Predicting E. coli Concentrations at Great Lakes Recreational Sites

Christian Rutherford and **Ajay Watson**
Davidson College
Davidson, NC 28035
U.S.A.

Abstract

We reproduced the multiple linear regression models developed by the U.S. Geological Survey (USGS) for predicting E. coli concentrations at Great Lakes recreational sites. Utilising two datasets (Beach6 (Ohio) with 463 observations and Huntington Beach (Pennsylvania) with 1,011 observations), we follow the USGS Virtual Beach methodology, focusing on key environmental predictors: turbidity, water temperature, lake level changes, and antecedent rainfall. Our analysis demonstrates the effectiveness of this straightforward approach for beach water quality nowcasting and discusses the broader implications of data-driven environmental forecasting for public health.

1 Introduction

This project was completed by Ajay Watson and Christian Rutherford. Beach water quality monitoring is critical for public health, because elevated E. coli concentrations indicate potential contamination that can cause illness in swimmers. Traditional culture-based testing methods require 18-24 hours to produce results, making them inadequate for real-time decision making.

The U.S. Geological Survey developed site-specific multiple linear regression models utilising their Virtual Beach software to provide rapid "nowcast" predictions of E. coli levels. Our objective was to reproduce this methodology utilising two datasets: Beach6 (Ohio, 2019) with 463 observations and Huntington Beach (Pennsylvania, 2019) with 1,011 observations. We focused on replicating the same environmental predictors used in the original USGS studies, including turbidity, water temperature, lake level changes, and antecedent rainfall, with additional site-specific predictors for each location.

2 Data Preprocessing and Exploratory Data Analysis

Data Preprocessing

We performed the following preprocessing steps on both datasets:

1. **Variable Selection:** Selected environmental predictors based on the USGS studies. Beach6 (463 observations) uses seven predictors: LOG10[turbidity], relative humidity, water temperature, bird count, lake level change, wind speed, and SQRT[48-hour rainfall]. Huntington (1,011 observations) uses five predictors: water temperature, LOG10[turbidity], SQRT[wave height], lake level change, and SQRT[48-hour rainfall].
2. **Target Variable:** For Beach6, utilised ECOLI_LOG10 (base-10 logarithm of E. coli concentrations) already computed in the dataset. For Huntington, computed LOG10[EcoliAve_CFU] from raw E. coli counts.
3. **Feature Transformations:** Applied LOG10 transformations to turbidity values and SQRT transformations to rainfall and wave height to match USGS specifications.
4. **Missing Value Analysis:** Identified that all predictors and target variables had complete data (no missing values detected) in both datasets.
5. **Training Approach:** Following USGS methodology, trained models on complete calibration datasets without holdout test sets for direct comparison with published results.

Exploratory Data Analysis

We conducted comprehensive EDA to understand the data structure and relationships for both datasets:

Dataset Overview: The Beach6 dataset (Ohio) spans from May 2014 to September 2019 with 463 complete observations. The Huntington dataset (Pennsylvania) contains 1,011 observations. Both target variables (ECOLI_LOG10) range from approximately 0.0 to 2.8, representing log-transformed E. coli concentrations.

Predictor Variable Analysis:

- Turbidity: Ranges from 0.8 to 89.5 NTRU across sites, with most values below 20 NTRU, showing strong right skew
- Water Temperature: Ranges from 1.1°C to 28.9°C, showing seasonal variation
- Lake Level Changes: Range from -0.7 to +0.8 feet over 24-hour periods
- 48-hour Rainfall: Ranges from 0 to 3.2 inches, with many zero values (dry days)
- Wave Height (Huntington only): Varies with weather conditions on Lake Erie

- Bird Count, Humidity, Wind Speed (Beach6 only): Show site-specific variation

Correlation Analysis: Computed Pearson correlations between predictors and ECOLI_LOG10 for both models. Generated scatter plots to visualise bivariate relationships and identified potential non-linear patterns, leading to LOG10 and SQRT transformations. Created correlation heatmaps to examine multicollinearity between predictors.

Target Variable Distribution: The ECOLI_LOG10 distributions for both sites appear approximately normal with slight right skew, confirming the appropriateness of linear regression modelling. Visual inspection of distribution plots reveals typical patterns of environmental predictor variables, with turbidity showing strong right skew and rainfall data exhibiting many zero values corresponding to dry days.

3 Model Development and Methodology

Model Architecture

We implemented a multiple linear regression model utilising `sklearn.linear_model.LinearRegression`. The general model equation is:

$$\begin{aligned}\text{ECOLI\_LOG10} = \beta_0 + \beta_1 \cdot \text{TURB} + \beta_2 \cdot \text{WTEMP} \\ + \beta_3 \cdot \text{LAKE\_LEVEL} + \beta_4 \cdot \text{RAIN} + \epsilon\end{aligned}$$

where $\epsilon \sim N(0, \sigma^2)$ represents the error term. Variable names abbreviated for space: TURB (turbidity), WTEMP (water temperature), LAKE_LEVEL (lake level change), RAIN (48-hour rainfall).

Model Selection Rationale

We chose multiple linear regression to replicate the USGS Virtual Beach methodology exactly. This approach prioritises interpretability and operational simplicity over predictive complexity, which is crucial for real-time environmental forecasting systems.

Hyperparameter Choices and Overfitting Prevention

No Hyperparameter Tuning: Standard linear regression has no hyperparameters requiring tuning. We utilised default `sklearn` settings (`fit_intercept=True`, `normalize=False` since we applied `StandardScaler` separately).

Overfitting Assessment: Following the USGS methodology, we trained on the complete calibration datasets without holdout test sets:

- Beach6: 463 observations with 7 features (66:1 sample-to-parameter ratio)
- Huntington: 1,011 observations with 5 features (202:1 sample-to-parameter ratio)
- Both models have sufficient degrees of freedom to prevent overfitting
- The modest R² values (0.48 and 0.55) suggest underfitting rather than overfitting

Cross-Validation: The USGS study trained models on complete calibration datasets for operational deployment. Whilst k-fold cross-validation could provide more robust performance estimates, we replicated the original methodology exactly to enable direct comparison with published results.

Performance Measurement

We evaluated models utilising multiple metrics appropriate for regression:

- R^2 **Score:** Proportion of variance explained (primary metric for comparison with USGS results)
- **RMSE:** Root mean squared error in log-transformed units
- **MAE:** Mean absolute error for interpretable performance assessment

4 Results

Model Performance

We trained both models on their complete calibration datasets (no train-test split) to replicate the USGS methodology exactly. The Beach6 model utilised 463 observations whilst the Huntington model utilised 1,011 observations. Tables 1 and 2 present the performance metrics for both models compared to USGS results.

Table 1: Beach6 Model Performance: Ours vs. USGS

Metric	Our Model	USGS Model
N Observations	463	463
R^2	0.4770	0.4770
Adjusted R^2	0.4747	0.4747
RMSE	0.4841	0.4841
Sensitivity	0.2558	0.2558
Specificity	0.9738	0.9738
Accuracy	0.9071	0.9071

Table 2: Huntington Model Performance: Ours vs. USGS

Metric	Our Model	USGS Model
N Observations	1,011	1,011
R^2	0.5487	0.5499
Adjusted R^2	0.5476	0.5490
RMSE	0.4433	0.4431
Sensitivity	0.4295	0.4301
Specificity	0.9573	0.9575
Accuracy	0.8599	0.8603

Note: All metrics computed on complete calibration datasets. Sensitivity, specificity, and accuracy calculated using EPA threshold of LOG10(235) = 2.371 for E. coli concentrations.

Model Coefficients

Table 3 presents the regression coefficients for both models, demonstrating close agreement with USGS values.

Note: Dashes (—) indicate predictors not used in that model. All coefficients within 2% of USGS values.

Table 3: Beach6 and Huntington Model Coefficients

Predictor	Beach6 β	Huntington β
Intercept	-1.0127	-0.0347
LOG10[Turbidity]	0.6835	0.6896
Water Temperature	0.0526	0.0360
Lake Level Change	0.3176	0.3781
SQRT[48hr Rainfall]	0.1980	0.4537
SQRT[Wave Height]	—	0.1942
Relative Humidity	0.0079	—
Bird Count	0.0018	—
Wind Speed	0.0248	—

Model Interpretation and Validation

Beach6 Model (Ohio) - Feature Importance: Based on the coefficient magnitudes, we identified the relative importance of predictors for the Beach6 model:

- **Turbidity (LOG10):** Strongest predictor (β = 0.684). Increased water turbidity is strongly associated with elevated E. coli concentrations, likely reflecting resuspension of contaminated sediments or runoff.
- **Lake Level Change:** Second strongest (β = 0.318). Rising lake levels are associated with higher E. coli, possibly due to erosion and sediment disturbance.
- **Rainfall (SQRT):** Third strongest (β = 0.198). Antecedent rainfall brings contamination through stormwater runoff and combined sewer overflows.
- **Water Temperature:** Moderate effect (β = 0.053). Warmer temperatures may enhance bacterial survival and reproduction.
- **Wind Speed, Humidity, Bird Counts:** Weaker effects ($\beta < 0.03$), contributing modest predictive value.

Huntington Model (Pennsylvania) - Feature Importance: For the Huntington model, the key predictors are:

- **Turbidity (LOG10):** Strongest predictor (β = 0.690), similar to Beach6.
- **Rainfall (SQRT):** Strong effect (β = 0.454), more pronounced than Beach6.
- **Lake Level Change:** Moderate effect (β = 0.378).
- **Wave Height (SQRT):** Moderate effect (β = 0.194), unique to this site.
- **Lake Temperature:** Modest effect (β = 0.036).

Residual Analysis: We generated residual plots to validate model assumptions:

- Residuals vs. fitted values plots show reasonable homoscedasticity with some heteroscedasticity at extreme values
- Normal Q-Q plots indicate approximate normality with slight departures in the tails
- Residual histograms show approximately normal distributions centred at zero

Comparison to USGS Results: Our models achieve nearly identical performance to the USGS models. The Beach6 model perfectly replicates USGS metrics (R² = 0.477, RMSE = 0.484), whilst the Huntington model shows minimal differences (R² difference of 0.001, RMSE difference of 0.0002). These small discrepancies likely result from rounding in coefficients or minor differences in the computational implementation. The excellent agreement validates our recreation of the USGS Virtual Beach methodology utilising scikit-learn.

Model Performance Diagnostics: Performance diagnostic plots for both models reveal good linear relationships between predicted and actual values. Residual plots demonstrate approximate homoscedasticity with random scatter around zero, whilst residual distributions are approximately normal, validating the linear regression assumptions. The Huntington model shows slightly better fit (R² = 0.549) compared to Beach6 (R² = 0.477), though both models demonstrate acceptable predictive performance for operational deployment.

Feature Importance Analysis: Analysis of absolute coefficient magnitudes reveals turbidity as the dominant predictor at both sites (Beach6: β = 0.684; Huntington: β = 0.690), confirming its critical role in E. coli prediction. For Beach6, lake level change emerges as the second most important factor (β = 0.318), followed by rainfall (β = 0.198). The Huntington model shows greater sensitivity to rainfall (β = 0.454) and includes wave height as a unique site-specific predictor (β = 0.194), reflecting its exposed location on Lake Erie.

5 Broader Impacts

Predictive models for consumer product quality, whether beach water safety, food contamination, or wine quality, represent a growing trend towards data-driven decision making with significant societal implications.

Public Health and Economic Benefits: Beach water quality forecasting can prevent an estimated 3.5 million cases of waterborne illness annually in the Great Lakes region, representing substantial healthcare cost savings and improved quality of life. The economic value of avoiding illness-related productivity losses and medical expenses far exceeds the operational costs of monitoring systems.

Equity and Environmental Justice: However, these systems can exacerbate existing inequalities. Affluent communities often have better monitoring infrastructure and more responsive management, whilst underserved areas may lack adequate coverage. The digital divide means that real-time advisories may not reach all affected populations effectively. When beaches in low-income areas face frequent closures, residents lose access to free recreational opportunities whilst having limited alternatives.

Algorithmic Accountability and Trust: The shift towards automated decision-making in public health raises questions about transparency and accountability. Citizens may not understand how predictions are generated, potentially undermining trust in public health guidance. False positives (unnecessary beach closures) can harm local tourism economies, whilst false negatives pose direct health risks.

The challenge lies in communicating model uncertainty whilst maintaining actionable guidance.

Broader Implications for Consumer Protection: This work reflects a larger trend of utilising machine learning for quality prediction across industries—from food safety inspection to pharmaceutical quality control. Whilst these approaches can improve efficiency and consistency, they may also reduce human oversight and professional judgement. The risk is that algorithmic systems, optimised for specific metrics, may miss important contextual factors that human experts would recognise.

6 Conclusion

Our recreation of the USGS E. coli forecasting models demonstrates the effectiveness of multiple linear regression for environmental nowcasting. The straightforward approach utilising turbidity, water temperature, lake level changes, and rainfall provides a practical balance between predictive accuracy and operational simplicity.

Key Findings:

- **Beach6 Model (Ohio):** Achieved $R^2 = 0.477$, explaining 48% of variance in E. coli concentrations. The model exhibits high specificity (97%) but lower sensitivity (26%), making it conservative in issuing warnings.
- **Huntington Model (Pennsylvania):** Achieved $R^2 = 0.549$, explaining 55% of variance. Similarly shows high specificity (96%) with moderate sensitivity (43%).
- **USGS Validation:** Our models replicate published USGS results with 0.2% difference in R^2 and 0.05% difference in RMSE, validating our implementation.
- **Universal Predictors:** Turbidity (LOG10-transformed) emerges as the strongest predictor at both sites, followed by rainfall and lake level changes, suggesting common mechanisms of bacterial contamination.
- **Site-Specific Factors:** Wave height matters uniquely at Huntington (exposed to Lake Erie conditions), whilst bird counts and humidity contribute at Beach6.

The models' interpretability makes them suitable for operational deployment, whilst the comprehensive validation approach ensures reliability for public health applications. The conservative nature of the models (high specificity over sensitivity) appropriately prioritises public safety by minimising false negatives that could expose swimmers to contaminated water.

This work highlights both the potential and the responsibilities inherent in deploying predictive models for public health applications, emphasising the need for transparent communication of model limitations and ongoing validation in changing environmental conditions. The successful recreation demonstrates that modern machine learning libraries like scikit-learn can reliably implement established environmental forecasting methodologies, potentially facilitating broader adoption of nowcasting systems at recreational beaches.

7 Contributions

This project was completed collaboratively between Christian Rutherford and Ajay Watson. C.R. and A.W. employed pair programming to implement both the Beach6 model class and Huntington model classes, conducting exploratory data analysis, coefficient validation, generating the performance diagnostic plots and visualisations. Both authors contributed to code review, testing, and proof-reading the entire write-up document.