Predicting *E. coli* Levels in the Harpeth River Watershed

ABSTRACT

Efforts to monitor recreational waters for fecal contamination have shifted away from culturing pathogen indicator organisms, such as *Escherichia coli*, to developing real-time predictive models. Culturing *E. coli* requires 18 to 24 hours to complete; thus, “do NOT swim advisories” are based on previous day measurements– this traditional approach under-performs compared to predictive models. The purpose of this project was to determine whether predictive modeling methods using Virtual Beach software (USEPA; version 3.0.7) could reasonably predict *E. coli* levels in real-time for the Harpeth River. A comprehensive model was also developed to identify conditions of water quality that correlate with elevated levels of *E. coli*. Both models were developed using multiple-linear regressions to predict densities of *E. coli* in the Harpeth River, which has commonly exhibited densities higher than the U.S. Environmental Protection Agency’s (USEPA) Beach Action Value (BAV) of 235 colony forming units (CFUs)/100 mL.

Water quality data from three locations monitored by Franklin Water Reclamation Facility and weather data from National Oceanic and Atmospheric Association were used to create models. The model’s predictive variables were determined based on Pearson Correlation Coefficients, which indicate the degree of association between each independent variable against the dependent variable: *E. coli*. A 70.18% accurate real-time model and 72.67% accurate comprehensive model showing *E. coli*’s direct correlation with TSS (meaning more suspended sediment indicates more *E. coli*) were created. This approach could be a viable strategy to warn recreators of the safety of the Harpeth River.

BACKGROUND

- Recreational waters are threatened by fecal contamination from urban and agricultural drainage that may contain human pathogens associated with gastrointestinal and respiratory illness.
- The United States Environmental Protection Agency’s Beach Action Value (BAV) of 235 colony forming units (CFUs)/100 mL has been exceeded numerous times in the Harpeth River as shown in Municipal Separate Storm Sewer System (MS4) reports.
- ***Escherichia coli*** (abbreviated as *E. coli*) is bacteria found in the environment, bodies of water, and intestines of people and animals; it may cause illness and infection.
- **Current *E. coli* measurement** in the Harpeth River is done through culturing which takes 18-24 hours with only 50% accuracy.
- Virtual Beach linearizes independent variables such as total suspended sediment, pH, precipitation, river flow, and phosphorus to find the best fit against the dependent variable *E. coli*.
- A **comprehensive model** predicts *E. coli* but contains independent variables whose values require overnight lab analysis to obtain. This model shows the variables that have a strong relation to *E. coli*.

ADVISORY
HIGH BACTERIA LEVELS.
SWIMMING NOT RECOMMENDED
INCREASED RISK OF ILLNESS
AT THIS TIME.
FOR FURTHER INFORMATION, PLEASE CONTACT THE LOCAL COUNTY HEALTH DEPARTMENT, OR VISIT THE DEPARTMENT OF HEALTH'S BEACH WATER QUALITY WEBSITE
(SELECT "BEACH WATER QUALITY" FROM THE A-Z TOPICS LIST

Figure 1: This is a sign warning against recreation due to increased bacteria (*E. coli)* levels found in the body of water.

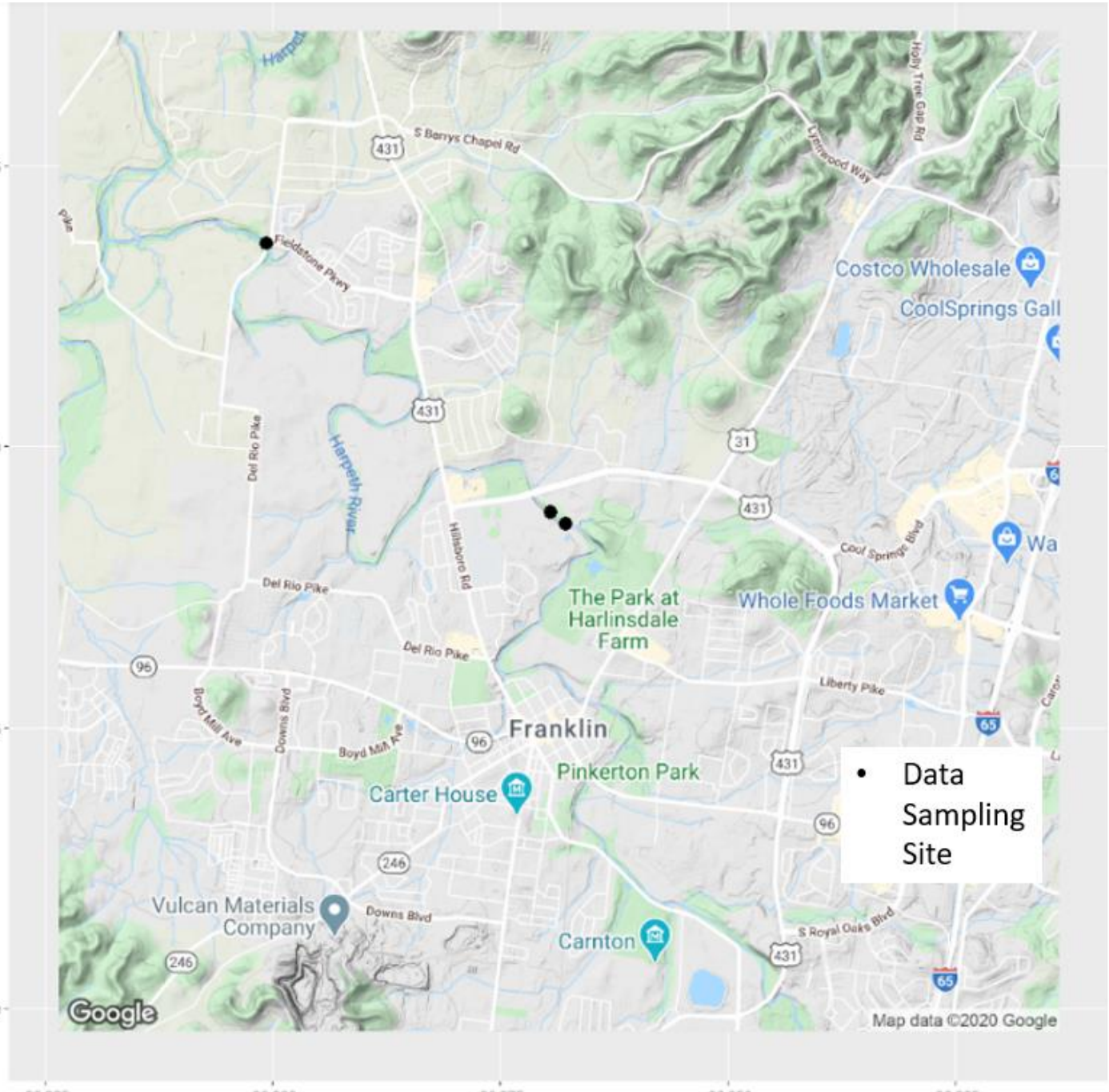


Figure 2: Shows the three sights on the Harpeth River from which data was collected.

GOALS

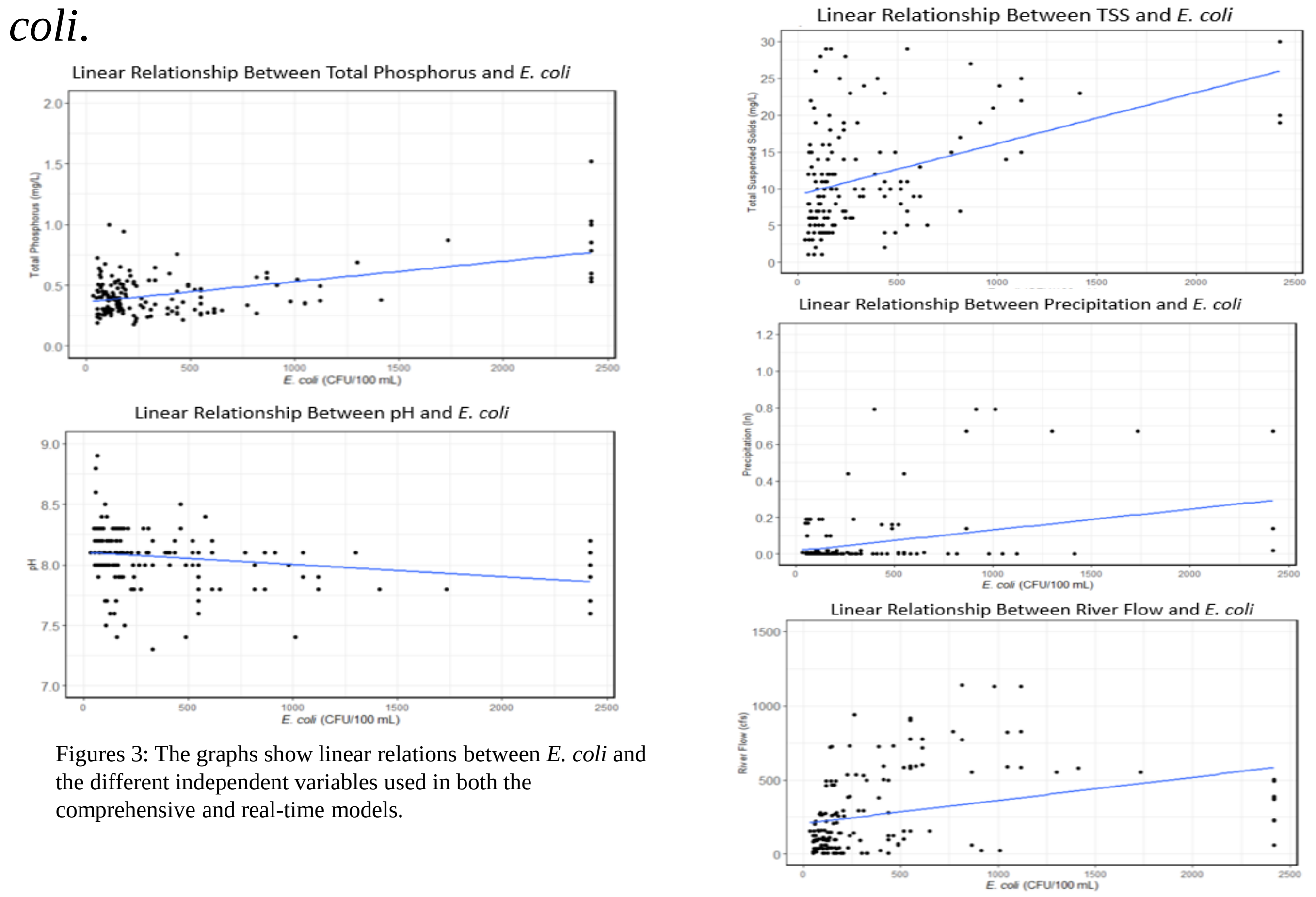
1 Comprehensive Model: Create a comprehensive model using Virtual Beach software that determines independent variables that strongly correlate to *E. coli* levels.

2 Real-time Model: Determine whether Virtual Beach software could be used to accurately predict *E. coli* levels of the Harpeth River in real-time that would increase safety for recreators of the Harpeth River.

METHODS

1. The same methods were used for both models with different variables targeted. Data were compiled from water monitoring stations run by Franklin Water Reclamation Facility and from the National Oceanic and Atmospheric Association. Data for 32 independent variables were plugged in to Virtual Beach software and validated.

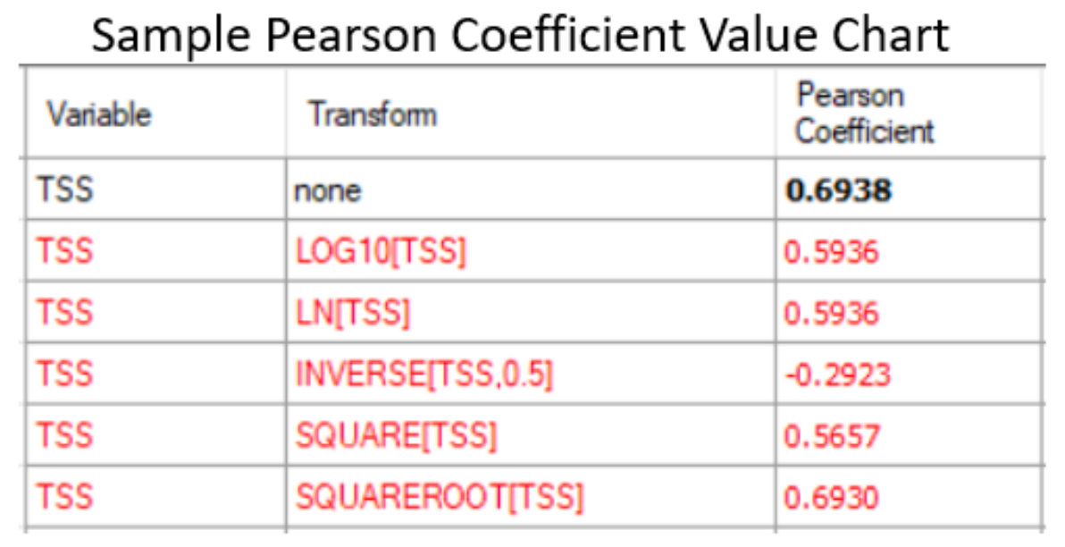
2. Pearson Correlation Results with respect to *E. coli* were then calculated for direct, logarithmic, natural logarithmic, inverse, square, and square root relationships. The Pearson Coefficients were used to determine the best fitting independent variables. A coefficient closer to +/- 1 indicates a strong fit against *E. coli*.



Figures 3: The graphs show linear relations between *E. coli* and the different independent variables used in both the comprehensive and real-time models.

3. 27 models were made and the best fit value, number of false positives and negatives, specificity, sensitivity, accuracy, and parameters were recorded. Percent accuracy was used to determine the best models.

Sample Pearson Coefficient Value Chart

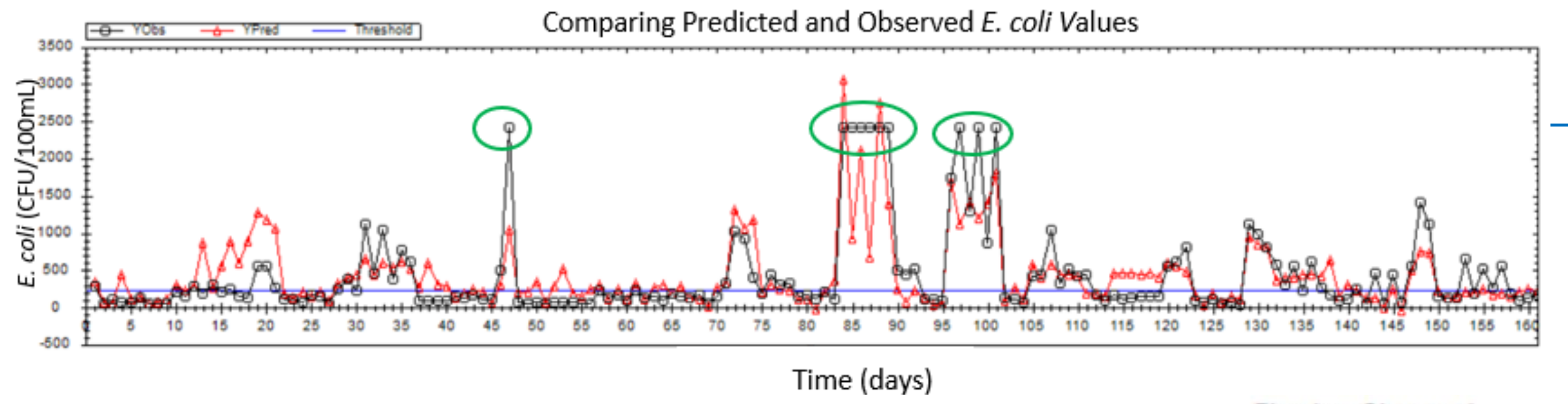


Variable	Transform	Pearson Coefficient
TSS	none	**0.6938**
TSS	LOG10[TSS]	0.5936
TSS	LN[TSS]	0.5936
TSS	INVERSE[TSS,0.5]	-0.2923
TSS	SQUARE[TSS]	0.5657
TSS	SQUAREROOT[TSS]	0.6930

Figure 4: This is a sample chart of the Pearson Coefficient values for different transformations of TSS. The strongest is indicated in black.

DATA AND RESULTS

1: Comprehensive Model



E. coli Beach Action Value (235 CFUs/100 mL of *E. coli*)

Figure 5: Comparison of the predicted and observed values of *E. coli* on given days.

Comprehensive Model Equation:

E.coli = 728.8 + 19.2*(TSS) + 1244*(SQUARE(Precip)) + 17.07*(SQUAREROOT(river.flow.cfs)) - 13.56*(SQUARE(pH))

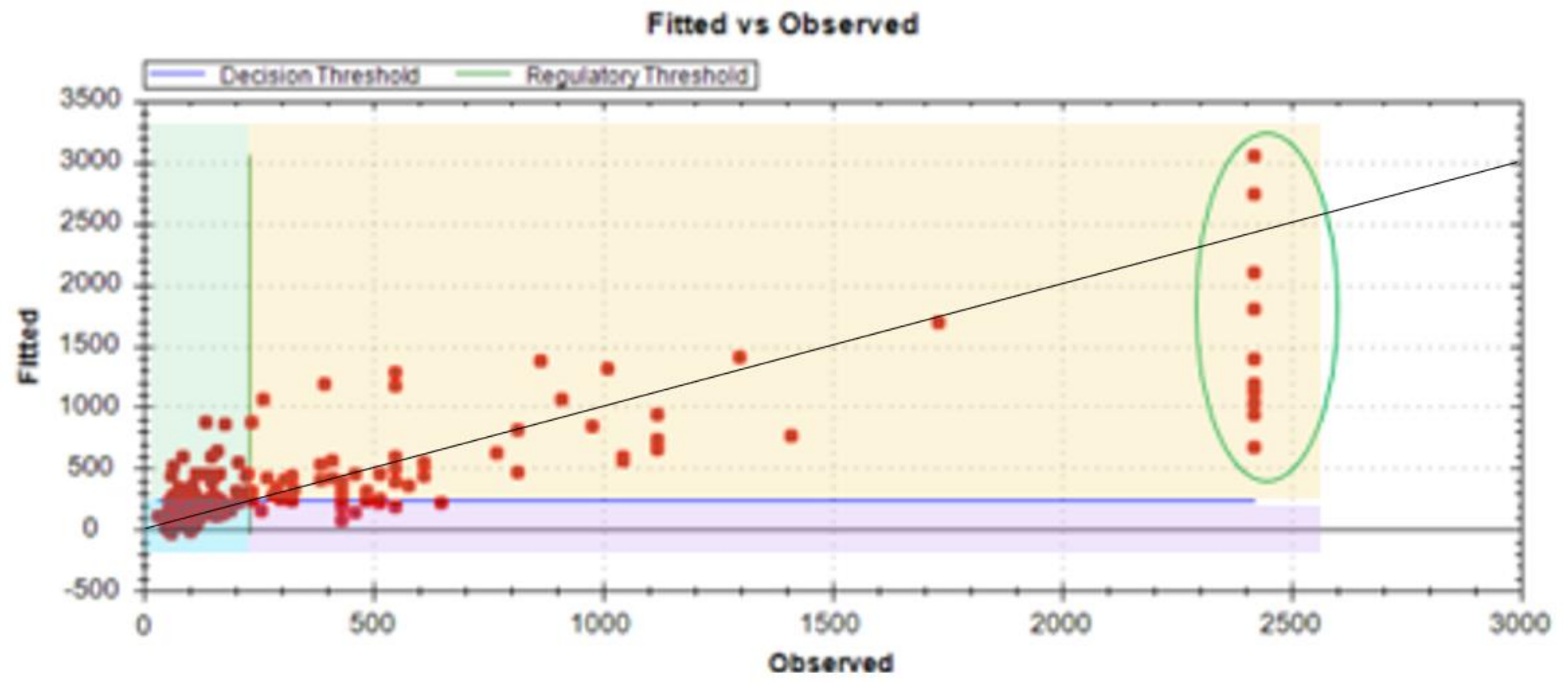
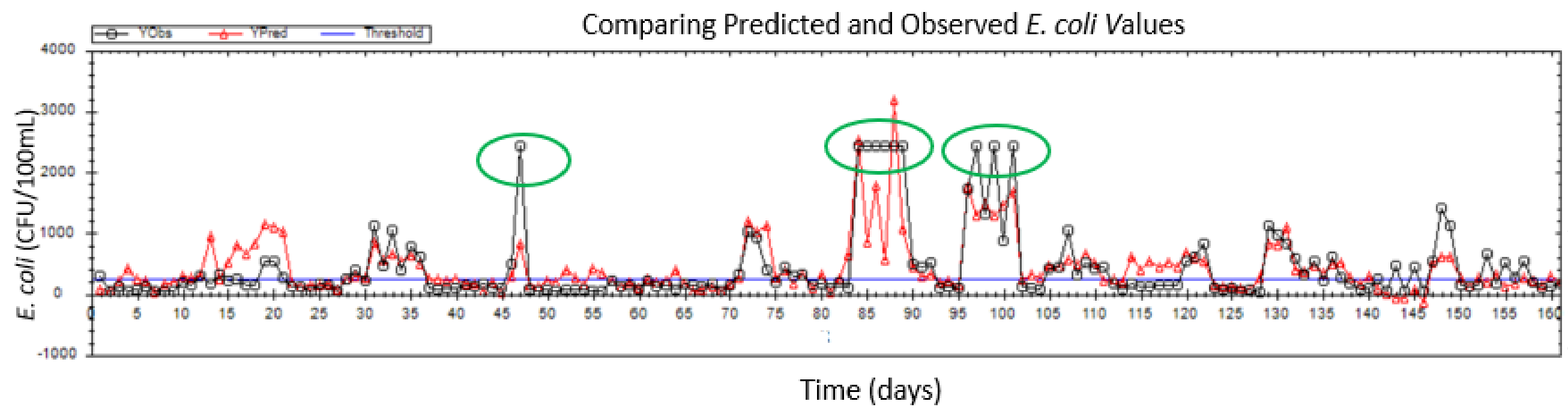


Figure 6: This graph compares the observed *E. coli* values collected by The Franklin Wastewater Treatment Plant to the fitted *E. coli* values that the comprehensive model predicted.

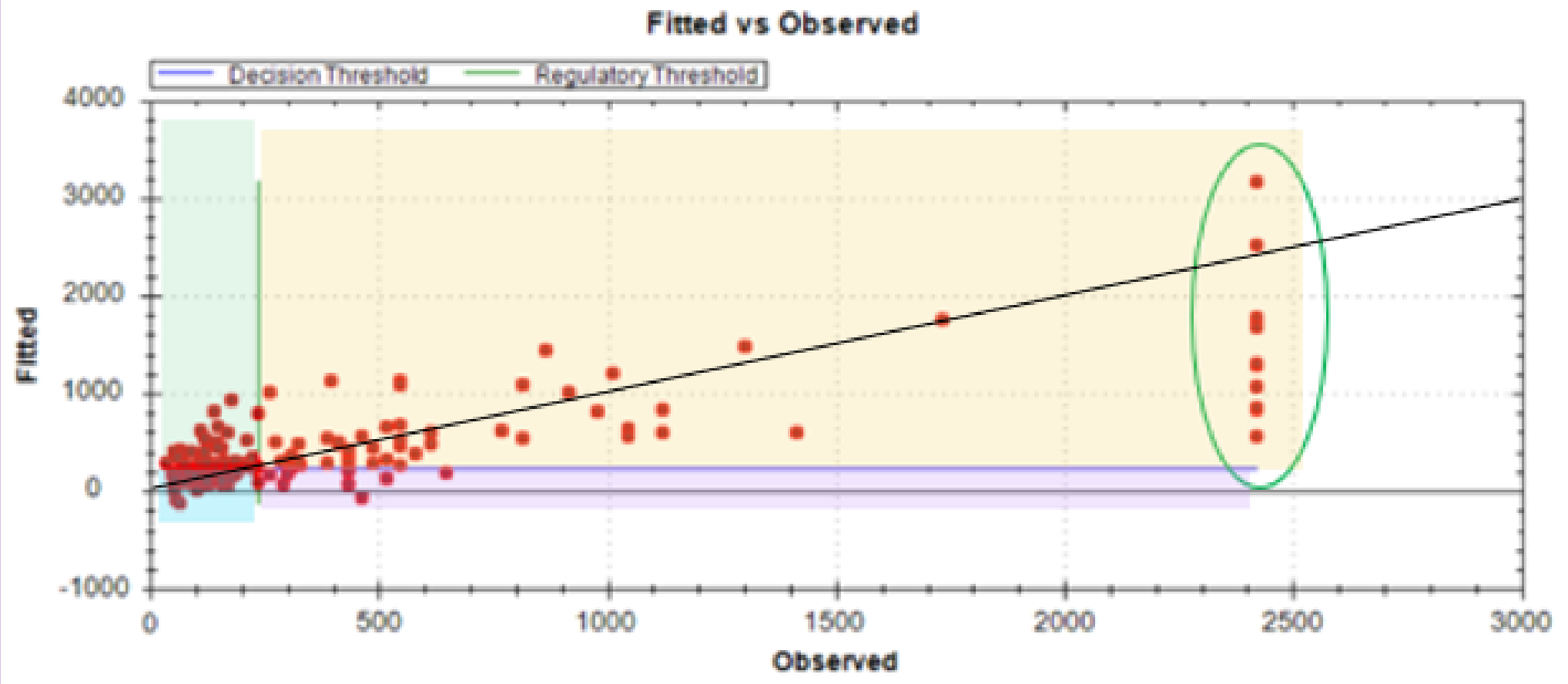
DATA AND RESULTS

2: Real-time Model



E. coli Beach Action Value (235 CFUs/100 mL of *E. coli*)

Figure 7: Comparison of the predicted and observed values of *E. coli* on given days.



Real-time Model Equation:

E.coli = 249.9 + 1337*(SQUARE(Precip)) - 11.91*(SQUARE(pH)) + 1150*(total.P) + 26.8*(SQUAREROOT(river.flow.cfs)

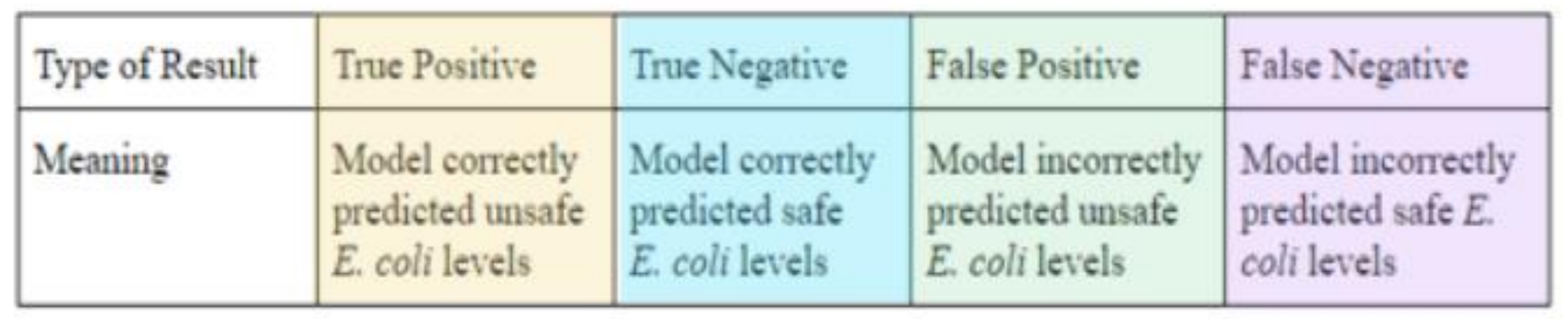
On both Fitted vs. Observed graphs, a best fit line with a slope of one and a y-intercept of zero (black line) would mean the model perfectly predicts *E. coli* levels.

Figure 8: This graph compares the observed *E. coli* values collected by The Franklin Wastewater Treatment Plant to the fitted *E. coli* values that the real-time model predicted.

○ On all graphs, This shape circles the maximum value (2420 CFU/100mL) that the IDEXX Colilert *E. coli* measurement test has the capacity to measure.

3: Overall Results

Explanation of Types of Results



Type of Result	True Positive	True Negative	False Positive	False Negative
Meaning	Model correctly predicted unsafe *E. coli* levels	Model correctly predicted safe *E. coli* levels	Model incorrectly predicted unsafe *E. coli* levels	Model incorrectly predicted safe *E. coli* levels

Figure 9: This table shows the type of results as they correspond with the Fitted vs. Observed graphs (Figures 6 and 8). Each color represents one type of result.

Overall Results from the Comprehensive and Real-Time Models

Model	Best Fit Value	False Positives	Specificity	False Negatives	Sensitivity	Accuracy	Parameters
Comprehensive	2071.7513	35	0.62765957	9	0.8656716	0.7267081	TSS, squareprecip, squarerootriverflow, squarepH
Real-Time	2085.9988	39	0.58510638	9	0.8656716	0.7018634	squareprecip, squarepH, totalp, squarerootriverflow

Sensitivity: Measure of the rate of true positive results

vs.

Specificity: Measure of the rate of false positive results

Figure 10: This table shows the overall statistics from both the comprehensive and real-time models.
*False positive and false negative points are out of 161 data points used.

CONCLUSIONS

- The same process that is being used to determine the safety of beaches can be used to determine the safety of river water.
- **The accuracy of identifying days where *E. coli* levels are high increased to 72.6% and 70.2% respectively.**
- Certain factors such as temperature that are perceived to have a strong relation to *E. coli* levels are proven to be weaker predictors than TSS, square of precipitation, square root of river flow, square of pH, and total phosphorus.
- Maximizing or minimizing the number of false positives or negatives holds different levels of importance for each group analyzing the data. Recreation advisors would want to minimize the number of false negatives, whereas businesses would want to minimize the number of false positives.
- It is worthwhile for cities to invest in and maintain water monitoring equipment so that the necessary variables can be measured to determine the quality of the water.
- One source of error comes from the maximum threshold value that the *E. coli* sampling test was able to measure. This created inaccurate data points that have an effect on the accuracy of the models.

REFERENCES

"About the Watershed." *Harpeth Conservancy*, HWRA, 2020, www.harpethconservancy.org/watershed/about. Accessed 19 Feb. 2020.

"Colilert." *IDEXX US*, IDEXX, www.idexx.com/en/water/water-products-services/colilert/. Accessed 6 Mar. 2020.

"Environmental Modeling Community of Practice." *United States Environmental Protection Agency*, www.epa.gov/ceam/virtual-beach-vb. Accessed 6 Mar. 2020.

"Water Quality State Operation Permit." *Tennessee Department of Environment and Conservation*, Department of Environment &Conservation, www.tn.gov/environment/permit-permits/water-permits1/water-quality-state-operation-permit.html.