NC WATER RESOURCES RESEARCH INSTITUTE

Report No. 510

FECAL CONTAMINATION SOURCE TRACKING AND FORECASTING TO SUPPORT RECREATIONAL AND CULTURAL DEVELOPMENT IN THE BLACK RIVER WATERSHED

By
Angela Harris[1], Ryan Emanuel[2], and Daniel Obenour[3]

1. Department of Civil, Construction, and Environmental Engineering
   North Carolina State University
   Raleigh, NC

2. Nicolas School of the Environment
   Duke University
   Durham, NC

wrri.ncsu.edu
NC WATER RESOURCES RESEARCH INSTITUTE

UNC-WRRI-510

The research on which this report is based was supported by funds provided by the North Carolina General Assembly and/or the US Geological Survey through the North Carolina Water Resources Research Institute.

WRRI Project No. 22-06-W
April 30, 2024

wrri.ncsu.edu
NC WATER RESOURCES RESEARCH INSTITUTE

Fecal contamination source tracking and forecasting to support recreational and cultural development in the Black River watershed

Final Report
30 April 2024

Investigators:

Angela Harris; Assistant Professor; Dept. of Civil, Construction, and Environmental Engineering; North Carolina State University (NCSU); Campus Box 7908, Raleigh NC 27695-7908; Phone: (919)-515-2402, Email: aharris5@ncsu.edu

Ryan Emanuel, Associate Professor, Nicholas School of the Environment; Duke University, Campus Box 90328, Durham NC 27708; Email: ryan.emanuel@duke.edu

Daniel Obenour, Associate Professor; Dept. of Civil, Construction, and Environmental Engineering; NCSU; Campus Box 7908, Raleigh NC 27695-7908; Phone: (919)-515-7702, Email: drobenou@ncsu.edu

Student Researchers:
Benjamin Clark (Lead, PhD student, NCSU)
Julia Harrison (PhD student, NCSU)
Andrew Daddone (Undergrad, NCSU)
Sean Daly (PhD student, NCSU)
Tanvir Pasha (PhD student, NCSU)
Sebastian Lopez (MS student, Morgan State University)
McKenna Crummey (BS student, NCSU)
Hiromu Koyama (BS student, NCSU)
Norah Patterson (BS student, NCSU)

Community Partners:
Philip Bell – Coharie River Initiative
Greg Jacobs – Coharie River Initiative
Kullen Bell – Coharie River Initiative

Fecal contamination source tracking and forecasting to support recreational and cultural development in the Black River watershed

1.0 Abstract

The Coharie Tribe expressed interest in learning more about specific CAFO and human impacts on water quality along the Great Coharie River, with special attention to spatial and temporal variability of these impacts. The Tribe's interest also aligns with opportunities to answer fundamental research questions about water quality in North Carolina's Coastal Plain and areas similarly impacted by industrialized agriculture. We proposed research questions and objectives intended to help meet these needs while also advancing fundamental science concerning biological water quality. What are the dominant sources of fecal contamination at 3 sites in the watershed? How do these sources vary across base flow and storm flow conditions, and warm and cold seasons? Can forecasting tools be developed to capture elevated levels of fecal contamination in the watershed to support Tribal decision-making related to the use of river? To address these questions, low frequency and high frequency (determined by base and storm flow) sampling events were conducted at three sites in the Great Coharie River watershed from March 2022 to March 2023. Each sample was evaluated for total coliforms, *Escherichia coli*, molecular microbial source tracking (MST) targets, Total suspended solids, Total Kjeldahl Nitrogen, the sum of Nitrate and Nitrite, Ammonia, Total Phosphorus, Orthophosphate, Temperature, pH, dissolved oxygen, specific conductivity, and turbidity. An automated DNA extraction kit was optimized for use with surface water filters and nucleic acid was extracted from each sample. Following DNA extraction, samples were analyzed using polymerase chain reaction microbial source tracking targets including LA35, Pig2bac and hCYTB484 which is sensitive and specific to swine and poultry combined, swine, and human fecal contamination, respectively. Results from the MST assays indicate that the watershed likely suffers from fecal contamination from the three target hosts. *Escherichia coli* data collected throughout the study indicate that the three sites have contamination levels that would deem the Great Coharie River as impaired per the United States Environmental Protection Agency's Recreational Water Quality Criteria. An explanatory model using multiple linear regression was developed with the available data. The preliminary model revealed that 2-day total rainfall is an important covariate to help explain *E. coli* concentrations in the watershed. In addition to the explanatory model, a random forest predictive model was constructed to predict whether the water was at an *E. coli* concentration above or below the EPA recommended threshold (i.e., 126 MPN/100 mL). When performing covariate importance analysis, it was determined that 2-day total rainfall was the most important covariate of any examined in this study when predicting *E. coli*. Lastly, a generalized estimating equation was developed, and this model also found 2-day total rainfall to be an important covariate. It is recommended that caution be used during primary contact recreation, which includes activities that involve increased likelihood of water ingestion or immersion like swimming, at each of the three sites due to elevated levels of *E. coli*- especially following an intense rainfall event. We also recommend that a land use study be undertaken to determine potential management strategies to minimize fecal contamination at the Marsh Swamp site, which experienced higher levels of *E. coli* contamination than other sites. Additionally, further studies into the human contributions to the watershed could be performed to help identify if further interventions related to human waste management need to be made.

2.0 Acknowledgments

This work was performed in part at the Environmental and Agricultural Testing Service laboratory (EATS), Department of Crop and Soil Sciences, at North Carolina State University. Special thanks are due to the Coharie Tribe who generously offered students working space to process samples during intensive sampling operations.

3.0 Introduction

3.1 The Coharie Tribe have a special connection with the river

The Coharie Tribe of North Carolina consider themselves "river people," and share their name with the Great Coharie River, in whose watershed they have lived for centuries. The river, which drains much of Sampson County, North Carolina, holds immense cultural value to the Tribe (Emanuel, 2024). Recently the Tribe has been concerned with the declining water quality and environmental injustices that have been documented in the area (Harris et al., 2021; Quist et al., 2022; Strickling & Obenour, 2018; Wing et al., 2000). The Coharie Tribe has expressed an interest in learning more about specific CAFO and human impacts on water quality along the Great Coharie River, with special attention to spatial and temporal variability of these impacts. The Tribe's interest also aligns with opportunities to answer fundamental research questions about water quality in North Carolina's Coastal Plain and areas similarly impacted by industrialized agriculture. We propose research questions and objectives intended to help meet these needs while also advancing fundamental science concerning biological water quality. What are the dominant sources of fecal contamination at 3 sites in the river shed? How do these sources vary across base flow and storm flow conditions, and warm and cold seasons? Can forecasting tools be developed to capture elevated levels of fecal contamination in the watershed to support Tribal decision-making related to the use of river?

3.2 United States Environmental Protection Agency Recommended Recreational Water Quality Criteria

In 2012 The United States Environmental Protection Agency (EPA) published new Recreational Water Quality Criteria (RWQC) (EPA, 2012). The new criteria were empirically formulated with the help of epidemiological studies, quantitative microbial risk assessments (QMRA), and experimentation with different indicators (EPA, 2012). In summary (Table 1), the EPA recommends that states decide on a level of risk that they are comfortable with and to adhere to those criteria. The EPA offers two estimated illness rates that states can create legislation on, gastrointestinal illness in 36/1000 recreators (recommendation 1) or in 32/1000 recreators (recommendation 2). Based on the risk decision, there are then geometric mean and statistical threshold values that are used to determine if the water is safe. Currently, the state of North Carolina uses the first recommendation for Enterococci in marine waters, but still uses the 1986 recommendation for fecal coliforms in freshwater systems, despite evidence that *E. coli* and Enterococci are better indicators and have finer precision when predicting risk (EPA, 2012; Resources & Administrative, 2019). For this study, criteria recommendation 1 will be used. Therefore, if the geometric mean of the *E. coli* quantified in the samples exceeds 126 CFU/100mL (i.e., RWQC GM) for a 30-day window or more than 10% of samples exceed the statistical threshold value of 410 CFU/100mL (i.e., RWQC STV) within a 30-day window, the water body would be considered impaired.

Criteria Elements	Recommendation 1 Estimated Illness Rate 36/1000		Recommendation 2 Estimated Illness Rate 32/1000	
Indicator	GM (cfu/100 mL)	STV (cfu/100 mL)	GM (cfu/100 mL)	STV (cfu/100 mL)
Enterococci (marine and fresh)	35	130	30	110
E. coli (fresh)	126	410	100	320

Table 1: EPA Recreational Water Quality Criteria [1].

4.0 Methods

4.1 Sampling Scheme

To answer our research questions, collaborators from the Coharie Tribe, North Carolina State University and Duke University constructed a sampling plan to include three sites (see Figure 1.) in the watershed with combinations of low frequency and high frequency temporal sampling schemes capturing both baseflow and stormflow conditions evenly distributed between the dry and wet seasons. Low frequency samplings (n = 72) were decided to be conducted fortnightly for a year with high frequency samplings to be conducted every other hour for 24 hours following a minimum of three days with no precipitation in the dry and wet season (baseflow n = 75) and for 48 hours following a substantial rainfall event in the dry and the wet season (stormflow n = 144). Each sample was tested for total coliforms, *Escherichia coli*, Total suspended solids, Total Kjeldahl Nitrogen, the sum of Nitrate and Nitrite, Ammonia, Total Phosphorus, Orthophosphate, Temperature, pH, dissolved oxygen, specific conductivity, and turbidity. Samples were collected in sterilized 1L polypropylene bottles using an extendable sampling pole. For every 24-hour sampling period field blanks were taken to the sites in the cooler and used as a control to test for contamination. Quantification of total coliforms and *Escherichia coli* was conducted using the IDEXX Colilert-18 assay. To ensure that the field blank had no background contamination we also processed a lab blank with distilled water for IDEXX only. Water quality parameters for temperature, pH, dissolved oxygen, specific conductivity, and turbidity were measured using a ProDSS Multiparameter Digital Water Quality Meter. 80 mL of sample was vacuum filtered through 0.4 µm Isopore filters and frozen at -

Figure 1: a map of sampling locations, drainage basins, the Coharie Tribal Center, a permanent water monitoring stage and a NC State Climate Office run weather station.

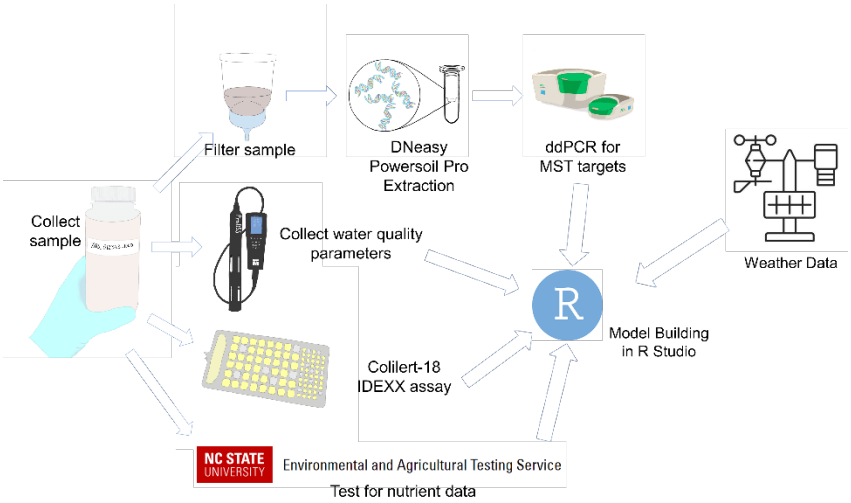

Figure 2: a visualization of the sample processing workflow.

80 Celsius in Qiagen PowerBead Pro tubes for long term storage prior to DNA extraction. An aliquot of water sample was also taken to the Environmental and Agricultural Testing Service laboratory for quantification of total suspended solids, total Kjeldahl nitrogen, the sum of nitrate and nitrite, ammonia, total phosphorus, and orthophosphate. Following all sampling events, filters were thawed on ice and extracted using Qiagen's DNeasy Powersoil Pro DNA extraction kit with the Qiacube HT. Samples were spiked with 10 µL of an internal positive control to measure extraction recovery efficiency. Following extraction, samples were tested for microbial source tracking targets for feces specific to swine and poultry (LA35), swine-only (pig2bac), and human-only (hCYTB484).

4.2 MST Assay Validation

Microbial source tracking is a methodology that has been employed to help researchers pinpoint the source of fecal pollution in the environment using molecular techniques (Boehm et al., 2013; Harwood et al., 2011; Pepper et al., 2015; Weidhaas & Lipscomb, 2013; Zhu et al., 2020). The theory behind the methodology is that specific microbes are associated with a specific host species' gut microbiota (Harwood et al., 2011). By targeting these specific microbes with molecular analysis, conclusions can be made about the sources of fecal pollution to a water system and the possible risks (Harwood et al., 2011; Soller et al., 2010). For this study, the concern is with industrialized agriculture (poultry and swine) and septic (human) contributions to the riverine system.

Feces (n=139) from cattle (12), deer (12), dogs (10), ducks (10), goats (11), geese (10), horses (10), humans (11 samples in the form of wastewater), poultry (39), sheep (15), swine (24) and a bird were collected and extracted with the DNeasy Powersoil Pro kit. These extracts were then used to test the sensitivity and specificity of animal (LA35 and Pig2Bac) and human (hCYTB484) specific targets for fecal contamination MST assays.

The LA35 MST target is present on the *Brevibacterium sp.*16S rRNA gene (Weidhaas & Lipscomb, 2013). This assay was first developed as a qPCR assay in West Virginia to target poultry and had a sensitivity of 76% and was 100% specific (Weidhaas & Lipscomb, 2013). To make this assay more sensitive, we have adapted the LA35 assay to ddPCR and validated it for our study area.

Pig2bac is an assay that was originally validated as a qPCR assay that targets the 16S rRNA gene in swine-specific *Bacteroidales* in France (Mieszkin et al., 2009). Using the same assay as Mieszkin et al., we also validated it with the fecal specimens collected in Eastern North Carolina. Lastly, the hCYTB484 assay was also validated. Having been originally developed on the qPCR platform targeting the cytochrome b gene in human mitochondrial DNA, Zhu et al., adapted the assay for use on the ddPCR platform (Caldwell et al., 2007; 2020). This target is specific to humans and has been used for source tracking in the environment (Zhu et al., 2020). The authors also performed a comparative analysis between their newly developed mitochondrial DNA assay and a different commonly used MST target for humans, HF183 (Zhu et al., 2020). The mitochondrial assay was 97% specific compared to 80% for HF-183. Cross reactivity for the mitochondrial assay occurred in samples from a cow and swine. Sensitivity of the assay was found to be 100% for the mitochondrial assay, but just 51% for HF-183 (Zhu et al., 2020).

Sensitivity was calculated using the following equation: $\frac{True\ Positives}{True\ Positives+False\ Negatives}$

Specificity was calculated using the following equation: $\frac{True\ Negatives}{True\ Negatives+False\ Positives}$

4.5 Multiple Linear Regression Model Construction

Log-transformed *E. coli* concentrations were modeled using multiple linear regressions for each site and cumulatively to determine which covariates explain the variability of *E. coli* concentration in the stream system. R studio was used to conduct this analysis with the randomForest R package (Liaw & Wiener, 2002; R Core Team, 2021). Covariates included water quality parameters, nutrient measurements, detect/non-detect of MST target concentrations, and cumulative rainfall amounts for the 48 hours prior to the sample grab (R Core Team, 2021). For the cumulative model, site was controlled for as a covariate. Rainfall data was provided by the NC State Climate Office. There are 97 observations at each site for a total of 291 observations. Covariates were removed manually to only be included in the final models if the p-value was less than 0.2 and had a variation inflation factor less than 5.

4.6 Random Forest Predictive Model Construction

E. coli concentrations were transformed into a binary variable as either "safe" (i.e., below the EPA standard threshold 126 MPN/100 mL) or "unsafe" (i.e., above 126 MPN/100). R studio was used to conduct this analysis with the randomForest R package (Liaw & Wiener, 2002; R Core Team, 2021)The predictive model was performed first using all available parameters including two-day rainfall, specific conductivity, the sum of nitrate and nitrite, site, pH, turbidity, dissolved oxygen, total suspended solids, temperature, total phosphorus, total Kjeldahl nitrogen, orthophosphate, ammonia and detect/non-detect data from the MST assays. The predictive model was also performed using strictly data that can be obtained in real-time including two-day rainfall, specific conductivity, site, pH, turbidity, dissolved oxygen, and temperature. The data was split into 80% training and 20% testing. The number of variables randomly sampled as candidates at each split was set to 4 for the model with all data variables and 3 for the real-time model (mtry = sqrt(variables). The number of trees to grow was set to 2001 due to marginal performance improvements in out-of-bag error beyond this forest size and an odd number was chosen in case the forest was in need of a tie breaker (Liaw & Wiener, 2002). Sensitivity and specificity equations were already described in the MST validation section and are consistent to the random forest modeling. Variable importance using mean decrease in accuracy were calculated for each model as described by the randomForest R package (Liaw & Wiener, 2002). The following equation was used to calculate misclassification (using the training dataset) and out-of-bag error (using the testing data set):

$$\frac{False\ Positive+False\ Negative}{True\ Positive+True\ Negative+False\ Positive+False\ Negative}.$$

4.7 Generalized estimating equation construction

Log-transformed *E. coli* concentrations were also predicted using generalized estimating equations with the R package geepack (Højsgaard et al., 2006). The model employed all water quality data collected and detect/non-detect data from microbial source tracking targets. This method was used to account for clustering at sample site (i.e., needed in our study due to repeat sampling at a site) using robust standard errors. Parameter estimates, standard error, Wald statistics and p-values were all calculated as described in the geepack R package (Højsgaard et al., 2006).

5.0 Results

5.1 MST Validation Results

The performance of the LA35 assay for detecting and distinguishing poultry feces was worse in our study sample compared to the original assay used in West Virginia. We found that the assay was 66.7% sensitive and 79.4% specific. However, when targeting poultry and swine together the performance of the assay improved to 79.4% sensitive and 98.0% specific; making the LA35 assay a good option as an MST target for industrialized agriculture in North Carolina. We found the Pig2bac assay to be 100% sensitive and 93% specific. The 8 non-swine fecal samples that cross reacted were from horse, sheep, human and duck. Lastly, hCYTB484 was found to be 100% sensitive and 92% specific to human feces, similar to the results found by Zhu et al. (2020). The 11 non-human fecal samples that cross reacted belonged to goats, poultry, geese, deer, dogs, and sheep.

5.1 Temporal *E. coli* data

Shown in Figure 3 are *E. coli* results from fortnightly sampling events throughout the sampling period. The geometric mean for Sevenmile Swamp, Marsh Swamp and Five Bridges Road are 120, 388, and 140 MPN/100 mL respectively. Sevenmile Swamp, Marsh Swamp and Five Bridges Road exceed the statistical threshold value 21.4%, 46.4% and 17.9% of the time respectively, indicating that the watershed frequently does not meet EPA recreational water quality standards.

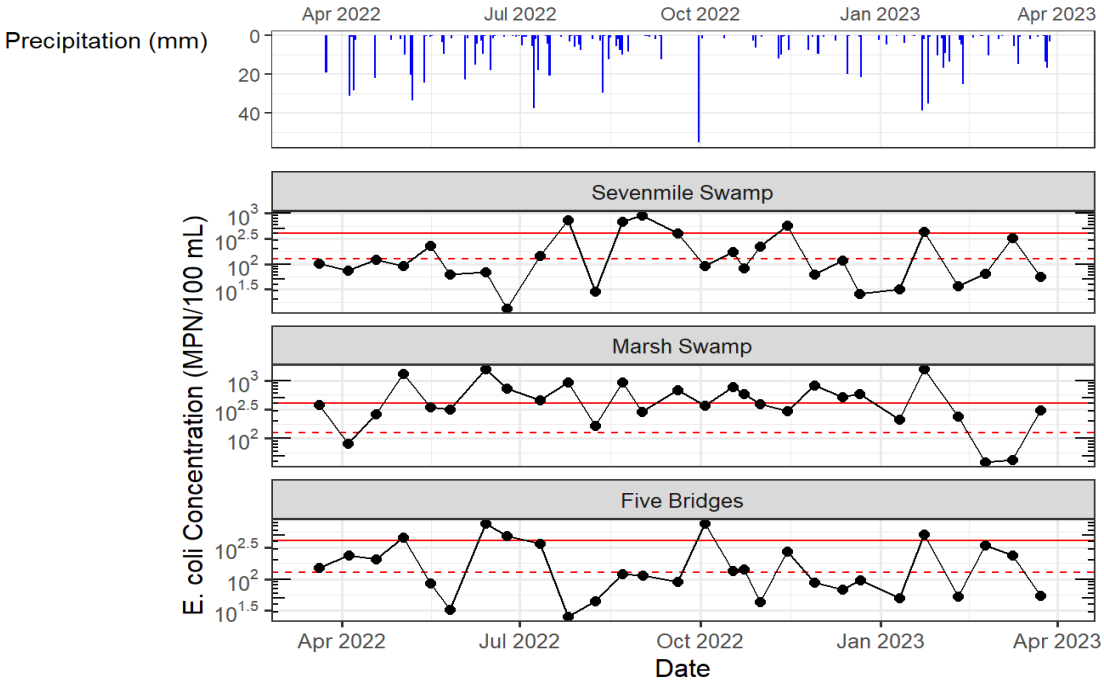

Figure 3: fortnightly samplings at the three sites from March 2022 to March 2023. Take notice to the Geometric Mean Threshold (dotted red line) and Statistical Threshold Value lines (solid red) which are indicative of RWQC recommendation 1 by the EPA. Daily precipitation is also plotted along the same x-axis along the top of the figure in millimeters.

Figure 4 illustrates the *E. coli* results from a high frequency sampling event (n=13 at each site) conducted from 20 March 2022 to 21 March 2022. The geometric mean for Sevenmile Swamp, Marsh Swamp and Five Bridges Road are 72, 278, and 92 MPN/100 mL respectively. Marsh Swamp is also seen to have exceeded the statistical threshold value for 15% of samples collected during this sampling campaign.

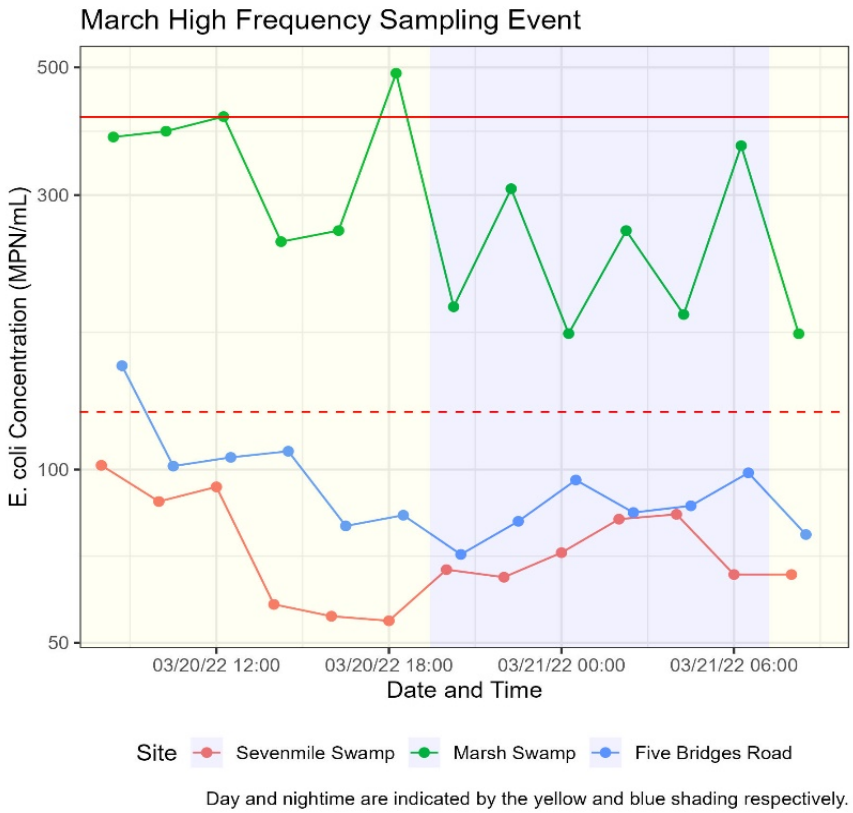

Figure 4: Summary of E. coli concentration over time during a high frequency sampling in March 2022. Take notice to the Geometric Mean Threshold (dotted red line) and Statistical Threshold Value lines (solid red) which are indicative of RWQC recommendation 1 by the EPA.

A storm flow sampling was conducted on 11 July 2022 after the stream stage located at the 3 Bridges site in the watershed was reported to have increased by 17 centimeters in a 24-hour window and approximately 1 centimeter of rainfall was recorded at the weather station. *E. coli* levels were above the RWQC geometric mean (126 MPN/100mL) for all but two samples across all sites (Figure 5). Sevenmile Swamp and Marsh Swamp saw all 24 of their samples for the 48-hour sampling period above the geometric mean with Marsh Swamp exceeding the statistical threshold value (410 MPN/100 mL) 45.8% of the time. Five Bridges Road had no samples

exceeding the statistical threshold value, but 22 out of the 24 samples were above the geometric mean value, indicating that all three sites were likely unsafe for recreation during the sampling.

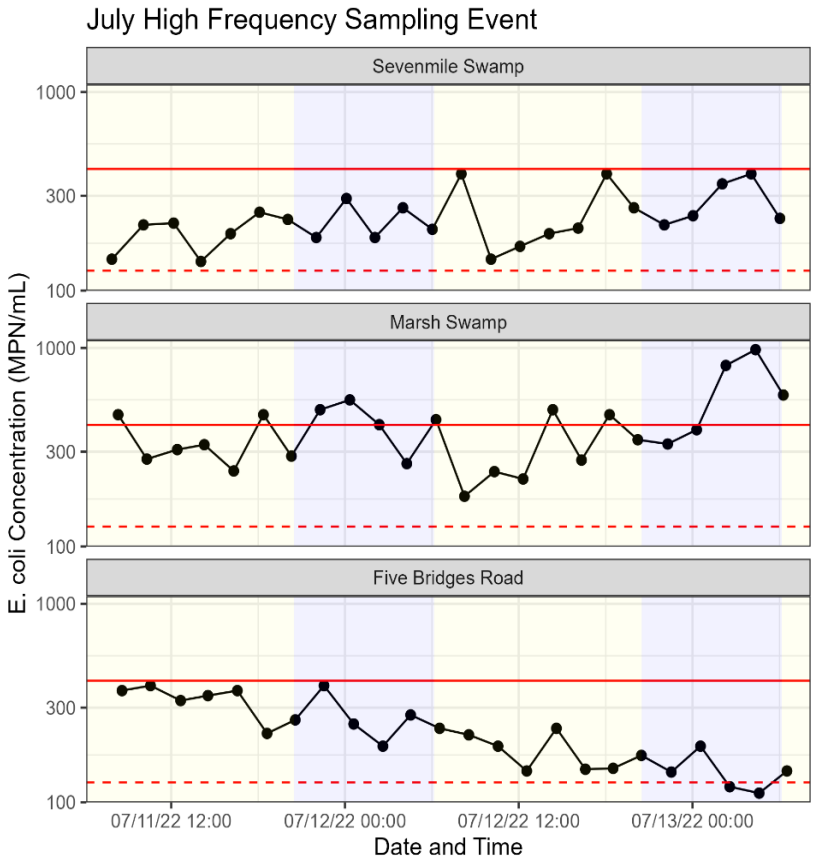

Figure 5: Summary of E. coli concentration over time during a high frequency sampling in July 2022. Take notice to the Geometric Mean Threshold (dotted red line) and Statistical Threshold Value lines (solid red) which are indicative of RWQC recommendation 1 by the EPA.

On 23 October 2022 a 24-hour high frequency baseline sampling was conducted. Results from the sampling are visualized in Figure 6. The geometric mean for Sevenmile Swamp, Marsh Swamp and Five Bridges Road are 63.8, 286, and 102 MPN/100 mL respectively. Marsh Swamp is also seen to have exceeded the statistical threshold value for 33% of samplings. These summary statistics align with what was observed in the March sampling earlier that year.

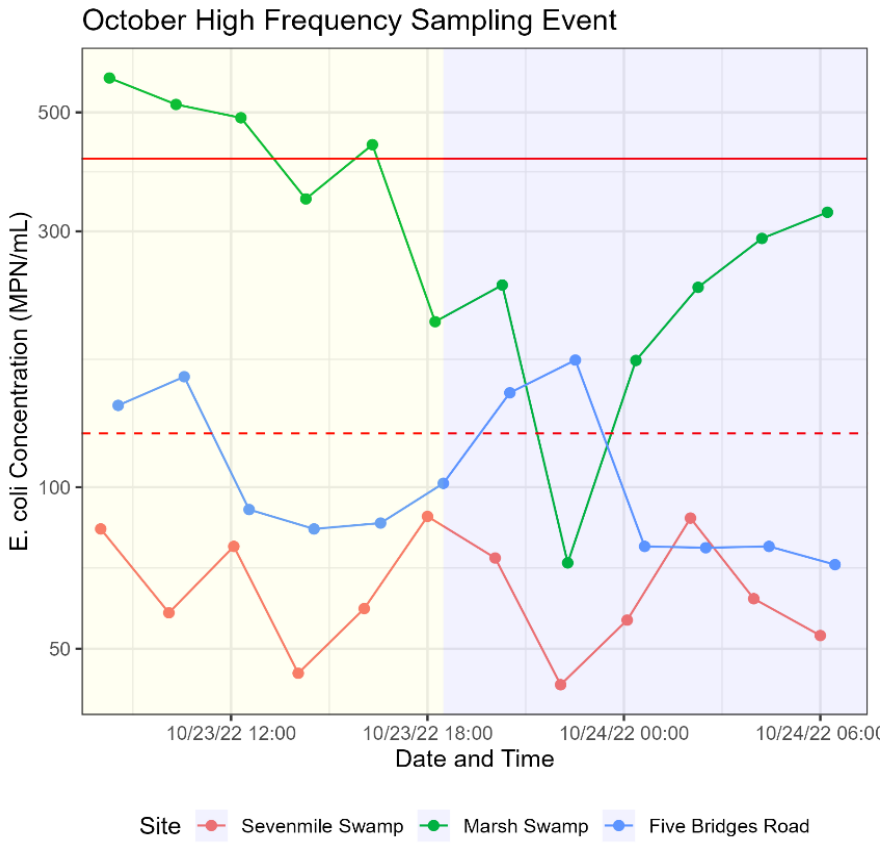

Figure 6: Summary of E. coli concentration over time during a high frequency sampling in October 2022. Take notice to the Geometric Mean Threshold (dotted red line) and Statistical Threshold Value lines (solid red) which are indicative of RWQC recommendation 1 by the EPA.

From January 22 to January 23, 40.1 mm of rainfall precipitation were logged at the weather station adjacent to the watershed and a sampling was started on 23 January 2023. The geometric mean of *E.coli* concentrations from samples collected at Sevenmile Swamp, Marsh Swamp and Five Bridges Road were 143, 336, and 296 MPN/100 mL respectively. Sevenmile Swamp, Marsh Swamp and Five Bridges Road samples exceed the statistical threshold value 8.33%, 45.8% and 25% of the time respectively. Data are visualized in Figure 7.

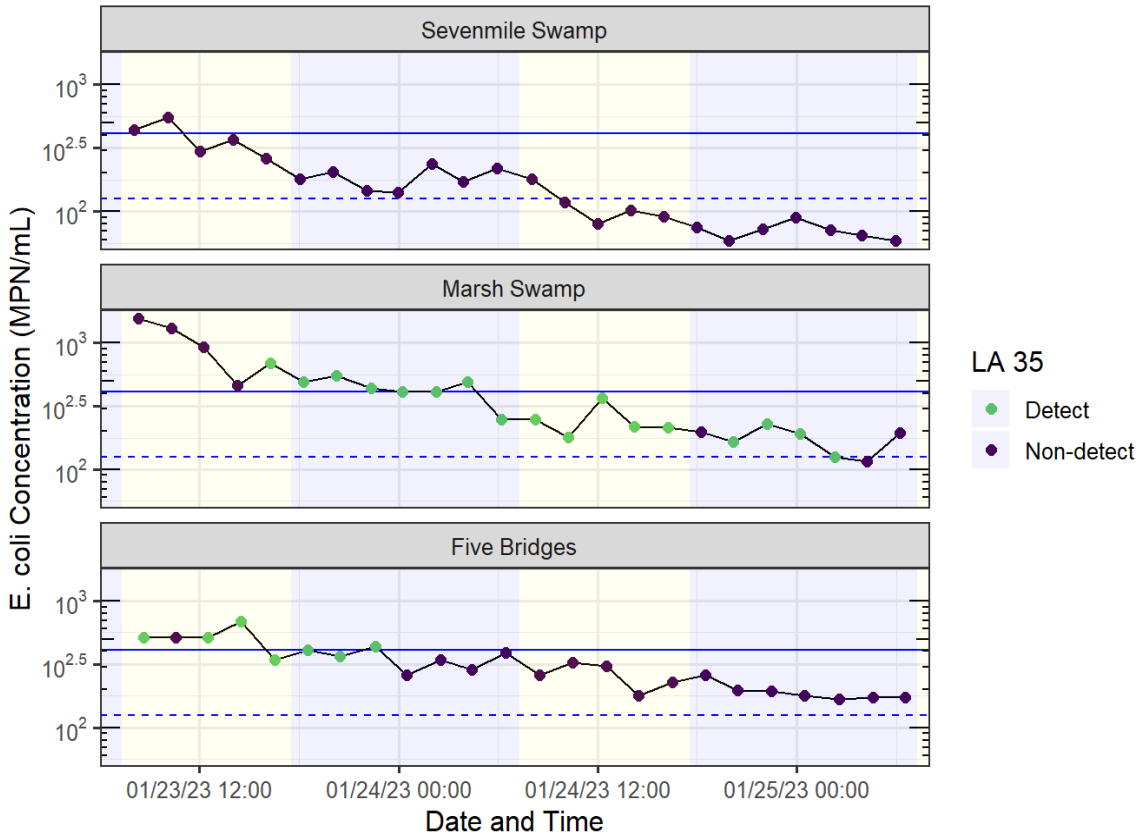

Figure 7: Summary of E. coli concentration over time during a high frequency sampling in January 2023. Take notice to the Geometric Mean Threshold (dotted blue line) and Statistical Threshold Value lines (solid blue) which are indicative of RWQC recommendation 1 by the EPA. This figure also shows samples that detected LA35 in green.

5.2 Combined swine and poultry fecal detection with LA35 MST assay

LA35 was detected in 28 of 291 samples collected over the year. 19 of the positive samples were from Marsh Swamp, 8 were from Five Bridges and 1 was from Sevenmile swamp. A vast majority of the positive samples were collected during the January 2023 high frequency sampling event which spanned January 23rd, 24th and 25th. Other dates that LA35 was detected include July 25th, 2022, at Marsh Swamp, October 23rd, 2022, at both Five Bridges and Sevenmile Swamp, and October 24th, 2022, at Marsh Swamp. The concentrations of LA35 in the samples range from 1008 to 5323 copies per liter of river water. The average concentration of LA35 in positive samples was 2426 copies per liter of river water. Log transformed *E. coli* concentrations were found to be statistically significantly higher when LA35 was detected versus when LA35 was not detected (Welch Two Sample t-test p-value = 0.0003). At Sevenmile Swamp LA35 was detected during 4% of baseflow samples, 0% of low frequency samples and 0% of storm flow samples. At Marsh Swamp LA35 was detected during 4% of baseflow samples, 4% of low frequency samples and 35% of storm flow samples. At Five Bridges LA35 was detected during 4% of baseflow samples, 0% of low frequency samples and 15% of storm flow samples.

5.3 Swine specific fecal detection with Pig2bac MST assay

Pig2bac was detected in 8% of samples collected from Sevenmile Swamp, 32% of the samples collected at Marsh Swamp and in 23% of samples at Five Bridges Road. Much like the detection of LA35, a majority of the detections occurred on samples that were collected during the high frequency sampling events. Of the 28 samples that were positive for LA35, 15 of them were also positive for pig2bac, this suggests that poultry may have been the source of fecal contamination for the samples that were positive for LA35 and not for pig2bac. Detection of Pig2Bac was not associated with higher concentrations of *E. coli*. At Sevenmile Swamp Pig2bac was detected during 12% of baseflow samples, 13% of low frequency samples and 4% of storm flow samples. At Marsh Swamp Pig2bac was detected during 52% of baseflow samples, 17% of low frequency samples and 38% of storm flow samples. At Five Bridges Pig2bac was detected during 44% of baseflow samples, 8% of low frequency samples and 23% of storm flow samples.

5.4 Human specific fecal detection with hCYT484B MST assay

The human target was detected in 44% of samples from Sevenmile Swamp, 43% of the samples at Marsh Swamp, and 41% of the samples at Five Bridges Road. Detections of hCYTB484 was not associated with higher concentrations of log-transformed *E. coli* (Welch Two Sample t-test p-value = 0.8). Suspected human fecal contamination was detected more frequently during high frequency sampling events than during low frequency sampling events. At Sevenmile Swamp hCYTB484 was detected in 36% of baseflow samples, 79% of low frequency samples and 31% of storm flow samples. At Marsh Swamp hCYTB484 was detected in 40% of baseflow samples, 46% of low frequency samples and 48% of storm flow samples. At Five Bridges hCYTB484 was detected in 40% of baseflow samples, 54% of low frequency samples and 40% of storm flow samples.

5.5 Quality assurance and control

All quality control measures including field blanks, extraction blanks, and no template controls, to check for contamination from laboratory technicians passed.

5.5 Explanatory Model

Results are summarized in Table 2 and Figure 8. Five Bridges Road had the model with the most variability in *E.coli* concentrations explained by the co-variates, with an adjusted R^2 value of 0.5053. All models identified 2-day rain to be positively associated with *E. coli* concentrations ($p<0.05$), which indicates that rainfall events result in more fecal contamination in the watershed. Only Sevenmile Swamp and the cumulative models had concentrations of targets for microbial source tracking as statistically significant covariates. The cumulative model had a positive relationship between log transformed *E. coli* and LA35 detections across all sites.

Model	Adjusted r^2	Covariates
Sevenmile Swamp	0.47	pH (-), specific conductivity (-), TSS (+), Nitrate+Nitrite (+), 2-day rain (+)
Marsh Swamp	0.28	2-day rain (+), TKN (-), dissolved oxygen (-), temperature (+)
Five Bridges Road	0.51	pH (-), specific conductivity (+), TSS (+), 2-day rain (+), dissolved oxygen (+)
Cumulative	0.47	pH (-), 2-day rain (+), TKN (-), temperature (+), site (+), LA35 Detect (+)

Table 2: summary of preliminary exploratory models.

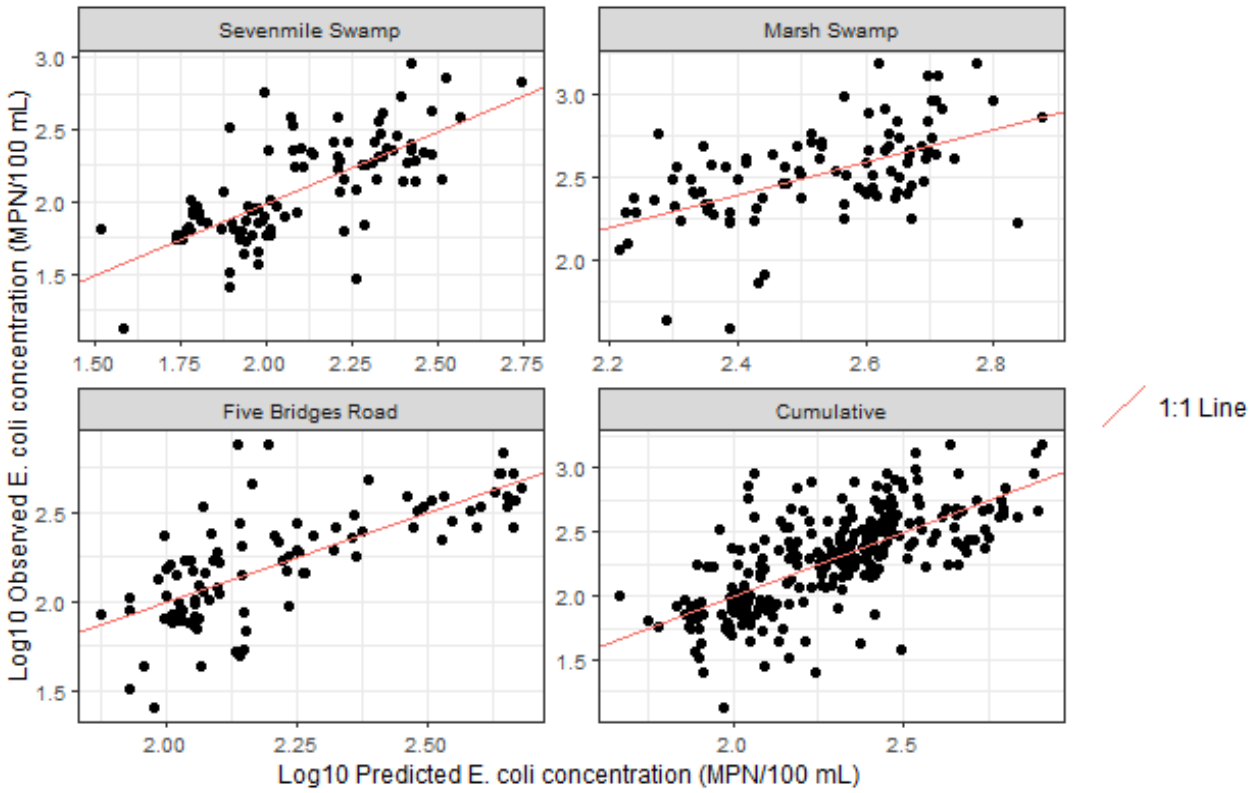

Figure 8: Preliminary explanatory models. Red line is a 1:1 line to help indicate goodness of fit.

5.6 Random Forest Predictive Model Results – Full data set

The random forest predictive model achieved an out of bag error of 9.44% during training, meaning that when the test data set was used in the predictive model, the model output misclassified 9.44% of the time. The testing data set yielded a sensitivity of 83.3% and a specificity of 87.5%. The misclassification rate for the model was 0% during training. Covariate importance analysis was performed (Figure 9) which determined that 2-day rain was the most important covariate in the predictive model and that inclusion of MST detect/non-detect data was least impactful to the model performance when predicting water safety level.

5.7 Random forest predictive model – real-time result

When utilizing only variables that can be available in real time, the random forest categorical model had an out-of-bag error of 12%. The testing data show a similar performance to the other random forest model that included all covariates, achieving a sensitivity of 83.3% and specificity of 87.5%. The misclassification rate for the model was 0% during training. Covariate importance was determined, and two-day rainfall, site, and specific conductivity were the three most important variables in the model.

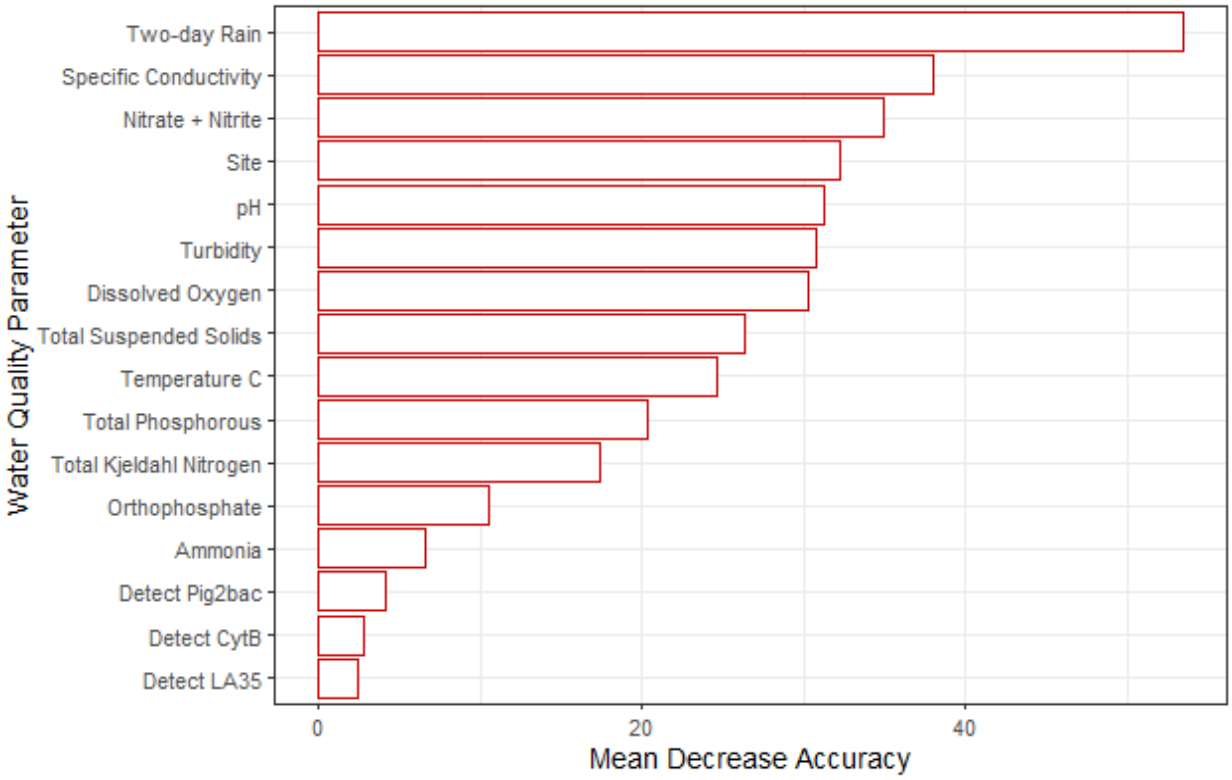

Figure 9: covariate importance analysis of the random forest predictive model including all covariates. When "Mean Decrease Accuracy" is high that means the model performs worse without that respective covariate.

5.8 Generalized Estimating Equation

The generalized estimating equation model yielded seven covariates that displayed statistical significance (p-value < 0.05) for estimating log transformed *E. coli* concentrations. Of particular note, two-day rainfall and water temperature were statistically significant co-variates in the GEE model.

Table 1: GEE model results including variable coefficients, standard errors, test statistic and p-values. Statistically significant covariates (p-value < 0.05) are bolded.

Variable	Parameter Estimates	Standard Error	Wald Test Statistic	p-value
(Intercept)	**1.873**	**0.177**	**111.969**	**< 0.001**
pH	**-0.087**	**0.042**	**4.373**	**0.037**
Specific Conductivity	**0.004**	**0.000**	**260.169**	**< 0.001**
TSS	0.007	0.007	1.066	0.302
Nitrate + Nitrite	**0.056**	**0.025**	**4.905**	**0.027**
2-day rain	**0.012**	**0.001**	**394.947**	**< 0.000**

TKN	-0.135	0.088	2.363	0.124
Ammonia	0.022	0.103	0.047	0.829
TP	0.765	0.887	0.743	0.389
DO %	0.001	0.002	0.112	0.738
Water Temperature C	**0.017**	**0.005**	**10.348**	**0.001**
LA35 detect	**0.044**	**0.014**	**9.995**	**0.002**
Pig2bac detect	**-0.099**	**0.038**	**6.785**	**0.009**
hCYTB484 detect	0.052	0.029	3.273	0.070

6.0 Discussion

Results from temporal water-quality sampling indicate that the Great Coharie watershed experiences frequently elevated levels of *E. coli* relative to EPA recommendations on recreational water quality. This finding indicates an increased risk of gastrointestinal illness associated with recreational use of these streams, but to what extent is still unknown. A study conducted by Soller et al. shows that the risk of gastrointestinal illness when exposed to water contaminated with pig, gull or chicken feces is less than that of waters contaminated with human or cow feces (2010). The addition of the MST markers has indicated that the watershed likely does experience swine, poultry, and human fecal contamination. While detections of MST targets were not found in all of the samples taken, it is important to note that the sensitivity of the assays may result in the presence of fecal contamination not being detected. Further, the source tracking targets may also decay at different rates than pathogens from feces as well as indicator *E. coli* (Dean & Mitchell, 2022; Oladeinde et al., 2014).

The explanatory models all had relatively good performance for predicting *E. coli* concentrations and could still benefit from the finetuning of covariates. Flow data in other studies has been shown to be an important covariate in modeling fecal indicator bacteria (Kay et al., 2007; Muirhead et al., 2004; Panidhapu et al., 2020). One issue that arose when constructing the explanatory model was the validity of the precipitation data. For the high frequency sampling event conducted in July of 2022 there was evidence of a rainfall event from a sharp incline on the stream stage and from personal accounts of collaborators who live in the watershed, but the weather station used for model construction did not follow those observations. Future modeling efforts would benefit from increased accuracy of rainfall in higher spatial resolution. Nonetheless, the models indicate that 2-day rainfall is an important covariate for explaining *E. coli* concentration in this system, which is in agreement with other studies that found rainfall to be an important covariate when predicting fecal indicator bacteria in surface waters (Guber et al., 2006; Tolouei et al., 2019; Vermeulen & Hofstra, 2014).

Sampson County has a large agricultural presence with substantial operations to support hog and poultry production. Prior work has identified these operations as potential sources of fecal contamination in the watershed (Guidry et al., 2018; Harden, 2015). North Carolina DEQ has

developed a general permit system that aims to minimize the amount of hog waste that is released into the environment. One of the measures that are outlined in the document include the design of the hog lagoon to be able to withstand a 25-year, 24-hour rainfall event, which happens to be 18.41 centimeters (90% confidence interval 16.15-21.03) in Clinton, NC (NOAA, 2017). There are also specified times for field applying (i.e. spraying should cease within 12 hours of the National Weather Service issuing a watch or warning), specified locations for spraying (i.e. not within 200 feet of a dwelling that is not owned by the permit holder), and not spraying during windy conditions that could cause waste to cross property lines or contaminate surface waters and wetlands (NCDEQ, 2019). The storms that were monitored in the present study were well below the 25-year, 24-hour rainfall amounts, but contamination was still present in the water samples collected. The results presented in this study indicate that the current regulations are not preventing contamination from entering the watershed, and further work to understand the risks posed by this contamination would be helpful for making of policy recommendations.

Policies that regulate the timing of field spraying related to large rainfall events focus on pollution that would result from overland transport, but it is documented that shallow groundwater contributions to surface water are sources of surface water contamination (Oldfield et al., 2020; Rozemeijer & Broers, 2007; Spoelstra et al., 2020). The consistent spraying of fields could lead to the accumulation of swine related pollutants in the shallow groundwater table that feeds these tributaries. Increased rainfall would increase the groundwater contributions to the rivers and therefore increase pollutant loadings. We recommend future studies to explore the accumulation, fate and transport of microbial source tracking targets in shallow groundwater tables that are impacted by lagoon spraying. Should evidence of groundwater transmission of these swine pollutants be found, then policies surrounding lagoon spraying could be changed to limit the total loading of waste per acre of farmland, to protect groundwater which eventually contributes to surface water.

Our study also found frequent detection of human fecal matter in the water samples. Prior work in other rural areas have also detected evidence of human sources of contamination. A study in Canada found that 91% of rural streams sampled in Ontario tested positive for artificial sweeteners, which is another indicator of septic contributions to surface water (Spoelstra et al., 2020). This same study also estimated that 13% of all septic effluent ended up in surface waters via groundwater transport. Human contamination was only detected at a higher rate during the storm flow sampling events at Marsh Swamp when compared to baseflow and low frequency samplings, which is consistent with the findings from another Canadian study that found increased concentrations of artificial sweeteners being contributed to surface waters during wet weather periods that contributed to higher discharges, but this phenomenon was not observed at the Sevenmile Swamp or Five Bridges sites (Oldfield et al., 2020). Another study performed in the same region of North Carolina used HF-183 as a human-associated fecal target and detected human-associated impacts in surface waters in 34% of samples collected downstream from commercialized hog operations and 20% of the time in background sites using a ddPCR platform (Christenson et al., 2022). This same study found that mean HF-183 concentrations were similar between background sites and downstream of commercialized hog operations (Christenson et al., 2022). It is important to note that the sensitivity and specificity of the HF-183 target was worse compared to the hCYTB484 target that was validated in our current study (Christenson et al., 2022). This could help explain our results between all three sites that seem to experience variable

detections of hog associated contamination, but more consistent detections of human associated contamination.

The modeling tools selected for this study were based on previous publications. Many linear regression model types have been used for identifying different environmental factors that are associated with *E. coli* and other pollutants in different locations (Francy & Darner, 2007; Guber et al., 2006; Nevers & Whitman, 2005; Sinha & Michalak, 2016). One attempt to nowcast water safety used turbidity, wave height, 2-day rainfall, satellite rainfall data and day of year to achieve an r^2 value of 0.43, which is comparable to the 0.46 r^2 value that was achieved in the cumulative model of our study (Francy & Darner, 2007). Another linear regression was developed for data collected in Lake Michigan and achieved r^2 values of 0.635, 0.32 and 0.465 at three sites using a different combination of wave height, wave period, turbidity (sometimes at two sites), chlorophyll, and 4-hour rainfall (Nevers & Whitman, 2005).

Our use of a random forest model to predict exceedances to regulatory thresholds is a previously used concept as well. A prior study employed an assortment of random forest models to predict exceedances in water quality thresholds at 7 sites in Lake Superior and Lake Michigan (Brooks et al., 2016). Their highest performing model across all sites achieved 80% specificity, 51% sensitivity, and made the correct decision 76% of the time. This model performed worse than our model likely due to their data inputs, which included limited water quality parameters like turbidity, water temperature, and sometimes specific conductance. This study also used some data that was publicly available, which is something that could potentially improve the performance of our model, increasing our model training data set by using data collected over multiple years at Five Bridges Road.

7.0 Conclusion

Our study found that the Great Coharie river frequently experiences elevated levels of *E. coli* and likely also experiences fecal contamination from industrialized agriculture and human waste in the area. To help the Coharie people make risk-based decisions on when to recreate on the river, predictive models show great potential. Using random forest machine learning approaches, unsafe water conditions can be predicted with just 12.9% out-of-bag error using parameters collected solely by immediately available data via a multiparameter sonde and a weather station that can be set up to report data measurements real-time.

Microbial source tracking data reveals that the Great Coharie River watershed likely experiences fecal contamination from swine and humans on a regular basis. The human targets were detected across the watershed at low concentrations showing that human development in the watershed warrants further investigation to improve management of human waste. This may require investments in stormwater infrastructure (e.g., such as nature-based solutions like riparian buffers and wetlands) and septic system maintenance and performance evaluation.

Further, our study found that prior rainfall results in elevated levels of fecal contamination, even with rainfall events that are less than that of a 25-year, 24-hour storm. In all the models constructed in this study, rainfall is indicated to be the most important variable to explaining *E. coli* variability in a positive relationship. These findings also demonstrate the potential for

subsurface contributions of lagoon and septic waste to surface waters driven by higher groundwater flows due to rainfall. Evidence of swine, poultry, and human impacted waters raise concerns surrounding the possibility for human exposure to human relevant pathogens. Contamination sources to monitor in the future include septic, swine and poultry feces that have been sprayed into the air or added to the river system due to policy failures that have led to subsurface inputs or contributions to this important tributary.

Future work can expand upon the findings of this study. To further understanding the spatial variability of contamination in the watershed, additional sample sites that are culturally relevant to the Tribe could be selected. Analyzing relationships between contamination across the watershed and landuse could help pinpoint the dominant sources of contamination. Additionally, a groundwater study investigating the fate and transport of MST targets in groundwater on sprayed lands adjacent to the watershed could shed further light on the transport of these microorganisms into the Great Coharie River. The same groundwater study could also investigate septic contributions to the River using the hCYTB484 assay paired with other human markers like acesulfame, which could improve the accuracy of source tracking. The overall groundwater study could assist NCDEQ in the modification of policy related to lagoon spraying practices and shed light on septic system performance in the area.

8.0 Recommendations

Based on the current data it is recommended that more frequent monitoring be undertaken at the Five Bridges Road site due to the recreational importance. We also recommend that a land use study be undertaken to determine potential management strategies to minimize fecal contamination at the Marsh Swamp site. For the time being, recreational swimming is highly discouraged at Marsh Swamp and increased caution should be taken when considering swimming at Sevenmile Swamp and Five Bridges Road. Further optimization and validation of predictive tools will help inform people of the best times to recreate on the river.

Works Cited

Boehm, A. B., Van De Werfhorst, L. C., Griffith, J. F., Holden, P. A., Jay, J. A., Shanks, O. C., Wang, D., & Weisberg, S. B. (2013). Performance of forty-one microbial source tracking methods: A twenty-seven lab evaluation study. *Water Research*, *47*(18), 6812–6828. https://doi.org/10.1016/j.watres.2012.12.046

Brooks, W., Corsi, S., Fienen, M., & Carvin, R. (2016). Predicting recreational water quality advisories: A comparison of statistical methods. *Environmental Modelling & Software*, *76*, 81–94. https://doi.org/10.1016/j.envsoft.2015.10.012

Caldwell, J. M., Raley, M. E., & Levine, J. F. (2007). Mitochondrial Multiplex Real-Time PCR as a Source Tracking Method in Fecal-Contaminated Effluents. *Environmental Science & Technology*, *41*(9), 3277–3283. https://doi.org/10.1021/es062912s

Christenson, E., Wickersham, L., Jacob, M., & Stewart, J. (2022). A watershed study assessing effects of commercial hog operations on microbial water quality in North Carolina, USA. *Science of The Total Environment*, *838*, 156085. https://doi.org/10.1016/j.scitotenv.2022.156085

Dean, K., & Mitchell, J. (2022). Identifying water quality and environmental factors that influence indicator and pathogen decay in natural surface waters. *Water Research*, *211*, 118051. https://doi.org/10.1016/j.watres.2022.118051

Emanuel, R. E. (2024). *On the Swamp: Fighting for Indigenous Environmental Justice*. University of North Carolina Press.

EPA. (2012). *2012 Recreational Water Quality Criteria*.

Francy, D. S., & Darner, R. A. (2007). *Nowcasting Beach Advisories at Ohio Lake Erie Beaches* (Open-File Report) [Open-File Report]. USGS. https://pubs.usgs.gov/of/2007/1427/pdf/ofr2007-1427.pdf

Guber, A. K., Shelton, D. R., Pachepsky, Y. A., Sadeghi, A. M., & Sikora, L. J. (2006). Rainfall-Induced Release of Fecal Coliforms and Other Manure Constituents: Comparison and Modeling. *Applied and Environmental Microbiology*, *72*(12), 7531–7539. https://doi.org/10.1128/AEM.01121-06

Guidry, V. T., Rhodes, S. M., Woods, C. G., Hall, D. J., & Rinsky, J. L. (2018). Connecting Environmental Justice and Community Health: Effects of Hog Production in North Carolina. *North Carolina Medical Journal*, *79*(5), 324–328. https://doi.org/10.18043/ncm.79.5.324

Harden, S. L. (2015). *Surface-Water Quality in Agricultural Watersheds of the North Carolina Coastal Plain Associated with Concentrated Animal Feeding Operations* (Scientific Investigations Report) [Scientific Investigations Report]. NCDEQ. https://pubs.usgs.gov/sir/2015/5080/pdf/sir2015-5080.pdf

Harris, A. R., Fidan, E. N., Nelson, N. G., Emanuel, R. E., Jass, T., Kathariou, S., Niedermeyer, J., Sharara, M., de los Reyes, F. L. I., Riveros-Iregui, D. A., & Stewart, J. R. (2021). Microbial Contamination in Environmental Waters of Rural and Agriculturally-Dominated Landscapes Following Hurricane Florence. *ACS ES&T Water*, *1*(9), 2012–2019. https://doi.org/10.1021/acsestwater.1c00103

Harwood, V. J., Ryu, H., & Santo Domingo, J. (2011). Microbial Source Tracking. In *Fecal Bacteria* (pp. 189–202). American Society for Microbiology (ASM). https://app-knovel-com.prox.lib.ncsu.edu/web/view/khtml/show.v/rcid:kpFB000001/cid:kt0090DEP1/viewerType:khtml//root_slug:9-microbial-source-tracking/url_slug:microbial-source-tracking?cid=kt0090DENH&b-toc-cid=kpFB000001&b-toc-title=Fecal%20Bacteria&b-toc-url-slug=microbial-source-tracking&view=collapsed&zoom=1&page=2&q=microbiota

Højsgaard, S., Halekoh, U., & Yan, J. (2006). The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software*, *15*, 1–11. https://doi.org/10.18637/jss.v015.i02

Kay, D., Aitken, M., Crowther, J., Dickson, I., Edwards, A. C., Francis, C., Hopkins, M., Jeffrey, W., Kay, C., McDonald, A. T., McDonald, D., Stapleton, C. M., Watkins, J., Wilkinson, J., & Wyer, M. D. (2007). Reducing fluxes of faecal indicator compliance parameters to bathing waters from diffuse agricultural sources: The Brighouse Bay study, Scotland. *Environmental Pollution*, *147*(1), 138–149. https://doi.org/10.1016/j.envpol.2006.08.019

Liaw, A., & Wiener, M. (2002). *Classification and Regression by randomForest*. *2*.

Mieszkin, S., Furet, J.-P., Corthier, G., & Gourmelon, M. (2009). Estimation of Pig Fecal Contamination in a River Catchment by Real-Time PCR Using Two Pig-Specific Bacteroidales 16S rRNA Genetic Markers. *Applied and Environmental Microbiology*, *75*(10), 3045–3054. https://doi.org/10.1128/AEM.02343-08

Muirhead, R. W., Davies-Colley, R. J., Donnison, A. M., & Nagels, J. W. (2004). Faecal bacteria yields in artificial flood events: Quantifying in-stream stores. *Water Research*, *38*(5), 1215–1224. https://doi.org/10.1016/j.watres.2003.12.010

NCDEQ. (2019). *Swine Waste Management System General Permit*. https://www.deq.nc.gov/general-permit-swine-2019/open

Nevers, M. B., & Whitman, R. L. (2005). Nowcast modeling of *Escherichia coli* concentrations at multiple urban beaches of southern Lake Michigan. *Water Research*, *39*(20), 5250–5260. https://doi.org/10.1016/j.watres.2005.10.012

NOAA. (2017). *PF Map: Contiguous US*. https://hdsc.nws.noaa.gov/pfds/pfds_map_cont.html?bkmrk=nc

Oladeinde, A., Bohrmann, T., Wong, K., Purucker, S. T., Bradshaw, K., Brown, R., Snyder, B., & Molina, M. (2014). Decay of Fecal Indicator Bacterial Populations and Bovine-Associated Source-Tracking Markers in Freshly Deposited Cow Pats. *Applied and Environmental Microbiology*, *80*(1), 110–118. https://doi.org/10.1128/AEM.02203-13

Oldfield, L. E., Roy, J. W., & Robinson, C. E. (2020). Investigating the use of the artificial sweetener acesulfame to evaluate septic system inputs and their nutrient loads to streams at the watershed scale. *Journal of Hydrology*, *587*, 124918. https://doi.org/10.1016/j.jhydrol.2020.124918

Panidhapu, A., Li, Z., Aliashrafi, A., & Peleato, N. M. (2020). Integration of weather conditions for predicting microbial water quality using Bayesian Belief Networks. *Water Research*, *170*, 115349. https://doi.org/10.1016/j.watres.2019.115349

Pepper, I. L., Gerba, C. P., & Gentry, T. J. (2015). *Environmental microbiology*. Academic Press is an imprint of Elsevier. https://catalog.lib.ncsu.edu/catalog/NCSU3110516

Quist, A. J. L., Holcomb, D. A., Fliss, M. D., Delamater, P. L., Richardson, D. B., & Engel, L. S. (2022). Exposure to industrial hog operations and gastrointestinal illness in North Carolina, USA. *Science of The Total Environment*, *830*, 154823. https://doi.org/10.1016/j.scitotenv.2022.154823

R Core Team. (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing* [Computer software]. https://www.R-project.org/

Resources, N. C. D. O. E. Q. D. of W., & Administrative. (2019). *Administrative Code Section: 15A NCAC 02B .0100: Procedures for Assignment of Water Quality Standards 15A NCAC 02B .0200:*

Classifications and Water Quality Standards Applicable to Surface Waters and Wetlands of North Carolina 15A NCAC 02B .0300: Assignmen.

Rozemeijer, J. C., & Broers, H. P. (2007). The groundwater contribution to surface water contamination in a region with intensive agricultural land use (Noord-Brabant, The Netherlands). *Environmental Pollution*, *148*(3), 695–706. https://doi.org/10.1016/j.envpol.2007.01.028

Sinha, E., & Michalak, A. M. (2016). Precipitation Dominates Interannual Variability of Riverine Nitrogen Loading across the Continental United States. *Environmental Science & Technology*, *50*(23), 12874–12884. https://doi.org/10.1021/acs.est.6b04455

Soller, J. A., Schoen, M. E., Bartrand, T., Ravenscroft, J. E., & Ashbolt, N. J. (2010). Estimated human health risks from exposure to recreational waters impacted by human and non-human sources of faecal contamination. *Water Research*, *44*(16), 4674–4691. https://doi.org/10.1016/j.watres.2010.06.049

Spoelstra, J., Schiff, S. L., & Brown, S. J. (2020). Septic systems contribute artificial sweeteners to streams through groundwater. *Journal of Hydrology X*, *7*, 100050. https://doi.org/10.1016/j.hydroa.2020.100050

Strickling, H. L., & Obenour, D. R. (2018). Leveraging Spatial and Temporal Variability to Probabilistically Characterize Nutrient Sources and Export Rates in a Developing Watershed. *Water Resources Research*, *54*(7), 5143–5162. https://doi.org/10.1029/2017WR022220

Tolouei, S., Dewey, R., Snodgrass, W. J., Edge, T. A., Andrews, R. C., Taghipour, M., Prévost, M., & Dorner, S. (2019). Assessing microbial risk through event-based pathogen loading and hydrodynamic modelling. *Science of The Total Environment*, *693*, 133567. https://doi.org/10.1016/j.scitotenv.2019.07.373

Vermeulen, L. C., & Hofstra, N. (2014). Influence of climate variables on the concentration of Escherichia coli in the Rhine, Meuse, and Drentse Aa during 1985–2010. *Regional Environmental Change*, *14*(1), 307–319. https://doi.org/10.1007/s10113-013-0492-9

Weidhaas, J., & Lipscomb, E. (2013). A new method for tracking poultry litter in the Potomac Basin headwaters of West Virginia. *Journal of Applied Microbiology*, *115*(2), 445–454. https://doi.org/10.1111/jam.12231

Wing, S., Cole, D., & Grant, G. (2000). Environmental injustice in North Carolina's hog industry. *Environmental Health Perspectives*, *108*(3), 225–231. https://doi.org/10.1289/ehp.00108225

Zhu, K., Suttner, B., Pickering, A., Konstantinidis, K. T., & Brown, J. (2020). A novel droplet digital PCR human mtDNA assay for fecal source tracking. *Water Research*, *183*, 116085. https://doi.org/10.1016/j.watres.2020.116085