Big Data Analysis for Assessing Climate Change Impacts on Urban Stormwater Management and Water Quality in Montreal

Université Concordia University
GINA CODY SCHOOL OF ENGINEERING AND COMPUTER SCIENCE

Introduction, Objectives, and Methodology

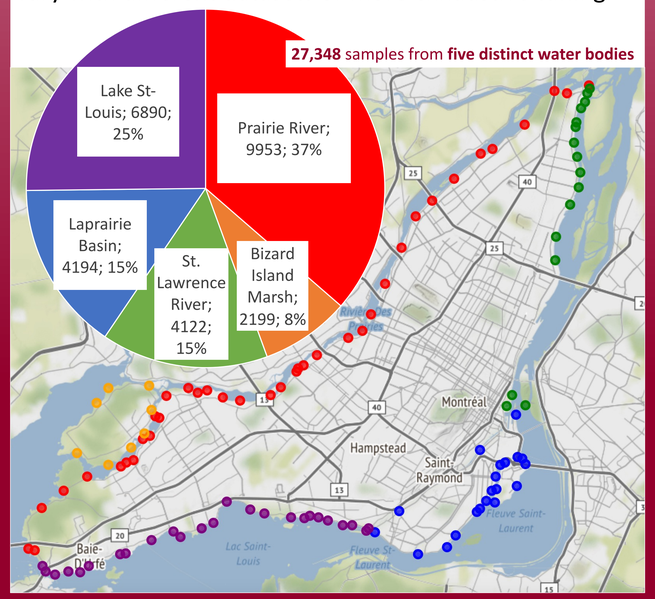
Introduction:
Climate change critically impacts urban stormwater management and local water quality. This study leverages big data and machine learning to evaluate these impacts in Montreal's stormwater systems.

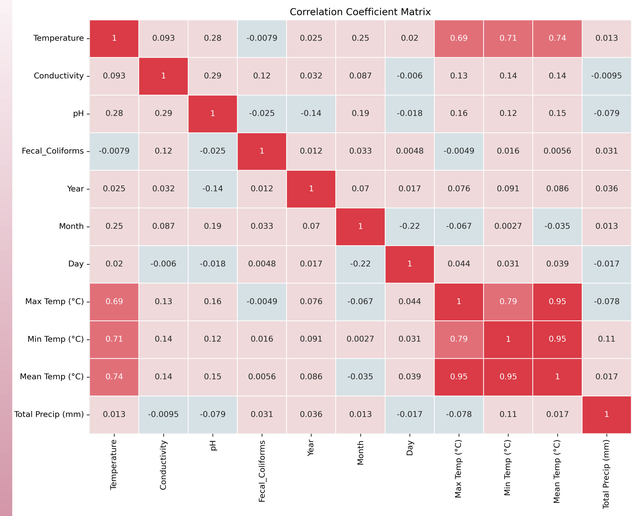
Objective: The research aims to create a predictive model estimating water quality in Montreal's urban stormwater systems under various climate change scenarios.

Expected Outcomes: The expected outcome is a robust model to assess water quality under diverse climate scenarios, facilitating the development of sustainable urban water quality management strategies.

Data Processing: Data from Montreal's "QUALO" and ECCC (2003-2022) underwent extensive preprocessing and integration, focusing on key features like **Body of Water, Temperature, Conductivity, pH, Fecal Coliforms, Year, Month, Day, Max Temp (°C), Min Temp (°C), Mean Temp (°C), and Total Precip (mm)**.

Model Selection and Tuning: The study compared several machine learning algorithms, with the XGBoost Regressor showing superior predictive abilities. Hyperparameter tuning was conducted to optimize algorithm performance, bearing in mind that results may vary with different datasets or more exhaustive tuning.





Exploratory Data Analysis

The analysis encompassed **27,348** samples from **five distinct water bodies**.

Temperature: Temperature correlates with mean, min, and max temperatures across all datasets. Notably, pH also correlates with temperature within specific water sources, though the strength varies (e.g., LaPrairie Basin: 0.41; Lake Saint Louis: 0.28), indicating local environmental factors might influence this relationship.

Conductivity: Correlation patterns are less consistent. Merged data show pH most strongly correlates with conductivity. In specific water sources, temperature and pH display correlations with conductivity, but these vary in strength. This could imply multiple factors influencing conductivity, with their relative importance differing per location.

pH: Conductivity and temperature show the highest correlations with pH, both in merged data and individual water sources. Some areas exhibit a significant correlation with the month, suggesting a potential seasonal pH effect. Yet, correlation with the year varies (e.g., Bizard Island: 0.40; merged data: 0.14), hinting at long-term trends influencing pH.

Fecal Coliforms: Conductivity most consistently correlates with fecal coliform levels. Total precipitation also shows correlation in individual datasets, suggesting rainfall might impact coliform levels through runoff, although this isn't evident in the merged dataset.

Results/Findings

Our analysis, leveraging various machine learning models on the **entire water quality dataset**, yielded key findings. The models' predictive capacity varied for different aspects of water quality, revealing the strengths and limitations of the current data and models.

Temperature

- Highly accurate predictions were achieved with the XGBoost model, demonstrating a training accuracy of 93.61% and a test accuracy of 91.73%. High R-squared values (0.9811 for training, 0.9425 for testing) suggest that the chosen features largely explain the temperature variance.

pH

- The XGBoost model was also successful in predicting pH levels, with training accuracy of 98.60% and test accuracy of 91.20%. The high R-squared values (0.9960 for training, 0.8303 for testing) highlight a strong relationship between pH levels and dataset features.

Conductivity

- Moderate success was noted in predicting conductivity using the XGBoost model, yielding training accuracy of 40.30% and test accuracy of 33.19%. R-squared values (0.9203 for training, 0.5718 for testing) imply that other influential factors not included in the dataset may affect conductivity.

Fecal Coliforms

- This aspect proved challenging to predict. The best performing model, Random Forest, yielded low accuracies (training: 1.27%, test: 1.23%) and R-squared values (training: 0.2848, testing: 0.0686), indicating that current models and features may not fully capture the complexity of factors influencing fecal coliform levels.

In summary, the models effectively predicted certain water quality aspects (i.e., temperature and pH) while performing less well on others (i.e., conductivity and fecal coliforms). The same patterns were evident when models were tested on individual water body datasets, suggesting a need for further investigation into additional features or advanced modeling techniques for improved predictability, particularly for conductivity and fecal coliform levels.

Target Variable	Best Model	Train Accuracy	Test Accuracy	Train R-squared	Test R-squared	Test Adjusted R-squared
Temperature	XGBoost	93.61%	91.73%	0.9811	0.9425	0.9421
pH	XGBoost	98.60%	91.20%	0.9960	0.8303	0.8297
Conductivity	XGBoost	40.30%	33.19%	0.9203	0.5718	0.5713
Fecal Coliforms	Random Forest	1.27%	1.23%	0.2848	0.0686	0.0673

Video of pilot study

LinkedIn

Research Conducted by: Bowen Xu
Research Advisor: Dr. Samuel Li
Department: Building, Civil and Environmental Eng.
Email: x_bowe@live.concordia.ca
Please feel free to reach out with any questions or inquiries regarding the research presented.