ensemble learners are proven to perform as well as or better than any single candidate algorithm and minimizing cross-validated risk controls for over-fitting of the final ensemble model (van der Laan et al., 2007; Polley and van der Laan, 2010). Modeling and feature design respected an internal four-fold stratified cross-validation structure balanced on risk category.

Model features are the independent variables that can predict fecal contamination risk category. The primary explanatory feature was the retrieved voltage from the in-situ sensor, standardized and normalized. Relative voltage was explained by the z-score, difference from rolling 7-day average, and percentile. The signal voltage is dependent on the input bias voltage and this variable was included as a potential effect modifier.

As biofouling and mineral scaling are likely to impact the sensor signal, the number of hours since the optical sensor was cleaned was included as a feature in the model as a proxy for fouling (Coble et al., 2014). An estimate of daily municipal rainfall was retrieved from NOAA Physical Sciences Laboratory and included as a model feature. It is well known fecal contamination varies seasonally with rainfall (Kostyla et al., 2015). Water temperature was used as a feature and corrected to 20°C using Eq. (3.1). The temperature of the SiPM will impact it's output, for this reason temperature inside the sensor enclosure was included as a feature (Kuznetsov, 2018). Field experimentation was conducted from June - September, 2021 so a variable of consecutive days since start of the experiment attempts to capture any remaining unexplained seasonality during this summer.

The ability to identify instances of high risk contamination from the sensor data and other input variables is evaluated by a Receiver Operating Characteristic (ROC) curve and area under the curve (AUC) and cross-validated accuracy and rates of misclassification. The ROC curve is a graph displaying the performance of a classification model at all classification thresholds. The curve plots the True Positive Rate (TPR) and True Negative Rate (TNR). AUC measures the area under the ROC curve. AUC provides an aggregate measure of performance across all possible thresholds (Mandrekar, 2010). A perfect classifier has an AUC of 1 and a random classifier has an AUC of 0.5. Past studies have considered AUC values of 0.7 to 0.8, 0.8 to 0.9, and 0.9 and greater as acceptable, excellent, and outstanding, respectively (Hosmer and Lemeshow, 2013).

Discretized models output the predicted probability of a data point belonging to an outcome category. Predicted probabilities are assigned to categories based on whether they are above or below a set threshold. Determining the threshold is a trade-off between sensitivity and specificity. For the binary classification model, sensitivity is the proportion of correctly identified high risk events while specificity is the proportion of correctly identified instances of non-high risk. The ROC curve plots sensitivity by one minus specificity and visually demonstrates the compromise between sensitivity and specificity. Youden's J statistic is often used as a threshold value and assumes both are equally important. The AUC demonstrates how good the model is at discriminating high risk overall despite the choice of threshold while accuracy is the percent of data points where predicted and observed high risk agree after implementing the threshold. The null or no information rate is the expected accuracy based on the prevalence of the outcome.

The false negative rate (FNR) – one minus sensitivity – is relative to the number of underestimated instances of actual high risk. Conversely, one minus specificity is equal to the false positive rate (FPR) and is the proportion of wrongly identified cases of high risk use among observed lower than high risk contamination.

Additionally, a variable importance plot indicates the relative information gained from each feature. Variable importance was measured by the risk ratio between the full model and a model on modified data where any dependence between the outcome and the respective feature is removed by permutation (random sampling without replacement).

The performance of the multinomial model was evaluated with accuracy and true positive and negative rates.

All data management and analysis was conducted in R statistical computing software (R Core Team, 2019).

4.3 Results

4.3.1 Field Validation

Measured *E. coli* in Boulder Creek ranged from 0 to 9580 CFUs/100 mL with a mean of 120 CFUs/100 mL over 298 independent observations. The data was balanced between categories indicative of less than high risk contamination and high risk or above contamination (Table 4.1). Few observations were in the low and very high risk categories. The extreme *E. coli* measurement of 9580 CFUs/100 mL was after known fecal exposure in the creek.

Table 4.1: Prevalence and distribution of fecal contamination risk categories observed during field experimentation.

WHO Risk Category	*E. coli* CFUs/100 mL	# in Sample	% in Sample
Very low	0 - 1	19	6.4%
Low	1 - 10	18	6.0%
Intermediate	11 - 100	103	35%
High	101 - 1000	145	49%
Very high	1000+	13	4.4%
Total		298	100%

Enumeration of *E. coli* occurred at 298 independent observations (between 69 and 82 from each sampling site). The time the *E. coli* sample was taken was matched with the sensor measurement taken contemporaneously. This subset formed the training and testing datasets for modeling. Categorical *E. coli* presence in the water was the desired response variable and several inputs were hypothesized to explain whether contamination was present, namely continuous TLF measurements from the in-situ sensor.

Raw voltage readings below zero and above 2.8 volts were dropped due to sensor non-functionality or over saturation, respectively. At each reading, the sensor recorded voltage outputs at four current levels to the LEDs: 10, 50, 100, and 200 mA. For this context it was discovered that none of the current input levels performed better than the other, thus to summarize one characteristic voltage per *E. coli* enumeration, voltage output normalized by current level was averaged. Voltage from the in-situ sensor was standardized to 20°C using water temperature measurements prior to being normalized. A binary model was developed to predict whether a sample was at least at the WHO high risk level ($\geq$ 100 CFUs *E. coli* / 100 mL vs. <100 CFUs / 100 mL). Another model for multinomial classification investigated performance at predicting the correct of five WHO risk categories. Thus, continuous *E. coli* measurements where characterized by number of CFUs per 100 mL of water at very low risk (0 – 1), low risk (1 – 10), intermediate risk (11 – 100), high risk (101 – 1000), and very high risk ($\geq$1000) (WHO, 2017).

4.3.1.1 Dichotomized High Risk (100+ CFUs/100 mL) of Contamination

A machine learning model with TLF from in-situ sensors as the principal feature identified high risk of fecal contamination in natural waters with impressive skill. The ROC curve (Figure 4.3) demonstrates that both high sensitivity and specificity were achievable and that, from the AUC, the model had a probability of 86% of accurately discriminating high contamination risk. The sensitivity and specificity at the chosen predicted probability threshold were 80% and 86%, respectively.

Accuracy of detection of high risk contamination (Figure 4.4) was 83% (95% CI: 78% - 87%) and significantly different from the null information rate (53%, p-value < 0.001). True positive and true negative rates were 80% and 86%, respectively. Therefore, false positive and false negative rates were correspondingly 20% and 14%. Given the sensor alerted to high risk contamination, it

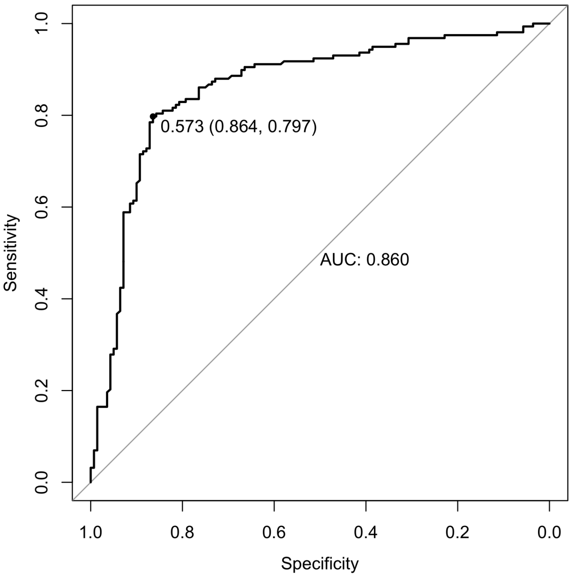
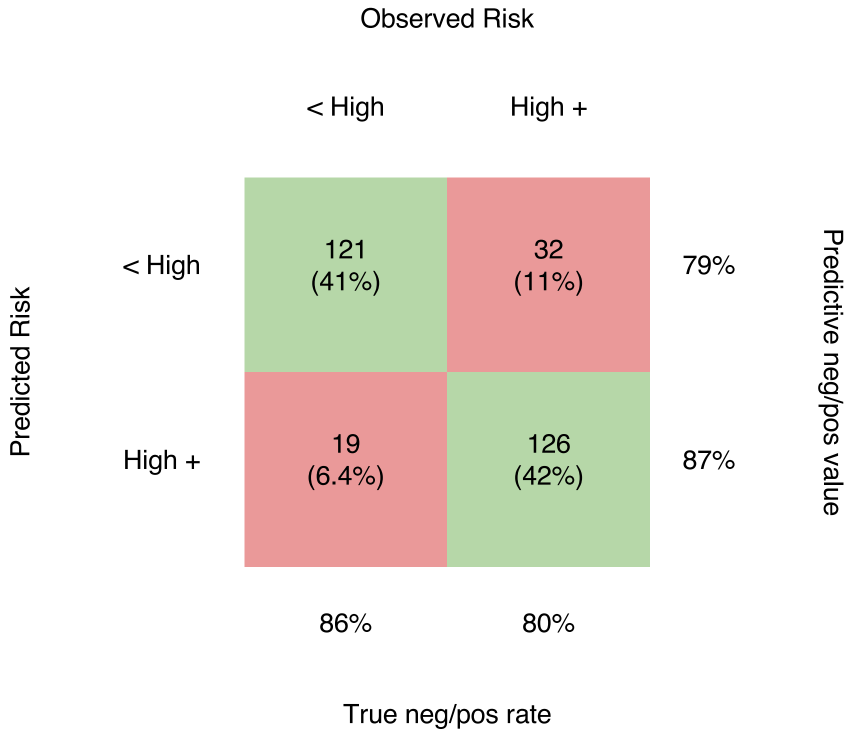


Figure 4.3: Receiver Operating Characteristic (ROC) curve. The point on the curve indicates the predicted probability threshold used to categorize high fecal contamination risk and the test specificity and sensitivity at Youden's J statistic. The area under the curve (AUC), a measure of test discrimination, is stated on the graph.

was 87% likely that the water sample contained *E. coli* at 100 CFUs/100 mL or greater, i.e. the positive predictive value.

Figure 4.5 displays the ranking of model variables, from most to least important for predictive performance. The variable importance plot confirms that TLF is highly informative to the identification of high fecal contamination risk in water when using a remote, in-situ, uninterrupted monitoring sensor. Randomizing sensor voltage increased model error by nearly 45% compared to a model preserving the observed sensor reading. The relative importance of cleaning of the sensor to remove biofilm and scaling (an increase in model risk of about 16%) indicated that these phenomena also impacted the field experiment. Relative voltage, temperature of the internal sensor board, seasonality, and the bias voltage demonstrated importance to detection of high risk contamination. The sensor voltage z-score, amount of rainfall, and water temperature did not appear informative (a risk ratio close to or less than 1.0). The effect of water temperature was incorporated when standardizing the sensor voltage to 20°C.



Observed Risk

Predicted Risk	< High	High +	Predictive neg/pos value
< High	121 (41%)	32 (11%)	79%
High +	19 (6.4%)	126 (42%)	87%
True neg/pos rate	86%	80%	

Figure 4.4: Two-by-two risk matrix for dichotiomized detection of the high risk of fecal contamination in water. The downward diagonal (green) indicates instances of accurate detection. The upward diagonal (red) are cases of over and under estimated risk.

The effect of daily streamflow on *E. coli* prevalence was investigated, but due to high missingness (53% in Boulder Creek during the summer of 2021) and the sparsity of this type of data in many low-resource contexts, it was not retained in the final model. When it was included with an additional missingness indicator (analysis not shown), accuracy to detect high risk contamination improved, but not significantly (85% vs. 83%, 95% CI: 78% - 87%).

4.3.1.2 Categorical Risk of Contamination

Overall accuracy of a model detecting one of five contamination risk categories simultaneously was 64% (95% CI: 58% - 70%) and significantly better than null accuracy (49%, p-value < 0.001). This model was still best at differentiating the high risk category with a true positive rate of 83% and

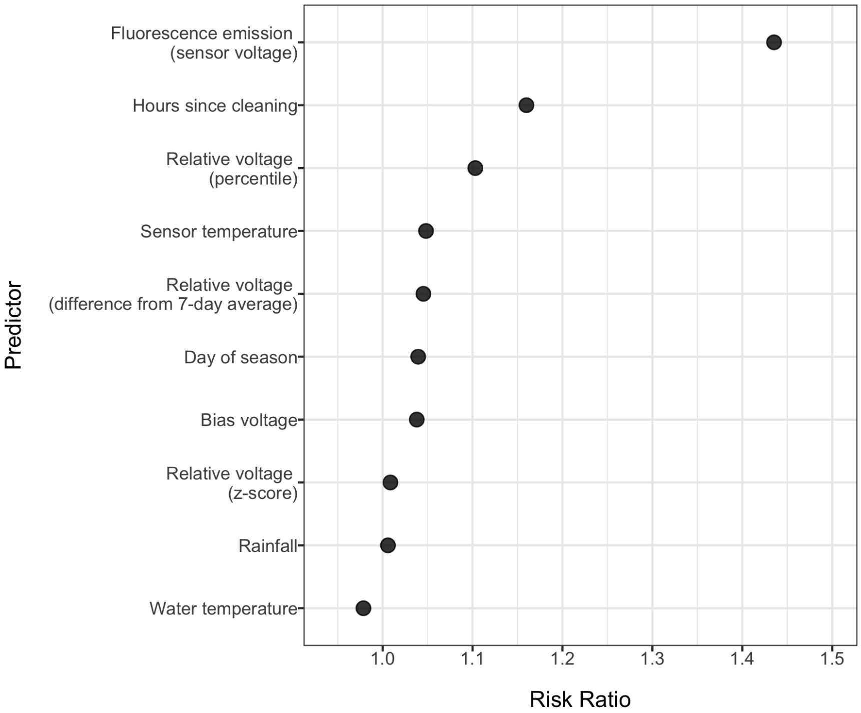
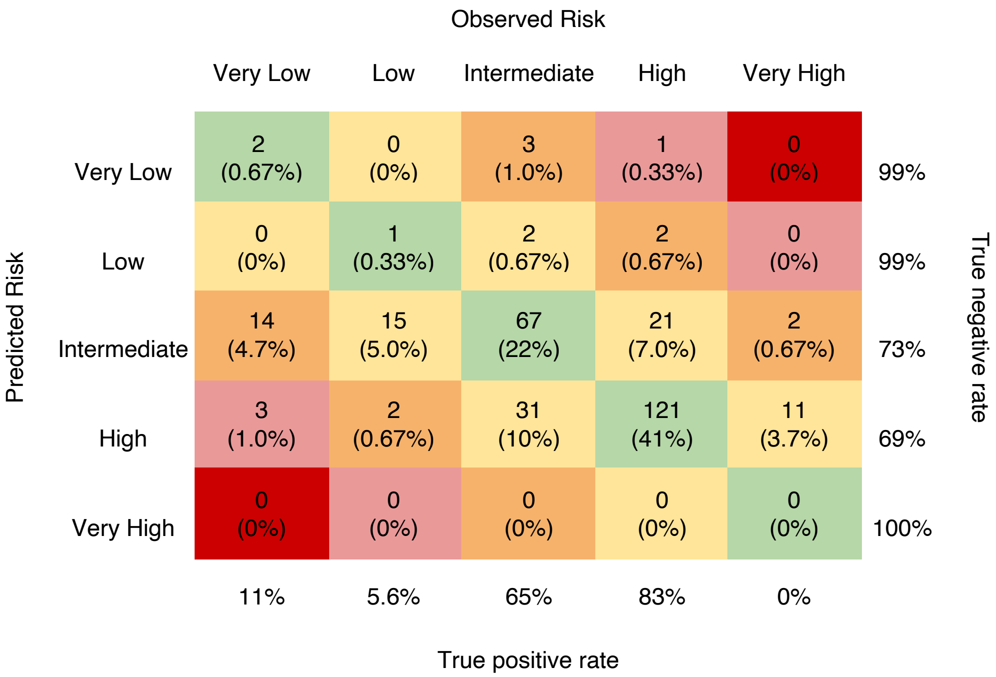


Figure 4.5: Variable importance plot and relative importance of model features in determining dichotiomized high risk detection.

a true negative rate of 69% (Figure 4.6). It performed moderately well at detecting intermediate risk with true positive and true negative rates of 65% and 73%, respectively. The combined sensor and machine learning algorithm system was not able to detect very low, low, or very high contamination. True positive rates were never greater than 11% and true negative rates were not revealing of actual performance, despite being very high, because of low sample prevalence of E. coli at these risk levels (Table 4.1). However, when very high risk was incorrectly specified, it was usually prescribed to the next closest category: 85% of incidence of E. coli over 1000 CFUs/100 mL was classified as high to very high risk.

4.4 Conclusion

Despite multiple factors interfering with the TLF signal, an ML model was able to predict high risk contamination with FPR and FNR of 20% and 14%, respectively. With an AUC of 0.86,



		Observed Risk					
		Very Low	Low	Intermediate	High	Very High	True negative rate
Predicted Risk	Very Low	2 (0.67%)	0 (0%)	3 (1.0%)	1 (0.33%)	0 (0%)	99%
	Low	0 (0%)	1 (0.33%)	2 (0.67%)	2 (0.67%)	0 (0%)	99%
	Intermediate	14 (4.7%)	15 (5.0%)	67 (22%)	21 (7.0%)	2 (0.67%)	73%
	High	3 (1.0%)	2 (0.67%)	31 (10%)	121 (41%)	11 (3.7%)	69%
	Very High	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	100%
	True positive rate	11%	5.6%	65%	83%	0%	

Figure 4.6: Risk matrix for categorical detection of the risk of fecal contamination in water.

this model would be considered to have outstanding performance. These values are higher than demonstrated by Sorenson et al. who found a FPR and FNR of 4% and 17%, respectively, but that study was done using a field portable, submersible fluorimeter that could be cleaned before each use (Sorensen et al., 2018b). The ability of the sensor to predict high risk contamination is promising because the relationship between *E. coli* and disease has exhibited a dose-response relationship. Fecal contamination's impact on public health becomes more prevalent at the intermediate, high, and and very high risk categories Hodge et al. (2016).

An online, in-situ, remotely reporting TLF sensor coupled with a ML model provides an instantaneous assessment of fecal contamination risk determined by fecal indicator organisms (FIO), namely *E. coli.* I have shown that fecal contamination risk can be assessed in near-time with high accuracy. This information could rapidly be communicated to consumers to prevent exposure, lowering rates of water quality induced disease. The integration of these sensors into a water system should not completely replace traditional FIO detection methods, but exist as an early warning system to reduce exposure while traditional testing takes place to validate contamination. This sensor combined with an ML model has the potential to revolutionize the way microbial water quality testing is conducted. If utilized by utilities, the sensor and data analysis package could assist them to expand their water quality programs because the data can be collected rapidly with minimal training requirements and no consumables for additional testing. If utilized by consumers, the sensor package could empower them to take control of their water quality and treatment system.

Chapter 5

Conclusion

5.1 Summary

This dissertation presents research investigating the design, characterization, and validation of a tryptophan-like fluorescence sensor coupled with a machine learning model to predict fecal contamination in water.

In one study, this dissertation presents design iterations to achieve high sensitivity from a flow through TLF sensor including component selection, electrical design, mechanical design, and optical optimization.

I investigated the impacts of various components and their placement on the sensor's sensitivity to tryptophan concentrations. A design limit of 1 ppb tryptophan in DI water was established as a design goal. With the integration of a SiPM and high powered LEDs, a sensitivity of 0.05 ppb tryptophan was achieved. The electrical and firmware systems were designed to provide consistent current to the LEDs and take a significant number of readings for an accurate measurement. The sensor was designed to have variable gain and range by allowing variable current inputs to the LEDs. This way the user can decide if the sensor should be more sensitive or have a higher range depending on the make up of the water being tested.

The sensor was validated in the lab by testing its sensitivity to lab grown *E. coli* and *E. coli* concentrations in wastewater dilutions. The sensor had a significant sensitivity to 33 CFU/100mL of lab grown *E. coli* and 10 CFU/100mL of *E. coli* in wastewater.

This research shows that a TLF sensor is able to significantly detect at least high risk fecal contamination in water when the sensor has been cleaned and is operating in it's optimal environment.

Next, I characterized the sensor's response to various environmental parameter impacts. Increasing water temperature decreased the signal output due to fluorescence quenching. pH outside of 6-8 significantly impacted the sensor's signal. Increasing turbidity increased the sensor's signal up to 200 NTU. Turbidity impacts could be due to significant light scattering, however the Fuller's Earth clay used in the experiment might not be indicative of turbidity occurring in surface, ground, or piped water.

Biofilm and mineral scaling impact were also characterized. Running wastewater effluent through the sensor for extended periods of time increased the sensor's signal over time even when contamination in the water was removed. A method was developed to measure the fluorescence on the surface of the cuvette using a benchtop laboratory fluorimeter. These measurements were used as a qualitative proxy for biofouling. Increased fluorescence signal in a "fouled" cuvette with DI water inside compared to a clean cuvette was inferred to be biofilm growth.

Similarly, running mineral scaling through the sensor for an extended period decreased the sensor's sensitivity. Decreased fluorescence output and increased absorbance were measured from benchtop instruments after scaling had occurred.

In a final study to validate the online functionality of the sensor to report contamination in real-time, four sensors were installed on Boulder Creek. A machine learning model was developed using a training set built by sampling the sensor locations and validating with membrane filtration. The ML model performed with great skill, with an 86% probability of accurately discriminating high risk contamination, a sensitivity of 80% and a specificity of 86%.

5.2 Key Takeaways

The studies described in this dissertation show that a TLF sensor has the capability to provide additional information to water service providers and consumers about their microbial water quality. The sensor developed requires no calibration and no reagents or dyes like many in-situ fluoresence sensors on the market. The development of a robust and reliable online TLF sensor for monitoring fecal contamination risk levels in drinking water has the potential to help avoid the need for cumbersome, time-consuming, expensive filtration, enumeration, or presence/absence techniques. It will assist in avoiding challenges with sample storage and coarse sampling designs with low temporal and spacial resolutions. The sensor can provide the ability to obtain rapid, high-quality, and highly sensitive measurements. Access to this information can reduce response times to contamination, provide information about the source of contamination, and help keep water service providers accountable for providing clean water to their customers.

Utilizing sensing technology to determine microbial water quality in a rapid, reliable, and robust manner in an indispensable advancement that can revolutionize water quality management globally. There is no perfect way to monitor microbial water quality, and there never will be, but lowering the time and cost it takes to discover contamination could save countless lives not only by reducing exposure to harmful pathogens but also providing extra funds to expand water access. Many water utilities cannot expand their piped systems because they lack the funds to ensure good water quality, purely because traditional water quality monitoring methods are so expensive.

5.2.1 Sensor Design

The development of the sensor was made possible by recent advances in silcon technology as well as miniaturization and power reduction advances in LEDs and photosensing technologies.

The most important factors of the physical sensor design were the integration of an SiPM and high powered LEDs. These components are relatively cheap, but have a significant impact on the sensitivity of the sensor to tryptophan concentrations. Using a SiPM as a highly sensitive photodiode instead of its default mode of photon counting allows for measurements that are sensitive through a large range of output vales, allowing for sensitivity and saturation control.

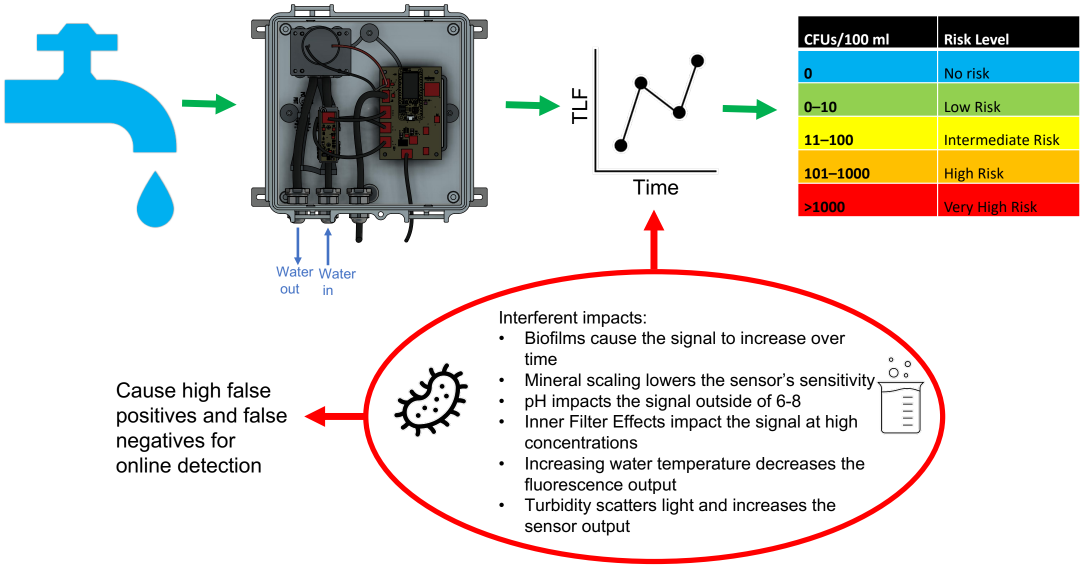
The optical components within the sensor, the cuvette and bandpass filter, are the most expensive and most complex. Both components are specialized for highly sensitive fluorescence readings. When the sensor is produced in greater quantities, it is assumed that the price of these products will drop significantly.

The integration of a signal amplifier also greatly increased the sensitivity of the sensor. Optimizing the gain of the amplifier to read the range appropriate for fecal contamination in varying waters was an important step for the functionality of the sensor.

5.2.2 Fluorescence Interferents

Environmental and physical factors have the potential to interfere with the TLF signal from the sensor. Characterizing these parameters helps us understand the extent to which these parameters should be monitored or when operational limits need to be set.

Not only did these characterization experiments help to explain noise within the sensor's field experiment but also gave input into important covariant inputs for the ML model. Fig. 5.1 shows the key takeaways from the characterization experiments and how they might impact an online sensor.


Water out
Water in
TLF
Time
Cause high false positives and false negatives for online detection

CFUs/100 ml	Risk Level
0	No risk
0–10	Low Risk
11–100	Intermediate Risk
101–1000	High Risk
>1000	Very High Risk

Interferent impacts:

- Biofilms cause the signal to increase over time
- Mineral scaling lowers the sensor's sensitivity
- pH impacts the signal outside of 6-8
- Inner Filter Effects impact the signal at high concentrations
- Increasing water temperature decreases the fluorescence output
- Turbidity scatters light and increases the sensor output

Figure 5.1: Flow diagram showing how varying physical, chemical, and environmental parameters may impact the sensor's real-time signal

There is also still potential for TLF fluorophores to originate from contamination that is not related to fecal sources such as diesel and fuel derivatives, food waste, paper mills and pesticides

(Carstea et al., 2020). The assumption is that the ML model will be able to assist in distinguishing increases related to fecal contamination and those related to other fluorophores, to some degree. Nevertheless, a certain false positive rate will always exist and an extremely high false positive rate may be alarm that another form of anthropogenic waste is present.

5.2.3 Machine Learning Capability

Many studies have suggested and implied that a data analysis system is needed to obtain maximum benefit from TLF outputs. I have shown that utilizing ensemble machine learning enables the sensor to output fecal contamination risk levels with high accuracy. Without ML, the sensor data by itself would be unusable for the prediction of fecal contamination. Since the ML model utilizes relative output from the sensor, an absolute reading from the sensor would be useless in determining contamination risk. Surface water should theoretically be one of the most challenging use cases for this sensor technology as it will have the most background fluorescence noise. Utilizing the sensor combined with the ML model in groundwater or piped, treated water should an even higher accuracy rate.

The FPR and FNR of 20% and 14%, respectively, was chosen from maximizing the Youden's index, corresponding to the point on the ROC curve with the highest vertical distance from the 45°diagonal line. A different point on the ROC curve could possibly be chosen in increase the FPR while decreasing the FNR. This circumstance may be of interest if the end user would rather experience a false positive than a false negative to have more insurance on the quality of water. For health reasons, users may prefer to be falsely alarmed of contamination than to possibly miss a contamination event.

5.3 Reflections, Recommendations, and Future work

(1) *E. coli's* utility as an indicator of contamination

Recent literature challenges *E. coli* as the preferred microbial water quality indicator (Nowicki et al., 2021). Initial findings around *E. coli* as an indicator for fecal contamination relied on the fact that *E. coli* was mainly associated with the gastrointenstinal tract and thus associated with feces from humans and animals. The secondary habitat for *E. coli*, water, sediment, soil, and flora, was thought to induce a net negative growth rate, which implied short term host persistence (Savageau, 1983). In the last few decades, this assumption has been challenged by the increased discovery and characterization of naturalized *E. coli* populations (Bergholz et al., 2011). Nowicki et al. 2021 recommends TLF sensors along side sanitary inspections and traditional water quality monitoring to begin to differentiate between naturalized *E. coli* and contamination events (Nowicki et al., 2021). Since the ML model developed in this dissertation uses *E. coli* as a training set, this model may not assist in differentiating between naturalized *E. coli* and contamination events. Further testing would need to be conducted to investigate if it would be possible to train the ML model with other forms of ground truth that may more accurately depict fecal contamination that would be harmful to human health.

(2) More extensive turbidity testing and filtration

The characterization of the sensor's response to turbidity was inconclusive. We can conclude that the sensor responds to a certain type of turbidity (Fuller's Earth Clay) by increasing its signal, but since we didn't take turbidity measurements for the field trial, we cannot conclude how the sensor responds to environmental turbidity values. More testing should be conducted with natural and treated waters with varying levels and types of turbidity to more wholly understand how turbidity impacts the sensor's signal.

No filtration was used in the field trial to filter out larger particles. This decision was made to reduce clogging or maintenance needed. In future designs, a course filter should be integrated into the design to reduce particle interference through light scattering or absorption in the readings.

(3) Surface water vs. treated drinking water

Validating the sensor's functionality on surface water is one of the more challenging use cases.

Attempting to use the sensor on groundwater or treated piped water should theoretically increase it's sensitivity, specificity, and overall accuracy. Further studies should be conducted to analyze the sensor's capability to predict contamination levels with different types of waters. The multivariate performance of the sensor to characterize all contamination risk levels was not ideal for this situation, but the results were encouraging for waters that may contain less background noise. If piped water is monitored, the model may have a different number of features. If the sensor is installed inside, rainfall would not be a contributing feature.

(4) Overall health impacts

Further testing should be conducted to determine health impacts related to predicted contamination events associated with the sensor. There is some disagreement in the literature concerning associations between *E. coli* and thermotolerant coliforms (TTCs) and diarrheal disease. Gruber et al. (2014) reviewed multiple studies and found that *E. coli* was associated with an increased risk of diarrheal disease, but that the presence of TTCs was not (Gruber et al., 2014). Hodge et al. (2016) analyzed individual-level data from seven studies and found significant increase in odds of diarrhea with increasing $log_1 0$ TTC in drinking water. They found no evidence of increased odds of diarrhea with contamination levels between 1-10 TTC/100mL (Hodge et al., 2016). Literature studying the impacts of fecal contamination on human health share a common conclusion: better water quality characterization is needed to capture the high temporal variability in water quality. With more data on the temporal variation of fecal contamination, better assessment of exposures and their effect on human health will be possible (Daly and Harris, 2021). These requests from multiple studies are encouraging for our sensor's ability monitor contamination in real time and to predict high risk spikes of contamination because they are most harmful to human health. A study on health impacts related to contamination discovered by the sensor would be an interesting investigation.

(5) Feedback informed accountability

Just as important as the functionality of the technology is its ease of use and ability to integrate into current water management practices. Now that the sensor's functionality in a certain context has been realized, the integration of the sensor and its data platform into water systems needs to be considered. Whether the sensor is used for utilities, distribution systems, or individual consumers, a sensor installation and maintenance procedure needs to be developed. Along with that, a web-based data interpretation program needs to be designed to alert water service providers or users of possible contamination in a matter that is clear and makes sense to the end user. The end goal of the sensor is to be implemented into a drinking water system to improve the service provider or consumer's capacity to detect, manage, and mitigate fecal contamination risk levels. Water quality improvement based on sensor data requires considerable on-going technical and planning support in order to maximize the benefits of the investment. However, to achieve this goal, service providers must be trained on how to read the sensor data and have the ability to understand the results produced, while also being able to apply that information to decision making around improvements in the water systems and have sufficient financial and human resources to carry out repairs, when deemed necessary.

Thomas and Brown (2020) proposed eight Characteristics of Feedback of Greatest Utility in Environmental Health and Engineering for Global Development. The first of these is that the feedback should be developed in partnership with communities and service providers. The design of the sensor feedback mechanism should be developed in partnership with end users. Their water quality information needs should be assessed and prioritized. These needs may change based on the use case of the sensor. The other feedback criterion are that that the usage of the data is incentivized, cost-effective, transparent, actionable, timely, objective, and relevant (Thomas and Brown, 2020). In all cases an alarm from the TLF sensor should trigger an action. If the sensor is being used in a water distribution or treatment system, the alarm would trigger more extensive, traditional testing to determine magnitude and origin of contamination. If the sensor is installed in a household system in combination with treatment technology, the alarm would trigger maintenance or filter replacement on the treatment technology. Incentives for these actions are clear, mitigating contamination to continue providing or consuming clean water. Utilization of sensor data to trigger contamination investigation is extremely cost effective when the costs to conduct traditional testing on a similar temporal scale is considered. The information provided from the sensor is clearly relevant, timely, and objective and the mitigation is actionable. The transparent nature of the sensor data would be up to the water service provider. Keeping this information transparent and available to consumers would support in incentivizing action or motivating consumers to treat their own water when needed.

(6) Future work: productizing the sensor for in home well water monitoring

Design work has begun on advancing the product from prototype to product. The next iteration of the design will be in partnership with a home treatment system. The ideal customer for this design will be people that utilize well water for their drinking water. Well water is particularly vulnerable to contamination through septic system malfunction and proximity to livestock. Private groundwater supplies are not regulated by the US EPA, putting the responsibly of water testing and treatment onto the homeowners (Murphy et al., 2020). Many point of use (POU) water treatment technologies have been developed that are proven to be effective at treating multiple contaminants, are highly commoditized, and cost effective, but have no way of offering real-time performance monitoring (Wu et al.). For these reasons, we are designing an in home sensor system to be paired with a treatment technology. Initial designs of this system have begun (Fig. 5.2). The sensor will be a miniturized version of the design built in this dissertation, be able to be installed in line with a houses main water line, and also have a cleaning function, where a user injects cleaning solution through the sensor when extensive fouling has occurred. The sensor's data will be sent through the user's wifi to an onine database where the ML model will be applied to predict risk level. A mobile app (Fig. 5.3) will be developed in order to alert the user if contamination is detected. Once contamination is detected, a maintenance or replacement procedure will need to take place with the treatment technology.

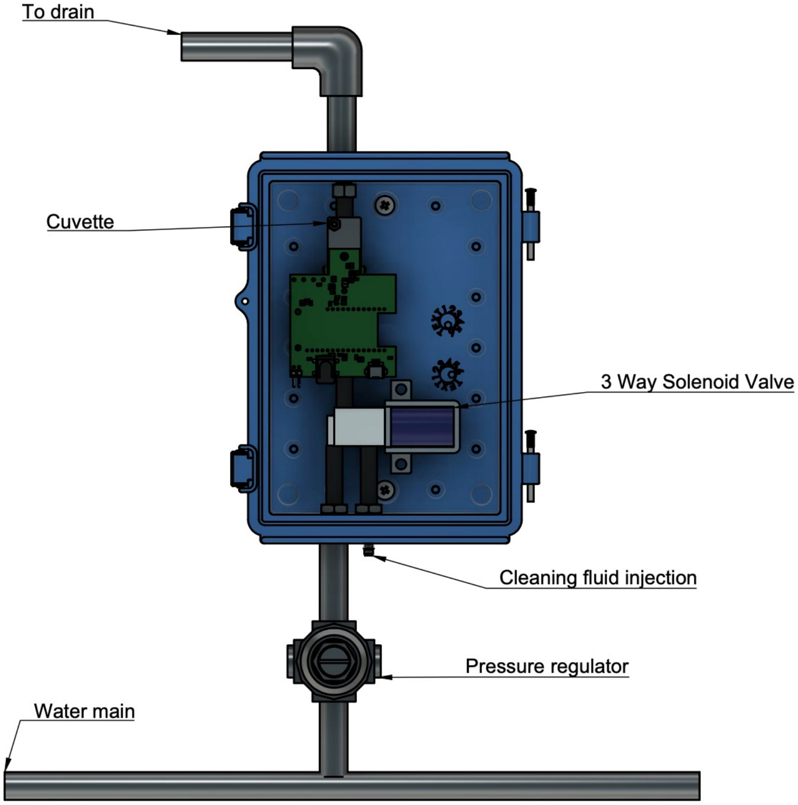


Figure 5.2: Next sensor design iteration for an in home system to be coupled with a treatment technology

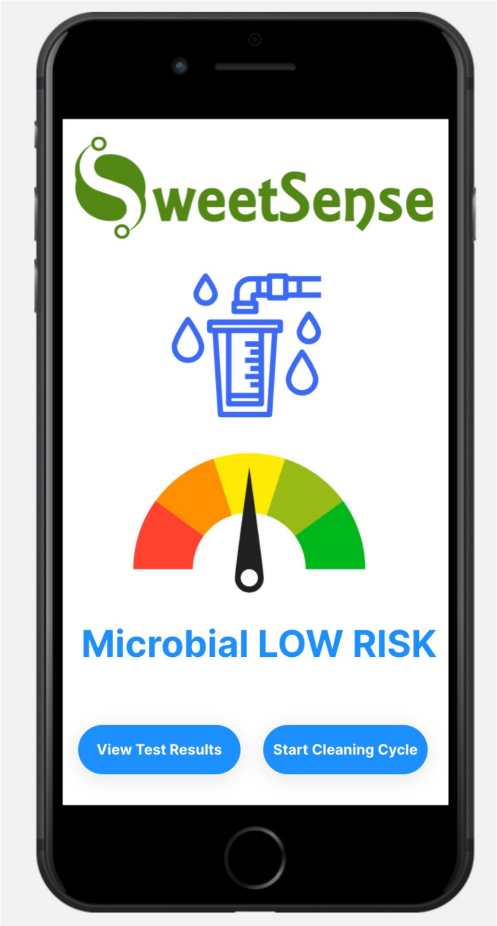


Figure 5.3: Application mock up to alert user if there is fecal contamination risk increases

Bibliography

et al. Abbafati. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet, 396(10258): 1204–1222, 2020. ISSN 1474547X. doi: 10.1016/S0140-6736(20)30925-9.

Alexandra Alimova, A. Katz, Masood Siddique, Glenn Minko, Howard E. Savage, Menhendra K. Shah, Richard B. Rosen, and Robert R. Alfano. Native fluorescence changes induced by bactericidal agents. IEEE Sensors Journal, 5(4):704–710, 2005. ISSN 1530437X. doi: 10.1109/JSEN.2005.845521.

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. Statistics Surveys, 4:40–79, 2010. ISSN 19357516. doi: 10.1214/09-SS054.

Joana Azeredo, Nuno F Azevedo, Romain Briandet, Nuno Cerca, Tom Coenye, Ana Rita Costa, Mickaël Desvaux, Giovanni Di Bonaventura, Michel Hébraud, Zoran Jaglic, Miroslava Kačániová, Susanne Knøchel, Anália Lourenço, Filipe Mergulhão, Rikke Louise Meyer, George Nychas, Manuel Simões, Odile Tresse, and Claus Sternberg. Critical review on biofilm methods. Critical reviews in microbiology, 43(3):313–351, 5 2017. ISSN 1549-7828. doi: 10.1080/1040841X.2016.1208146. URL `http://www.ncbi.nlm.nih.gov/pubmed/27868469`.

Robert Bain, Jamie Bartram, Mark Elliott, Robert Matthews, Lanakila McMahan, Rosalind Tung, Patty Chuang, and Stephen Gundry. A Summary Catalogue of Microbial Drinking Water Tests for Low and Medium Resource Settings. International Journal of Environmental Research and Public Health, 9(5):1609–1625, 5 2012. ISSN 1660-4601. doi: 10.3390/ijerph9051609. URL `http://www.ncbi.nlm.nih.gov/pubmed/22754460http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3386575http://www.mdpi.com/1660-4601/9/5/1609`.

Robert Bain, Ryan Cronk, Rifat Hossain, Sophie Bonjour, Kyle Onda, Jim Wright, Hong Yang, Tom Slaymaker, Paul Hunter, Annette Prüss-Ustün, and Jamie Bartram. Global assessment of exposure to faecal contamination through drinking water based on a systematic review. Tropical Medicine and International Health, 19(8):917–927, 2014a. ISSN 13653156. doi: 10.1111/tmi.12334.

Robert Bain, Ryan Cronk, Jim Wright, Hong Yang, Tom Slaymaker, and Jamie Bartram. Fecal Contamination of Drinking-Water in Low- and Middle-Income Countries: A Systematic Review and Meta-Analysis. PLoS Medicine, 11(5), 2014b. ISSN 15491676. doi: 10.1371/journal.pmed.1001644.

Andy Baker. Thermal fluorescence quenching properties of dissolved organic matter. Water Research, 39(18):4405–4412, 2005. ISSN 00431354. doi: 10.1016/J.WATRES.2005.08.023.

Andy Baker, Sarah Elliott, and Jamie R. Lead. Effects of filtration and pH perturbation on freshwater organic matter fluorescence. Chemosphere, 67(10):2035–2043, 5 2007. ISSN 0045-6535. doi: 10.1016/J.CHEMOSPHERE.2006.11.024. URL `https://www.sciencedirect.com/science/article/pii/S0045653506015360`.

Andy Baker, Susan A. Cumberland, Chris Bradley, Chris Buckley, and John Bridgeman. To what extent can portable fluorescence spectroscopy be used in the real-time assessment of microbial water quality? Science of the Total Environment, 532(June):14–19, 2015. ISSN 18791026. doi: 10.1016/j.scitotenv.2015.05.114.

Peter W Bergholz, Jesse D Noar, and Daniel H Buckley. Environmental Patterns Are Imposed on the Population Structure of Escherichia coli after Fecal Deposition †. APPLIED AND ENVIRONMENTAL MICROBIOLOGY, 77(1):211–219, 2011. doi: 10.1128/AEM.01880-10. URL `http://www.r-project.org`.

Marie-Claude Besner, Michèle Prévost, and Stig Regli. Assessing the public health risk of microbial intrusion events in distribution systems: Conceptual model, available data, and challenges. Water Research, 45(3):961–979, 1 2011. ISSN 0043-1354. doi: 10.1016/J.WATRES.2010.10.035. URL `https://www.sciencedirect.com/science/article/pii/S004313541000744X`.

Broadcom. AFBR-S4NxxC013-44P163 Brief Introduction to Silicon Photomultipliers. Technical report, 2019. URL `https://docs.broadcom.com/doc/AFBR-S4NxxC013-44P163-AN`.

J. M. Brown, S. Proum, and M. D. Sobsey. Escherichia coli in household drinking water and diarrheal disease risk: Evidence from Cambodia. Water Science and Technology, 58(4):757–763, 2008. ISSN 02731223. doi: 10.2166/wst.2008.439.

Joe Brown and Thomas Clasen. High Adherence Is Necessary to Realize Health Gains from Water Quality Interventions. PLoS ONE, 7(5):e36735, 5 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0036735. URL `http://www.ncbi.nlm.nih.gov/pubmed/22586491http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3346738https://dx.plos.org/10.1371/journal.pone.0036735`.

Elfrida M. Carstea, Cristina L. Popa, Andy Baker, and John Bridgeman. In situ fluorescence measurements of dissolved organic matter: A review. Science of the Total Environment, 699: 134361, 2020. ISSN 18791026. doi: 10.1016/j.scitotenv.2019.134361. URL `https://doi.org/10.1016/j.scitotenv.2019.134361`.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly Detection : A Survey. Technical report, 2009.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, volume 13-17-August-2016, 2016. doi: 10.1145/2939672.2939785.

Thomas Clasen, Wolf Peter Schmidt, Tamer Rabie, Ian Roberts, and Sandy Cairncross. Interventions to improve water quality for preventing diarrhoea: Systematic review and meta-analysis. British Medical Journal, 334(7597):782–785, 2007. ISSN 09598146. doi: 10.1136/bmj.39118.489931.BE.

P. G. Coble, J. Lead, M. Spencer, A. Baker, and D. M. Reynolds. Aquatic Organic Matter Fluorescence. 1991.

Paula Coble, Jaimie Lead, Andy Baker, Darren M. Reynolds, and Robert G.M. Spencer, editors. Aquatic Organic Matter Fluorescence. Cambridge University Press, Cambridge, 2014. ISBN 9781139045452. doi: 10.1017/CBO9781139045452. URL http://ebooks.cambridge.org/ref/id/CBO9781139045452.

Dim Coumou and Stefan Rahmstorf. A decade of weather extremes. Nature Climate Change, 2(7): 491–496, 2012. ISSN 1758678X. doi: 10.1038/nclimate1452.

Gunther F Craun, Joan M Brunkard, Jonathan S Yoder, Virginia A Roberts, Joe Carpenter, Tim Wade, Rebecca L Calderon, Jacquelin M Roberts, Michael J Beach, and Sharon L Roy. Causes of outbreaks associated with drinking water in the United States from 1971 to 2006. Clinical microbiology reviews, 23(3):507–28, 7 2010. ISSN 1098-6618. doi: 10.1128/CMR.00077-09. URL http://www.ncbi.nlm.nih.gov/pubmed/20610821http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2901654.

Susan Cumberland, John Bridgeman, Andy Baker, Mark Sterling, and David Ward. Fluorescence spectroscopy as a tool for determining microbial quality in potable water applications. Environmental Technology, 33(6):687–693, 2012. ISSN 09593330. doi: 10.1080/09593330.2011.588401.

Sean W. Daly and Angela R. Harris. Modeling Exposure to Fecal Contamination in Drinking Water due to Multiple Water Source Use. Environmental Science and Technology, 2021. ISSN 15205851. doi: 10.1021/ACS.EST.1C05683/SUPPL{_}FILE/ES1C05683{_}SI{_}001.PDF.

Stephanie DeFlorio-Barker, Abhilasha Shrestha, and Samuel Dorevitch. Estimate of burden and direct healthcare cost of infectious waterborne disease in the United States. Emerging Infectious Diseases, 27(8):2241–2242, 2021. ISSN 10806059. doi: 10.3201/eid2708.210242.

Caroline Delaire, Rachel Peletz, Emily Kumpel, Joyce Kisiangani, Robert Bain, and Ranjiv Khush. How Much Will It Cost To Monitor Microbial Drinking Water Quality in Sub-Saharan Africa? Environmental Science & Technology, 51(11):5869–5878, 6 2017. ISSN 0013-936X. doi: 10.1021/acs.est.6b06442. URL https://pubs.acs.org/doi/10.1021/acs.est.6b06442.

L. Delauney, C. Compare, and M. Lehaitre. Biofouling protection for marine environmental sensors. Ocean Science, 6(2):503–511, 2010. ISSN 18120784. doi: 10.5194/os-6-503-2010.

Sarah M Dorner, William B Anderson, Terri Gaulin, Heather L Candon, Robin M Slawson, Pierre Payment, and Peter M Huck. Pathogen and indicator variability in a heavily impacted watershed. doi: 10.2166/wh.2007.010. URL https://iwaponline.com/jwh/article-pdf/5/2/241/396706/241.pdf.

Edmund Optics. Bandpass filter.

Kyle S. Enger, Kara L. Nelson, Joan B. Rose, and Joseph N.S. Eisenberg. The joint effects of efficacy and compliance: A study of household water treatment effectiveness against childhood diarrhea. Water Research, 47(3):1181–1190, 3 2013. ISSN 18792448. doi: 10.1016/j.watres.2012.11.034.

EPA. Office of Water Definition and Procedure for the Determination of the Method Detection Limit, Revision 2. Technical report, 2016. URL www.epa.gov.

Matthias Fischer, Martin Wahl, and Gernot Friedrichs. Design and field application of a UV-LED based optical fiber biofilm sensor. Biosensors and Bioelectronics, 33(1):172–178, 3 2012. ISSN 0956-5663. doi: 10.1016/J.BIOS.2011.12.048. URL https://www.sciencedirect.com/science/article/pii/S0956566311008700.

B.G. Fox, R.M.S. Thorn, A.M. Anesio, and D.M. Reynolds. The in situ bacterial production of fluorescent organic matter; an investigation at a species level. Water Research, 125:350–359, 11 2017. ISSN 00431354. doi: 10.1016/j.watres.2017.08.040. URL https://linkinghub.elsevier.com/retrieve/pii/S0043135417307029.

Jerome H Friedman. Greedy Function Approximation: A Gradient Boosting Machine. Technical Report 5, 2001. URL https://www.jstor.org/stable/pdf/2699986.pdf?casa_token=aMtvY0Ov3gsAAAAA:XEnIL0LaCeJN4mJGpExaVt4vUOT_450I3_UcBkGoygYEbnSfUkfuEj0y46eIpz6yRaRGrttQongzWYghEnvJFMotXkSsg-UgG7sa8KlKpQOGNE0bXwU.

David Gillett and Alan Marchiori. A low-cost continuous turbidity monitor. Sensors (Switzerland), 19(14):1–18, 2019. ISSN 14248220. doi: 10.3390/s19143039.

Frederick G.B. Goddard, Amy J. Pickering, Ayse Ercumen, Joe Brown, Howard H. Chang, and Thomas Clasen. Faecal contamination of the environment and child health: a systematic review and individual participant data meta-analysis. The Lancet Planetary Health, 4(9):e405–e415, 2020. ISSN 25425196. doi: 10.1016/S2542-5196(20)30195-9. URL http://dx.doi.org/10.1016/S2542-5196(20)30195-9.

Joshua S. Gruber, Ayse Ercumen, and John M. Colford. Coliform bacteria as indicators of diarrheal risk in household drinking water: Systematic review and meta-analysis. PLoS ONE, 9(9), 9 2014. ISSN 19326203. doi: 10.1371/journal.pone.0107429.

George G. Guilbault. Practical fluorescence : theory, methods, and techniques. Marcel Dekker, 1973. ISBN 0824712633.

Charles N Haas and L D Betz. How to average microbial densities to characterize risk. Technical Report 4, 1996.

Charles N Haas, Joan B. Rose, and Charles P. Gerba. Quantitative microbial risk assessment. 2014.

Elayse M Hachich, Marisa Di, Bari ; Ana, Paula G Christ, ; Cláudia, C Lamparelli, Solange S Ramos, Maria Inês, and Z Sato. COMPARISON OF THERMOTOLERANT COLIFORMS AND ESCHERICHIA COLI DENSITIES IN FRESHWATER BODIES. Brazilian Journal of Microbiology, pages 675–681, 2012. ISSN 1517-8382.

A. C. Hambly, R. K. Henderson, M. V. Storey, A. Baker, R. M. Stuetz, and S. J. Khan. Fluorescence monitoring at a recycled water treatment plant and associated dual distribution system - Implications for cross-connection detection. Water Research, 44(18):5323–5333, 2010. ISSN 00431354. doi: 10.1016/j.watres.2010.06.003.

James Hodge, Howard H. Chang, Sophie Boisson, Simon M. Collin, Rachel Peletz, and Thomas Clasen. Assessing the Association between Thermotolerant Coliforms in Drinking Water and Diarrhea: An Analysis of Individual–Level Data from Multiple Studies. Environmental Health Perspectives, 124(10):1560–1567, 10 2016. ISSN 0091-6765. doi: 10.1289/EHP156. URL https://ehp.niehs.nih.gov/doi/10.1289/EHP156.

David W. Hosmer and Stanley Lemeshow. Applied Logistic Regression. Wiley & Sons, Inc., 2013.

Dibo Hou, Huimei He, Pingjie Huang, Guangxin Zhang, and Hugo Loaiciga. Measurement Science and Technology Detection of water-quality contamination events based on multi-sensor fusion using an extented Dempster-Shafer method Detection of water-quality contamination events based on multi-sensor fusion using an extented Dempster-Shafer method. Meas. Sci. Technol, 24: 55801–55819, 2013. doi: 10.1088/0957-0233/24/5/055801.

Naomi Hudson, Andy Baker, and Darren Reynolds. Fluorescence analysis of dissolved organic matter in natural, waste and polluted waters - A review. River Research and Applications, 23(6): 631–649, 2007. ISSN 15351459. doi: 10.1002/rra.1005.

Naomi Hudson, Andy Baker, David Ward, Darren M. Reynolds, Chris Brunsdon, Cynthia Carliell-Marquet, and Simon Browning. Can fluorescence spectrometry be used as a surrogate for the Biochemical Oxygen Demand (BOD) test in water quality assessment? An example from South West England. Science of the Total Environment, 391(1):149–158, 2008. ISSN 00489697. doi: 10.1016/j.scitotenv.2007.10.054.

Tao Jin, Shaobin Cai, Dexun Jiang, and Jie Liu. A data-driven model for real-time water quality prediction and early warning by an integration method. Environmental Science and Pollution Research, 26(29):30374–30385, 10 2019. ISSN 16147499. doi: 10.1007/s11356-019-06049-2.

Tyler D. Johnson, Kenneth Belitz, and Melissa A. Lombard. Estimating domestic well locations and populations served in the contiguous U.S. for years 2000 and 2010. Science of The Total Environment, 687:1261–1273, 10 2019. ISSN 0048-9697. doi: 10.1016/J.SCITOTENV.2019.06.036. URL `https://www.sciencedirect.com/science/article/pii/S0048969719325963?via%3Dihub`.

Kathleen Joslyn and John Lipor. A Supervised Learning Approach to Water Quality Parameter Prediction and Fault Detection. In 2018 IEEE International Conference on Big Data (Big Data), pages 2511–2514. IEEE, 12 2018. ISBN 978-1-5386-5035-6. doi: 10.1109/BigData.2018.8622628. URL `https://ieeexplore.ieee.org/document/8622628/`.

K. Khamis, J. P.R. Sorensen, C. Bradley, D. M. Hannah, D. J. Lapworth, and R. Stevens. In situ tryptophan-like fluorometers: Assessing turbidity and temperature effects for freshwater applications. Environmental Sciences: Processes and Impacts, 17(4):740–752, 2015. ISSN 20507895. doi: 10.1039/c5em00030k.

Caroline Kostyla, Rob Bain, Ryan Cronk, and Jamie Bartram. Seasonal variation of fecal contamination in drinking water sources in developing countries: A systematic review. Science of The Total Environment, 514:333–343, 5 2015. ISSN 0048-9697. doi: 10.1016/J.SCITOTENV.2015.01.018.

Daniel Krewski, John Balbus, David Butler-Jones, Charles N. Haas, Judith Isaac-Renton, Kenneth J. Roberts, and Martha Sinclair. Managing the microbiological risks of drinking water. In Journal of Toxicology and Environmental Health - Part A, volume 67, pages 1591–1617, 10 2004. doi: 10.1080/15287390490491909.

Mikael Kubista, Robert Sjöback, Svante Eriksson, and Bo Albinsson. Experimental correction for the inner-filter effect in fluorescence spectra. Analyst, 119(3):417–419, 1 1994. ISSN 1364-5528. doi: 10.1039/AN9941900417. URL `https://pubs.rsc.org/en/content/articlehtml/1994/an/an9941900417https://pubs.rsc.org/en/content/articlelanding/1994/an/an9941900417`.

Evgeny Kuznetsov. Temperature-compensated silicon photomultiplier. Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 912(November 2017):226–230, 2018. ISSN 01689002. doi: 10.1016/j.nima.2017.11.060. URL `https://doi.org/10.1016/j.nima.2017.11.060`.

Joseph R Lakowicz. Principles of Fluorescence Spectroscopy Third Edition. Technical report. URL `https://link.springer.com/content/pdf/10.1007%2F978-0-387-46312-4.pdf`.

Jin Hyung Lee, Thomas K. Wood, and Jintae Lee. Roles of indole as an interspecies and interkingdom signaling molecule. Trends in Microbiology, 23(11):707–718, 2015. ISSN 18784380. doi: 10.1016/j.tim.2015.08.001. URL `http://dx.doi.org/10.1016/j.tim.2015.08.001`.

Karen Levy, Alan E. Hubbard, Kara L. Nelson, and Joseph N.S. Eisenberg. Drivers of water quality variability in northern coastal Ecuador. Environmental Science and Technology, 43(6):1788–1797, 3 2009. ISSN 0013936X. doi: 10.1021/es8022545.

Gang Li and Kevin D. Young. Indole production by the tryptophanase TnaA in escherichia coli is determined by the amount of exogenous tryptophan. Microbiology (United Kingdom), 159(2): 402–410, 2013. ISSN 13500872. doi: 10.1099/mic.0.064139-0.

Runze Li, Dinesh Dhankhar, Jie Chen, Thomas C Cesario, and Peter M Rentzepis. A tryptophan synchronous and normal fluorescence study on bacteria inactivation mechanism. 116:18822–18826, 2019. doi: 10.1073/pnas.1909722116. URL `www.pnas.org/cgi/doi/10.1073/pnas.1909722116`.

Jie Liu, Peng Wang, Dexun Jiang, Jun Nan, and Weiyu Zhu. An integrated data-driven framework for surface water quality anomaly detection and early warning. 2019. doi: 10.1016/j.jclepro.2019.119145. URL `https://doi.org/10.1016/j.jclepro.2019.119145`.

Vadim B Malkov. Comparison of On-line Chlorine Analysis Methods and Instrumentation Built on Amperometric and Colorimetric Technologies. Technical report, 2009. URL `https://www.hach.com/cms-portals/hach_com/cms/documents/pdf/Application-CaseHistory-Whitepaper/ComparisonofOn-lineChlorineAnalysis.pdf`.

Jayawant N Mandrekar. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. Technical Report 9, 2010.

Derek V Manov, Grace C Chang, and Tommy D Dickey. Methods for Reducing Biofouling of Moored Optical Sensors. Technical report.

Michael Messner. An approach for developing a national estimate of waterborne disease due to drinking water and a national estimate model application. Journal of Water and Health, 4: 201–240, 2006. ISSN 14778920. doi: 10.2166/wh.2006.024.

Amirmasoud Mohtasebi, Andrew D. Broomfield, Tanzina Chowdhury, P. Ravi Selvaganapathy, and Peter Kruse. Reagent-Free Quantification of Aqueous Free Chlorine via Electrical Readout of Colorimetrically Functionalized Pencil Lines. ACS Applied Materials & Interfaces, 9(24): 20748–20761, 6 2017. ISSN 1944-8244. doi: 10.1021/acsami.7b03968. URL `https://pubs.acs.org/doi/10.1021/acsami.7b03968`.

Eric S Money, Gail P Carter, and Marc L Serre. Modern space/time geostatistics using river distances: data integration of turbidity and E. coli measurements to assess fecal contamination along the Raritan River in New Jersey. Environmental science & technology, 43(10):3736–42, 5 2009. ISSN 0013-936X. doi: 10.1021/es803236j. URL http://www.ncbi.nlm.nih.gov/pubmed/19544881http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2752213.

Heather M. Murphy, Shannon McGinnis, Ryan Blunt, Joel Stokdyk, Jingwei Wu, Alexander Cagle, Donna M. Denno, Susan Spencer, Aaron Firnstahl, and Mark A. Borchardt. Septic Systems and Rainfall Influence Human Fecal Marker and Indicator Organism Occurrence in Private Wells in Southeastern Pennsylvania. Environmental Science and Technology, 54(6):3159–3168, 3 2020. ISSN 15205851. doi: 10.1021/acs.est.9b05405.

Kathleen R. Murphy, Adam Hambly, Sachin Singh, Rita K. Henderson, Andy Baker, Richard Stuetz, and Stuart J. Khan. Organic matter fluorescence in municipal water recycling schemes: Toward a unified PARAFAC model. Environmental Science and Technology, 45(7):2909–2916, 4 2011. ISSN 0013936X. doi: 10.1021/es103015e.

Sheila F. Murphy. State of the watershed: Water quality of Boulder Creek, Colorado. Number 1284. 2006. ISBN 1411309545. doi: 10.3133/cir1284.

Gordon Nichols, Iain Lake, and Clare Heaviside. Climate change and water-related infectious diseases. Atmosphere, 9(10):1–60, 2018. ISSN 20734433. doi: 10.3390/atmos9100385.

Saskia Nowicki, Dan J. Lapworth, Jade S.T. Ward, Patrick Thomson, and Katrina Charles. Tryptophan-like fluorescence as a measure of microbial contamination risk in groundwater. Science of The Total Environment, 646:782–791, 1 2019. ISSN 00489697. doi: 10.1016/j.scitotenv.2018.07.274. URL https://linkinghub.elsevier.com/retrieve/pii/S0048969718327700.

Saskia Nowicki, Zaydah R. DeLaurent, Etienne P. De Villiers, George Githinji, and Katrina J. Charles. The utility of Escherichia coli as a contamination indicator for rural drinking water: Evidence from whole genome sequencing. PLoS ONE, 16(1 January), 2021. ISSN 19326203. doi: 10.1371/journal.pone.0245910. URL http://dx.doi.org/10.1371/journal.pone.0245910.

Takuya Okazaki, Tatsuya Orii, Akira Ueda, Akiko Ozawa, and Hideki Kuramitz. Fiber Optic Sensor for Real-Time Sensing of Silica Scale Formation in Geothermal Water. Scientific Reports, 7(1): 3387, 12 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-03530-1. URL http://www.nature.com/articles/s41598-017-03530-1.

Yu L. Pavlov. Random forests. Random Forests, pages 1–122, 2019. doi: 10.1201/9780429469275-8.

Amy J. Pickering, Clair Null, Peter J. Winch, Goldberg Mangwadu, Benjamin F. Arnold, Andrew J. Prendergast, Sammy M. Njenga, Mahbubur Rahman, Robert Ntozini, Jade Benjamin-Chung, Christine P. Stewart, Tarique M.N. Huda, Lawrence H. Moulton, John M. Colford, Stephen P. Luby, and Jean H. Humphrey. The WASH Benefits and SHINE trials: interpretation of WASH intervention effects on linear growth and diarrhoea. The Lancet Global Health, 7(8):e1139–e1146, 2019. ISSN 2214109X. doi: 10.1016/S2214-109X(19)30268-2. URL http://dx.doi.org/10.1016/S2214-109X(19)30268-2.

Eric C Polley and Mark J van der Laan. Super Learner in Prediction. U.C. Berkeley Division of Biostatistics Working Paper, (266), 2010.

Wendy Pons, Ian Young, Jenifer Truong, Andria Jones-Bitton, Scott McEwen, Katarina Pintar, and Andrew Papadopoulos. A systematic review of waterborne disease outbreaks associated with small non-community drinking water systems in Canada and the United States. PLoS ONE, 10 (10):1–17, 2015. ISSN 19326203. doi: 10.1371/journal.pone.0141646.

Annette Prüss-Ustün, Jennyfer Wolf, Jamie Bartram, Thomas Clasen, Oliver Cumming, Matthew C. Freeman, Bruce Gordon, Paul R. Hunter, Kate Medlicott, and Richard Johnston. Burden of disease from inadequate water, sanitation and hygiene for selected adverse health outcomes: An updated analysis with a focus on low- and middle-income countries. International Journal of Hygiene and Environmental Health, 222(5):765–777, 2019. ISSN 1618131X. doi: 10.1016/j.ijheh.2019.05.004. URL `https://doi.org/10.1016/j.ijheh.2019.05.004`.

Matthias Pucher, Urban Wünsch, Gabriele Weigelhofer, Kathleen Murphy, Thomas Hein, and Daniel Graeber. StaRdom: Versatile software for analyzing spectroscopic data of dissolved organic matter in R. Water (Switzerland), 11(11), 2019. ISSN 20734441. doi: 10.3390/w11112366.

R Core Team. R: A language and environment for statistical computing., 2019.

Anditya Rahardianto, Wen Yi Shih, Ron Wai Lee, and Yoram Cohen. Diagnostic characterization of gypsum scale formation and control in RO membrane desalination of brackish water. Journal of Membrane Science, 279(1-2), 2006. ISSN 03767388. doi: 10.1016/j.memsci.2005.12.059.

D.M. Reynolds. Rapid and direct determination of tryptophan in water using synchronous fluorescence spectroscopy. Water Research, 37(13):3055–3060, 7 2003. ISSN 00431354. doi: 10.1016/S0043-1354(03)00153-2. URL `http://www.ncbi.nlm.nih.gov/pubmed/14509692https://linkinghub.elsevier.com/retrieve/pii/S0043135403001532`.

Daniel Saboe, Hamidreza Ghasemi, Ming Ming Gao, Mirjana Samardzic, Kiril D. Hristovski, Dragan Boscovic, Scott R. Burge, Russell G. Burge, and David A. Hoffman. Real-time monitoring and prediction of water quality parameters and algae concentrations using microbial potentiometric sensor signals and machine learning tools. Science of the Total Environment, 764, 4 2021. ISSN 18791026. doi: 10.1016/j.scitotenv.2020.142876.

Mohamad Sakizadeh. Artificial intelligence for the prediction of water quality index in groundwater systems. Modeling Earth Systems and Environment, 2(1), 3 2016. ISSN 23636211. doi: 10.1007/s40808-015-0063-9.

Michael A Savageau. ESCHERICHIA COLI HABITATS, CELL TYPES, AND MOLECULAR MECHANISMS OF GENE CONTROL. Technical Report 6, 1983.

Seoul Viosys. Deep UV LED - 275nm Product Brief. Technical Report July, 2018.

Uferah Shafi, Rafia Mumtaz, Hirra Anwar, Ali Mustafa Qamar, and Hamza Khurshid. Surface Water Pollution Detection using Internet of Things. In 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT and IoT, HONET-ICT 2018, pages 92–96. Institute of Electrical and Electronics Engineers Inc., 11 2018. ISBN 9781538683545. doi: 10.1109/HONET.2018.8551341.

João Simões and Tao Dong. Continuous and Real-Time Detection of Drinking-Water Pathogens with a Low-Cost Fluorescent Optofluidic Sensor. Sensors (Basel, Switzerland), 18(7), 7 2018. ISSN 1424-8220. doi: 10.3390/s18072210. URL `http://www.ncbi.nlm.nih.gov/pubmed/29996477http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6068492`.

João Simões, Zhaochu Yang, and Tao Dong. An Ultrasensitive Fluorimetric Sensor for Pre-Screening of Water Microbial Contamination Risk. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 258:119805, 2021. ISSN 13861425. doi: 10.1016/j.saa.2021.119805.

Torben Lund Skovhus and Bo Højris, editors. Microbiological Sensors for the Drinking Water Industry. International Water Association, 10 2018. ISBN 9781780408699. doi: 10.2166/9781780408699. URL `https://iwaponline.com/ebooks/book/732/Microbiological-Sensors-for-the-Drinking-Water`.

J. P R Sorensen, D. J. Lapworth, B. P. Marchant, D. C W Nkhuwa, S. Pedley, M. E. Stuart, R. A. Bell, M. Chirwa, J. Kabika, M. Liemisa, and M. Chibesa. In-situ tryptophan-like fluorescence: A real-time indicator of faecal contamination in drinking water supplies. Water Research, 81:38–46, 2015. ISSN 18792448. doi: 10.1016/j.watres.2015.05.035. URL `http://dx.doi.org/10.1016/j.watres.2015.05.035`.

J. P.R. Sorensen, A. Vivanco, M. J. Ascott, D. C. Gooddy, D. J. Lapworth, D. S. Read, C. M. Rushworth, J. Bucknall, K. Herbert, I. Karapanos, L. P. Gumm, and R. G. Taylor. Online fluorescence spectroscopy for the real-time evaluation of the microbial quality of drinking water. Water Research, 137(March):301–309, 2018a. ISSN 18792448. doi: 10.1016/j.watres.2018.03.001. URL `https://doi.org/10.1016/j.watres.2018.03.001`.

James P. R. Sorensen, Andrew F. Carr, Jacintha Nayebare, Djim M. L. Diongue, Abdoulaye Pouye, Raphaëlle Roffo, Gloria Gwengweya, Jade S. T. Ward, Japhet Kanoti, Joseph Okotto-Okotto, Laura van der Marel, Lena Ciric, Seynabou C. Faye, Cheikh B. Gaye, Timothy Goodall, Robinah Kulabako, Daniel J. Lapworth, Alan M. MacDonald, Maurice Monjerezi, Daniel Olago, Michael Owor, Daniel S. Read, and Richard G. Taylor. Tryptophan-like and humic-like fluorophores are extracellular in groundwater: implications as real-time faecal indicators. Scientific Reports, 10(1):15379, 12 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-72258-2. URL `http://www.nature.com/articles/s41598-020-72258-2`.

James P.R. Sorensen, Andy Baker, Susan A. Cumberland, Dan J. Lapworth, Alan M. MacDonald, Steve Pedley, Richard G. Taylor, and Jade S.T. Ward. Real-time detection of faecally contaminated drinking water with tryptophan-like fluorescence: defining threshold values. Science of the Total Environment, 622-623:1250–1257, 2018b. ISSN 18791026. doi: 10.1016/j.scitotenv.2017.11.162. URL `https://doi.org/10.1016/j.scitotenv.2017.11.162`.

J.P.R. Sorensen, A. Sadhu, G. Sampath, S. Sugden, S. Dutta Gupta, D.J. Lapworth, B.P. Marchant, and S. Pedley. Are sanitation interventions a threat to drinking water supplies in rural India? An application of tryptophan-like fluorescence. Water Research, 88:923–932, 1 2016. ISSN 0043-1354. doi: 10.1016/J.WATRES.2015.11.006. URL `https://www.sciencedirect.com/science/article/pii/S0043135415303341?via%3Dihub`.

Robert G.M. Spencer, Lucy Bolton, and Andy Baker. Freeze/thaw and pH effects on freshwater dissolved organic matter fluorescence and absorbance properties from a number of UK locations. Water Research, 41(13):2941–2950, 2007. ISSN 00431354. doi: 10.1016/j.watres.2007.04.012.

Muhammad Syafrudin, Ganjar Alfian, Norma Latif Fitriyani, and Jongtae Rhee. Performance analysis of IoT-based sensor, big data processing, and machine learning model for real-time monitoring system in automotive manufacturing. Sensors (Switzerland), 18(9), 9 2018. ISSN 14248220. doi: 10.3390/s18092946.

J E T Akinsola, Akinsola Jet, and Hinmikaiye J O. Supervised Machine Learning Algorithms: Classification and Comparison Supervised Machine Learning View project The Use Of BIG DATA in Mobile Analytics View project Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology, 48, 2017. ISSN 2231-2803. doi: 10.14445/22312803/IJCTT-V48P126. URL http://www.ijcttjournal.org.

Pam Tallon, Brenda Magajna, Cassandra Lofranco, and Kam Tin Leung. MICROBIAL INDICATORS OF FAECAL CONTAMINATION IN WATER: A CURRENT PERSPECTIVE. Technical report. URL https://link.springer.com/content/pdf/10.1007%2Fs11270-005-7905-4.pdf.

Evan Thomas and Joe Brown. Using Feedback to Improve Accountability in Global Environmental Health and Engineering. Environmental Science & Technology, page acs.est.0c04115, 12 2020. ISSN 0013-936X. doi: 10.1021/acs.est.0c04115. URL https://pubs.acs.org/doi/10.1021/acs.est.0c04115.

UNICEF. Target Product Profile - Rapid E. coli Detection Tests. 2019. URL https://www.unicef.org/supply/media/2511/file/Rapid-coli-detection-TPP-2019.pdf.

UNICEF/WHO. Progress on Household Drinking Water ,. 2021.

US EPA. Analytical Methods Approved for Drinking Water Compliance Monitoring under the Revised Total Coliform Rule. Technical report, 2017.

Mark J. van der Laan, Eric C Polley, and Alan E. Hubbard. Super Learner. Statistical Applications in Genetics and Molecular Biology, 6(1), 1 2007. ISSN 1544-6115. doi: 10.2202/1544-6115.1309. URL https://www.degruyter.com/view/j/sagmb.2007.6.issue-1/sagmb.2007.6.1.1309/sagmb.2007.6.1.1309.xml.

P. Vidon, L. P. Tedesco, J. Wilson, M. A. Campbell, L. R. Casey, and Mark Gray. Direct and Indirect Hydrological Controls on ¡i¿E. coli¡/i¿ Concentration and Loading in Midwestern Streams. Journal of Environmental Quality, 37(5):1761–1768, 9 2008. ISSN 00472425. doi: 10.2134/jeq2007.0311. URL http://doi.wiley.com/10.2134/jeq2007.0311.

Jade S.T. Ward, Daniel J. Lapworth, Daniel S. Read, Steve Pedley, Sembeyawo T. Banda, Maurice Monjerezi, Gloria Gwengweya, and Alan M. MacDonald. Large-scale survey of seasonal drinking water quality in Malawi using in situ tryptophan-like fluorescence and conventional water quality indicators. Science of the Total Environment, 744:140674, 2020. ISSN 18791026. doi: 10.1016/j.scitotenv.2020.140674. URL https://doi.org/10.1016/j.scitotenv.2020.140674.

C.J. Watras, P.C. Hanson, T.L. Stacy, K.M. Morrison, J. Mather, Y.-H. Hu, and P. Milewski. A temperature compensation method for CDOM fluorescence sensors in freshwater. Limnology and Oceanography: Methods, 9(7):296–301, 7 2011. ISSN 15415856. doi: 10.4319/lom.2011.9.296. URL http://doi.wiley.com/10.4319/lom.2011.9.296.

WHO. The World Health Report 2005. Technical report, 2005. URL https://www.who.int/whr/2005/whr2005_en.pdf?ua=1.

WHO. Guidelines for Drinking Water Quality 2017. 2017. ISBN 9783540773405.

WHO. Drinking-water Fact Sheet. Technical report, WHO, 2019. URL https://www.who.int/en/news-room/fact-sheets/detail/drinking-water.

WHO/OECD. Assessing Microbial Safety of Drinking Water IMPROVING APPROACHES AND METHODS Published on behalf of the World Health Organization and the Organisation for Economic Co-operation and Development by. Technical report, 2002. URL `http://www.oecd.org/pdf/M000014000/M00014623.pdf`.

Jishan Wu, Miao Cao, Draco Tong, Zach Finkelstein, and Eric M V Hoek. A critical review of point-of-use drinking water treatment in the United States. doi: 10.1038/s41545-021-00128-z. URL `https://doi.org/10.1038/s41545-021-00128-z`.

Zhongheng Zhang. Introduction to machine learning: K-nearest neighbors. Annals of Translational Medicine, 4(11), 6 2016. ISSN 23055847. doi: 10.21037/ATM.2016.03.37.