# Reproducing and Evaluating USGS Predictive Models for E. Coli at Great Lakes Beaches

**Riana Doctor** and **David Yoder**
{ridoctor,dayoder}@davidson.edu
CSC 371 Machine Learning
Dr. Raghu Ramanujan
Davidson College

**Abstract**

In this project, we reproduced and evaluated site-specific regression models originally developed by the U.S. Geological Survey (USGS) to predict concentrations of *Escherichia coli* (E. coli) at recreational beaches on the Great Lakes in Pennsylvania and Ohio. Using the CSVs provided by the USGS (calibration data and 2019 validation files), we built two regression models in Python with the scikit-learn library, using water quality and environmental data from the USGS. While our reproduced models did not perfectly match the official USGS models, they demonstrated similar trends and still captured the essential relationships between environmental predictors and E. coli concentrations, with our best models achieving $R^2$ scores of 0.476 and 0.556.

## 1 Introduction

In the Great Lakes region, elevated levels of E.coli can pose serious risks to swimmers and require quick advisories from public health officials, making accurate and timely estimates of these E.coli concentrations critical. Traditional water testing methods don't "accurately reflect current water-quality conditions", because results with these methods can only be obtained 18 to 24 hours post-sampling (Francy, Brady, and Zimmerman 2021). To address these inaccuracies, USGS and collaborators developed predictive models that use environmental and water quality measurements to estimate E.coli concentrations quickly. In this paper, we reproduce those models using the calibration and validation datasets provided by the USGS and compare our results to their official models. We then evaluate how closely our reproduced models match the USGS benchmark in terms of fitted coefficients and predictive accuracy ($R^2$, Root Mean Squared Error (RMSE)). By comparing our models to the original, we are able to assess the reproducibility of the USGS approach as well as reflect on model quality and its broader impact. The dataset used comes from model archives, more specifically, the 2019 validation calibration files and validation data files were used, which are both published on ScieneBase.

## 2 Background

The datasets we analyze are part of a U.S. Geological Survey (USGS) ScienceBase model archive summary report. These calibration data files contain observations collected at recreational Great Lakes beaches, including E. coli concentrations and a range of environmental, meteorological, and biological factors that potentially influence bacterial levels. Furthermore, the data was collected in cooperation with regional health agencies and the U.S. Environmental Protection Agency's Great Lakes Restoration Initiative. Discrete water-quality samples were collected weekly for five days per week from May–September 2019. Traditional culture-based methods require 18–24 hours for E. coli results, motivating the use of rapid predictive models. The NowCast approach replaces persistence (previous-day concentration) with site-specific multiple linear regression models developed in EPA's Virtual Beach software (Francy, Brady, and Zimmerman 2021). These models use easily measured environmental surrogates (such as turbidity, water temperature, rainfall, and lake level change) to estimate E. coli concentration or exceedance probability in near real-time, enabling daily public advisories for recreational sites.

### Features

We summarize the broad categories of features typically collected for NowCast model calibration datasets. These features fall into three main categories: Environmental Conditions, Meteorological Variables, and Observational/Biological Variables.

**Environmental Conditions**:

- **RHUM_PCT:** Relative humidity (%).
- **WTEMP_CEL:** Water temperature (°C).
- **CHANGELL_FT:** Lake level change over the last 24 hours (ft).

**Meteorological Variables**:

- **AirportWindSpInst_mph:** Wind speed (mph).
- **AirportRain48W_in:** Weighted 48-h rainfall (in).

**Observational/Biological Variables**:

- **BIRDS_NO:** Number of birds observed at the beach.
- **TURB_NTRU:** Turbidity (NTRU); requires $\log_{10}$ transform.

### Comparative Features

Although both datasets measure similar environmental conditions, the variable names and transformations differ

slightly. Table 1 summarizes the predictors available at each site.

Table 1: Comparison of predictor variables and transformations across Beach 6 (Erie, PA) and Huntington (Cleveland, OH).

| Category | Beach 6 (Beach6_2019) | Huntington (Huntington_2019) |
|---|---|---|
| Target | LAB_ECOLI → $\log_{10}$ | EcoliAve_CFU → $\log_{10}$ |
| Environmental | WTEMP_CEL (water temp, °C) | Lake_Temp_C (water temp, °C) |
| | CHANGELL_FT (lake level change, ft) | LL_PreDay (lake level change, ft) |
| | TURB_NTRU (NTRU) → $\log_{10}$ | Lake_Turb_NTRU (NTRU) → $\log_{10}$ |
| | — | WaveHt_Ft (ft) → $\sqrt{\cdot}$ |
| Meteorological | RHUM_PCT (%) | — |
| | AirportWindSpInst_mph (mph) | — |
| | AirportRain48W_in (in) | AirportRain48W_in (in) |
| Observational | BIRDS_NO (bird counts) | — |

## 3 Experiments

For this project, we focus on two site-specific calibration datasets released by the U.S. Geological Survey (USGS) in 2021 (Francy, Brady, and Zimmerman 2021):

- `Beach6_2019_calibration_data.csv` (Presque Isle Beach 6, Erie, PA)
- `Huntington_2019_calibration_data.csv` (Huntington Reservation, Cleveland Metroparks, OH)

Presque Isle Beach 6 corresponds to USGS station 420839080081801 in Erie, Pennsylvania. Huntington Reservation comprises three sampling sites in Cleveland, Ohio (Central: 412928081560220; West: 412929081561100; Composite: 412928081560215). In both cases, discrete E. coli samples were paired with environmental measurements collected on-site and meteorological data compiled from NOAA's National Centers for Environmental Information and Tides & Currents programs.

The predictive models in the original USGS archive were developed using Virtual Beach version 3.07, with explanatory variables mathematically transformed to improve linearity (e.g., $\log_{10}$ turbidity, square-root rainfall and wave height, untransformed lake temperature and lake-level change). Model performance was compared against the baseline persistence method, with results informing the public via the Great Lakes NowCast system.

The Beach 6 calibration dataset includes 463 water samples, while the Huntington calibration dataset contains 1,001 water samples, each spanning May–September 2019. In addition, the Beach 6 validation dataset includes 100 observations, while the Huntington validation data inlcudes 103 observations. Our target variable, **LAB_ECOLI** (Beach 6) and **EcoliAve_CFU** (Huntington), is transformed to $\log_{10}$(CFU/100 mL) for modeling. Both sites use a threshold of 2.37 $log_{10}$ CFU/100 mL as the water-quality exceedance benchmark.

## 4 Pre-Processing

Before building the Ridge regression models for both Beach 6 and Huntington, we created several graphs with the assistance of ChatGPT to help us visualize the data and discover patterns to understand the nature of the dataset.

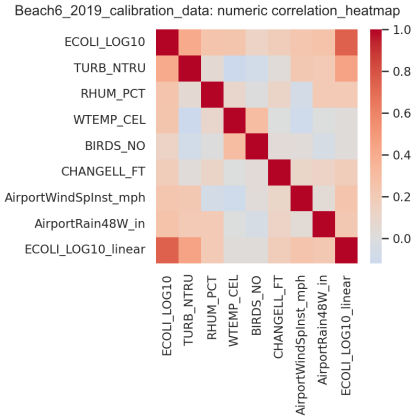


Figure 1: Beach 6 heatmap of variable correlation

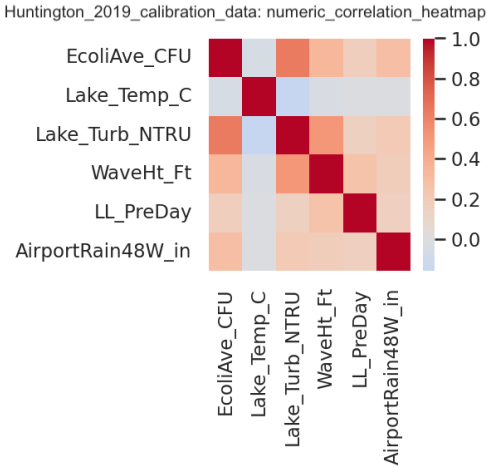


Figure 2: Huntington heatmap of variable correlation

Figures 1 and 2 visualizes a heatmap, where red indicates a positive variable correlation, and blue indicates a negative correlation. It displays the correlations between E. coli and factors like turbidity, rainfall, and water temperature, highlighting the most influential predictors. They also reveal correlations among predictors, point to possible multicollinearity.

Figures 3 and 4 display several noticeable spikes across both sites, which likely correspond to rainfall events, changes in lake level, or increases in turbidity. These peaks are important for identifying when water quality deteriorates and for linking environmental conditions to bacterial concentrations. The visualizations provide temporal context that supports the selection of predictors in the regression model.

Figures 5 and 6 demonstrate that turbidity is strongly associated with E. coli at both Beach 6 and Huntington. The

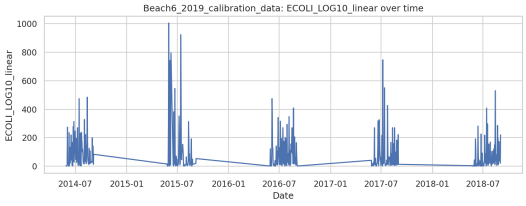


Figure 3: Beach 6 time series plot that shows how E. coli concentrations change across the sampling period.

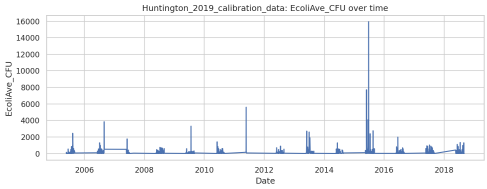


Figure 4: Huntington time series plot that shows how E. coli concentrations change across the sampling period.

consistent positive correlation highlights a key linear relationship driving the regression model. This is important because turbidity serves as one of the most influential environmental predictors.

## Data Cleaning

For the Beach 6 calibration, we started by reading the calibration dataset and identifying the target variable, ECOLI_LOG10. Several features were transformed for consistency with the existing USGS models. More specifically, turbidity values were log-transformed (LOG_TURB), and rainfall over the past 48 hours was square-root transformed (SQRT_RAIN48).

Furthermore, the validation dataset for Beach 6 went through similar feature transformations. However, once we began inspecting the validation dataset for Beach 6, we noticed that the dataset contained non-numeric values. Some

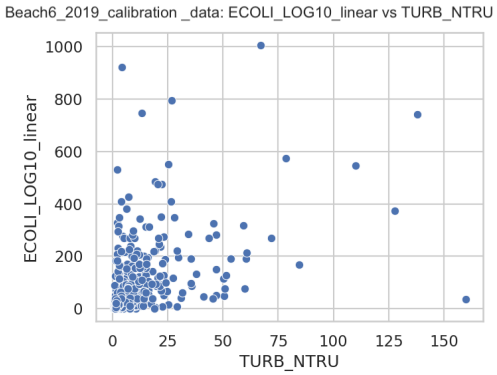


Figure 5: Beach 6 scatter plot that shows the relationship between turbidity and E. coli.

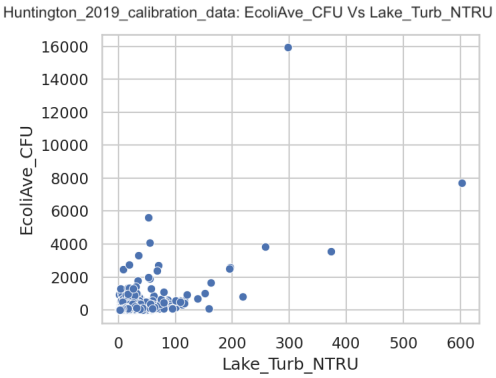


Figure 6: Huntington scatter plot that shows the relationship between turbidity and E. coli.

E. coli measurements were reported as < 1, representing values below the detection limit. Since the model requires numerical input and cannot interpret the < symbol, these values were replaced with 0.5 prior to applying the log10 transformation.

Next, for the Huntington calibration data, the preprocessing steps were similar. The calibration dataset's target variable, EcoliAve_CFU, was log-transformed to EcoliAve_CFU_log10. Rainfall and wave height were square-root transformed, and lake turbidity was log-transformed.

During model evaluation, both the Huntington validation features (X_valid) and the target variable (y_valid) contained missing values. Specifically, X_valid had missing values in Lake_Turb_NTRU_log10 (1 missing), SQRT_WaveHt (3 missing), and LL_PreDay (7 missing), while y_valid had 3 empty entries. Because Ridge regression cannot handle NaN values directly, we applied the `SimpleImputer` function from scikit-learn to replace missing feature values with the mean of the corresponding variable. However, since missing values in the target variable cannot be imputed, the three rows with missing y_valid entries were removed. This approach ensured that the model could be trained and evaluated only on complete observations.

# 5 Model Implementation: Beach 6 and Huntington

Ridge Regression models were used for both the Beach 6 and Huntington data. This was implemented using Python's scikit-learn library. For both models, the calibration dataset was used to train the model, while the validation dataset was reserved for testing and assessing generalization.

First, we used hyperparameter tuning by using GridSearchCV with a 5-fold cross-validation approach. The dataset was split into five subsets, iteratively training the model on four folds and validating on the fifth. The regularization parameter alpha was tested over a range of values [0.01, 0.1, 1, 10, 100]. This ensured that we generated a robust estimate of model performance.

Next, based on cross-validation results, the optimal alpha was determined to be 1 for both, and the final Ridge model

was trained on the full calibration dataset. Model performance on the calibration dataset was assessed using $R^2$ to measure explained variance and RMSE to quantify prediction error. Additionally, scatter plots of predicted versus actual values were generated to visualize model accuracy.

Lastly, to evaluate the models' generalization ability, the same feature transformations applied to the calibration dataset were applied to the validation sets, and the trained model was used to generate predictions. Calculating $R^2$ and RMSE on the validation datasets allowed us to determine how well the models performed on unseen data and to identify any potential overfitting.

## 6 Results

### Model 1: Beach 6

The results for Beach 6 showed that our Ridge regression model closely mirrored the original published model in both structure and performance. The fitted coefficients for key predictors, including turbidity, relative humidity, and water temperature, were nearly identical to those in the original model. On the calibration dataset, the model achieved an $R^2$ of 0.476 and an RMSE of 0.480, which are values that closely match the original model's reported performance. This confirmed that the fitted coefficients and overall predictive structure remained stable. However, when applied to the validation dataset, performance declined, with $R^2$ dropping to 0.253 and RMSE increasing slightly to 0.489. This reduction in predictive accuracy highlights potential overfitting and indicates that while the model is effective within the calibration sample, it does not generalize as strongly to unseen data. An important factor is that the validation dataset contained only 100 observations, compared to the calibration data, which contains 463 water samples. So, the smaller dataset is most likely the cause of the higher variance estimates and less stable $R^2$ values. The scatter plot of predicted versus observed E. coli concentrations for the calibration dataset Figure 7 shows points clustering closely around the 1:1 line, while the corresponding validation scatter plot Figure 8 displays greater dispersion, reflecting the weaker performance on unseen data.

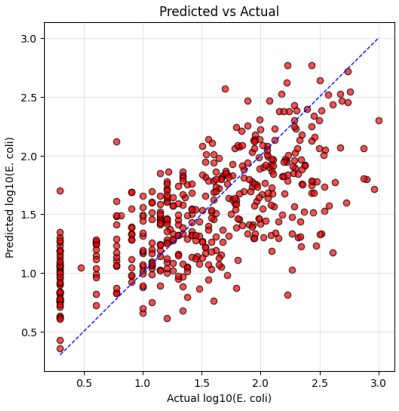


Figure 7: Predicted vs Actual Values for calibration data (Beach 6)

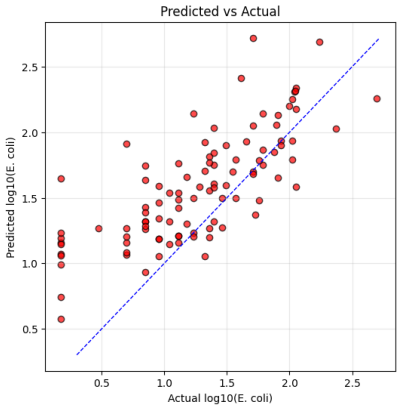


Figure 8: Predicted vs Actual Values for validation data (Beach 6)

### Model 2: Huntington

The Huntington model also closely mirrored the original published model in both structure and performance. The fitted coefficients were again closely aligned with those reported in the original study, specifically for turbidity, rainfall, and lake temperature. The model achieved an $R^2$ of 0.556 and an RMSE of 0.424 on the calibration dataset, slightly outperforming the original model in terms of predictive accuracy. But, as with Beach 6, validation performance revealed weaker generalization, with $R^2$ decreasing to 0.288 and RMSE rising to 0.569. Similar to the Beach 6 validation data, the Huntington validation data contained fewer than 103 observations after cleaning, compared to the calibration data, which contains 1,011 water samples. So, the smaller dataset is most likely the cause of the higher variance estimates and less stable $R^2$ values. The scatter plot for calibration Figure 9 shows predictions aligning more closely with observed values along the 1:1 line. However, the validation scatter plot Figure 10 illustrates a wider spread of points, highlighting greater error and weaker performance on unseen data.

Together, the results show that the Ridge regression successfully replicated the coefficient patterns and predictive framework of the original models while also revealing the challenges of generalizing to new conditions.

## 7 Broader Impacts

Predictive beach "nowcast" models have clear public-health value by reducing the 24-hour culture lag in traditional E. coli testing. When accurate, they can improve same-day advisories and potentially reduce illness (Heasley 2021; Guo 2021; Searcy 2018; 2023). However, these systems also raise equity considerations. Firstly, access to clean, open beaches is not evenly distributed. Studies document landscape-scale inequities in coastal access and amenities, with neighborhoods of color and lower-income communities often facing more barriers or reputational harms from recurrent closures (Montgomery 2015; Twichell 2022). Additionally, water-quality burdens and infrastructure gaps disproportionately affect socially vulnerable populations across

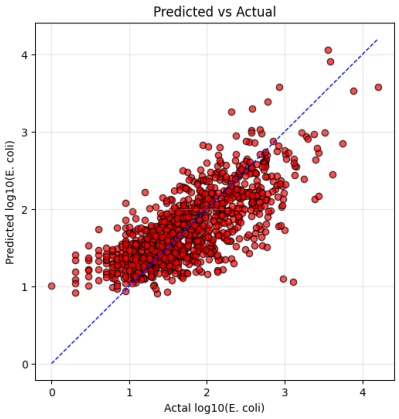


Figure 9: Predicted vs Actual Values for scikit-learn training data (Huntington)

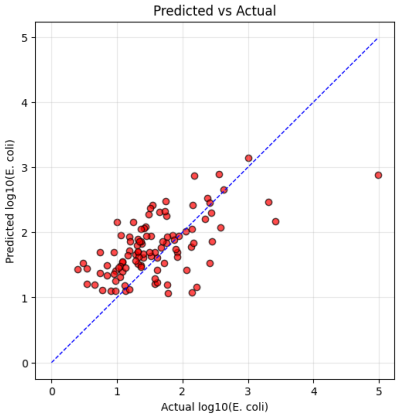


Figure 10: Predicted vs Actual Values for scikit-learn test data (Huntington)

the U.S. (Gochfeld and Burger 2011; Neville 2022; US Water Alliance 2017). So, considering that these communities may already face a lack of investment in infrastructure and monitoring, if predictive systems consistently flag certain beaches as "unsafe", then these areas risk being stigmatized. Communities may be labeled as "dirty" or unsafe for recreation. This can reinforce stereotypes tied to poverty or racial demographics, rather than reflecting inequities in environmental management. While these methods can provide critical real-time information, they also risk shifting responsibility away from systemic solutions and toward communities already burdened by environmental risk. Overall, predictive models can improve public health outcomes, but their deployment should be used with attention to equity, community, and trust.

## 8 Conclusions

Overall, we successfully reproduced USGS regression models to predict E. coli at Great Lakes beaches. Our Ridge models captured key predictor relationships and closely matched original coefficient patterns, though validation showed reduced generalization. These results highlight the models' usefulness for rapid water-quality estimation while emphasizing the need for ongoing recalibration and monitoring under changing conditions.

## 9 Acknowledgements

Chatgpt assisted us in our initial visualizations during the pre-processing phase. We also used Claude as guidance on conceptual ideas when we found ourselves stuck. Also, ChatGPT assisted with LaTeX formatting during the writing of this paper.

## References

Francy, D. S.; Brady, A. M.; and Zimmerman, T. M. 2021. Data for multiple linear regression models for estimating escherichia coli (e. coli) concentrations or the probability of exceeding the bathing-water standard at recreational sites in ohio and pennsylvania as part of the great lakes nowcast, 2019.

Gochfeld, M., and Burger, J. 2011. Disproportionate exposures in environmental justice and other populations at risk. *Environmental Justice*.

Guo, J. e. a. 2021. Development of predictive models for "very poor" beach water quality. *Environmental Science & Technology*.

Heasley, C. e. a. 2021. Systematic review of predictive models of microbial water quality at freshwater recreational sites. *Environmental Modelling & Software*.

Montgomery, M. C. e. a. 2015. An environmental justice assessment of public beach access in california. *Applied Geography*.

Neville, J. A. e. a. 2022. Water quality inequality: a non-targeted hotspot analysis for the united states. *Hydrological Sciences Journal*.

Searcy, R. T. e. a. 2018. Implementation of an automated beach water quality nowcast system. *Journal of Environmental Management*.

Searcy, R. T. e. a. 2023. Data-driven beach water quality forecasting. Technical report, NOAA.

Twichell, J. H. e. a. 2022. Landscape-scale inequities in coastal access in rhode island. *Land*.

US Water Alliance. 2017. An equitable water future: A national briefing paper.