LEVERAGING SANITARY SEWER FLOW AND RAINFALL MONITORING DATA FOR SYSTEM INTELLIGENCE

Prepared by York Region for 2023 LIFT Intelligent Water Systems Challenge

york.ca/landl

York Region

LEVERAGING SANITARY SEWER FLOW AND RAINFALL MONITORING DATA FOR SYSTEM INTELLIGENCE

Prepared by York Region for 2023 LIFT Intelligent Water Systems Challenge

Contents

Meet the project team

This project is an interdepartmental collaborative effort within York Region. The project team is a group of engineers from York Region's Inflow and Infiltration Analysis and Strategy team, a data scientist from Data Analytics and Visualization Services and a data analyst from Technology Transformation.

Ranin Nseir (Team Lead) M.Eng., P.Eng.

Manager, System Sustainability Management (A)

York Region, Public Works, Infrastructure Asset Management (IAM)

Area of Expertise

- Inflow and Infiltration (I&I) strategy and analysis

Project Roles and Responsibilities

- Provide guidance to team members and ensure project goals are met within timelines
- Coordinate between team members
- Work with team members to ensure KPIs are met

Faruque Mia Ph.D., P.Eng

Infrastructure Engineer

York Region, Public Works, Infrastructure Asset Management

Area of Expertise

- Data analytics and technical support

Project Roles and Responsibilities

- Prepare historical data for model training
- Provide technical support for model developing
- Review and validate model results

Sophie Xu M.A.Sc.

Project Coordinator

York Region, Public Works, Infrastructure Asset Management

Area of Expertise

- I&I analysis

Project Roles and Responsibilities

- Download, review and prepare model input data
- Coordinate with the data scientist to debug/improve models
- Expand scripts to build models for all nine local cities and towns

Meet the project team (continued)

Cassie Sining Liu

M.Eng., P.Eng., PMP

Project Manager, Inflow and Infiltration Strategy and Analysis (A)

York Region, Public Works, Infrastructure Asset Management

Area of Expertise

- I&I reduction strategy and programming, I&I management in new developments

Project Roles and Responsibilities

- Provide technical support on I&I strategy and methodology
- Coordinate with Water Environment Association of Ontario (WEAO) committee

Peter Chu Su

Data Scientist

York Region, Corporate Services, Data Analytics and Visualization Services

Area of Expertise

- Machine learning model development and output visualization

Project Roles and Responsibilities

- Build and operate pilot machine learning model scripts for four local cities and towns
- Build pilot Tableau dashboard
- Maintain and debug models

Ibraheem Nuaaman

Ph.D., MMA

Data Analyst

York Region, Public Works, Technology Transformation

Area of Expertise

- Data analytics and business intelligence

Project Roles and Responsibilities

- Data visualization
- Operationalize model
- End-to-end automated solution

1. Problem statement

1.1 Background

Inflow and infiltration (I&I) occurs when water other than sewage enters the sanitary sewer system. Excessive I&I consumes system capacity and can lead to basement flooding and spills to the environment. This triggers early-stage servicing challenges and reduces asset life expectancy.

One of the fastest-growing communities in Canada, York Region is a two-tiered municipality with nine local cities and towns and 1.2 million residents. Wastewater servicing within the Region is multi-jurisdictional, so managing wastewater assets requires extensive collaboration between the Region and its nine local cities and towns.

In 2011, York Region developed its I&I Reduction Strategy in collaboration with its nine local cities and towns to meet a condition of approval of the Southeast Collector Individual Environmental Assessment. The Strategy set a target to reduce I&I by 40 million litres per day (MLD) in the York Durham Sewage System by 2031.

1.2 Existing conditions

Data collection

As part of the I&I Reduction Strategy, the Region established a sanitary sewer flow and rainfall monitoring program, with real-time data from over 450 monitoring locations, covering 90% of the wastewater collection system.

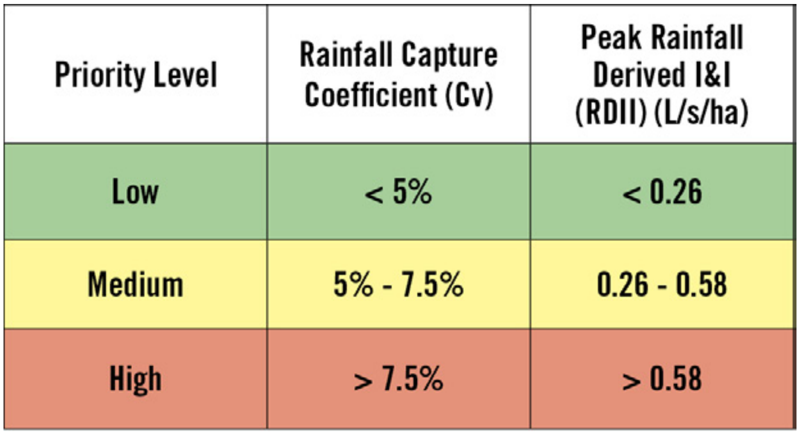
Each flow monitor takes a reading of the sanitary flow data every five minutes and transmits data to a cloud-based data management tool and an on-premise data warehouse. These datasets form the basis for I&I analysis. Results are used to identify and map areas with high I&I to inform investigation/rehabilitation efforts. Since 2013, the program has collected and managed over 900 million data points across the Region.

Current analysis approach

York Region's I&I reduction team conducts I&I analysis using various key performance indicators (KPIs). KPIs commonly used for wet weather analysis are:

Rainfall capture coefficient (Cv), the percent of rainfall volume that enters the sanitary sewer system.

Peak RDII (L/s/ha), the maximum (peak) flow minus diurnal flow normalized by the contributed basin area.



Priority Level	Rainfall Capture Coefficient (Cv)	Peak Rainfall Derived I&I (RDII) (L/s/ha)
Low	< 5%	< 0.26
Medium	5% - 7.5%	0.26 - 0.58
High	> 7.5%	> 0.58

I&I analysis informs I&I reduction programming and helps generate maps to support infrastructure asset management and planning work. Data has also been used to develop I&I reduction targets to each of the nine local cities and towns. However, the current analysis process requires over 14 days for an area with 40 monitoring locations.

1.3 Problem statement

The current I&I analysis process is lengthy, resource-intensive and uses third-party software. **This project seeks to support and advance data-driven decision-making by transforming our 10-year flow and rainfall dataset into actionable insights about I&I in the wastewater system.** The project uses machine learning to automate I&I analysis procedures and develop scenario planning and forecasting capabilities.

Due to climate change, major rainfall events are expected to occur at increased frequency, volume, and intensity. This project prepares the Region for the increased risk of flooding and system surcharges into the environment resulting from severe storms.

1.4 Goals and metrics

The goals and metrics for measuring successes for this project are:

1. Streamline I&I data analysis and decrease manual involvement by reducing processing time by at least 50% and automating backend data transfer and extraction processes
2. Build on the success of the pilot model and deploy a full-scale solution to cover the entire Region while maintaining an accuracy level above 85%
3. Augment proof of concept of the pilot model to improve the scenario planning model and integrate it with a time series model to predict the system response to simulated wet weather events
4. Enhance robustness and predictability of the tool by building a time series model
5. Create a sharable, user-friendly, interactive dashboard with embedded priority area mapping

2. Solution

2.1 Pilot Model

In 2018, York Region's I&I team started streamlining the annual I&I analysis process using an in-house machine learning pilot project (presently known as the Intelligent I&I tool) in coordination with York Region's Data Analytics and Visualization Services (DAVS) and Technology Transformation (TT) teams. Scripts for four of the Region's cities and towns were built in R, a programming language, using machine-learning algorithms. Figure 1 depicts the pilot project progress timeline.

Rainfall and sanitary sewer flow time series data are the model inputs, as well as external factors such as predictable human behavioral factors. Data points are feature engineered into variables for:

- a proxy variable that accounts for the relationship between wet weather and dry weather flow
- flow adjustments based on the previous day's storm events
- rain volume
- wet and dry weather flows

The model predicts system response to storms in terms of Cv and RDII, using a scenario planning model, and I&I flows, using a time series model.

At the beginning of the proof-of-concept stage, several algorithms were tested, but random forest performed best in all three metrics, especially computing power. Random forest is widely used in traditional machine learning and is less likely than other models to overfit data when outliers are present.

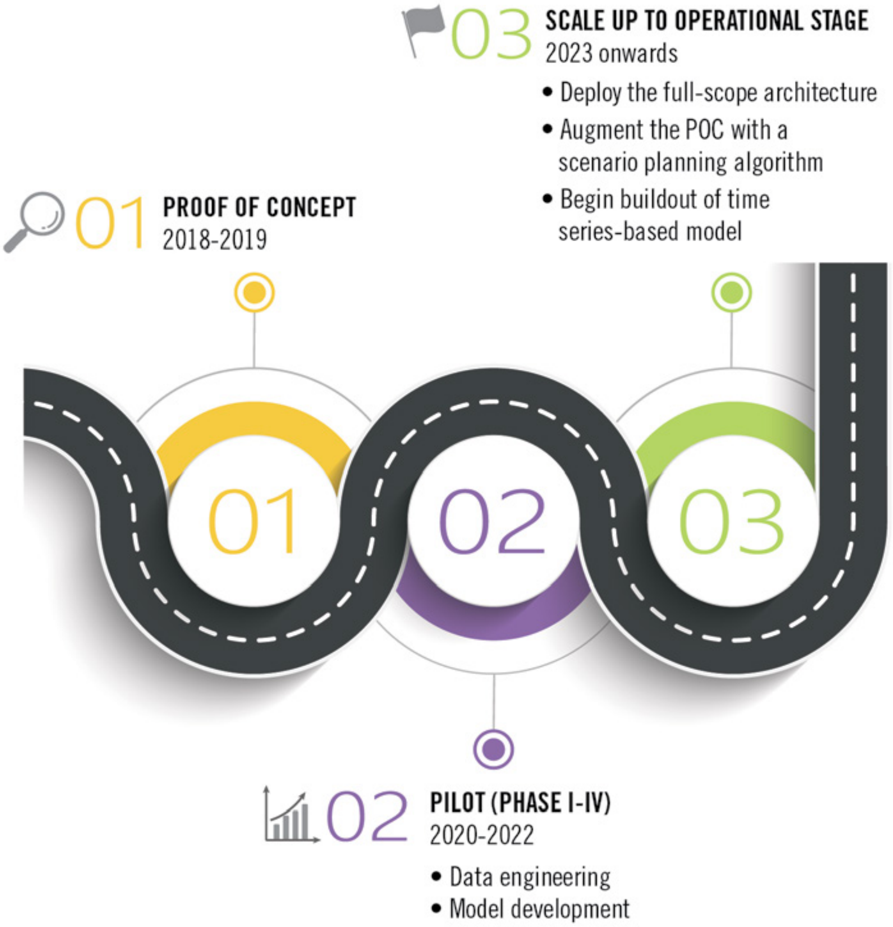

03
SCALE UP TO OPERATIONAL STAGE
2023 onwards
• Deploy the full-scope architecture
• Augment the POC with a scenario planning algorithm
• Begin buildout of time series-based model
01
PROOF OF CONCEPT
2018-2019
01
02
03
02
PILOT (PHASE I-IV)
2020-2022
• Data engineering
• Model development

Figure 1. I&I machine learning pilot project progress timeline

The pilot model achieved over 90% accuracy and reduced time to produce I&I priority maps by 78%. The model uses resampling to validate results. Model results are compared to results obtained from the Region's I&I analysis tool, then the comparison process is repeated to verify model predictions and ensure that performance remains consistent. The model is expected to continue to learn and improve as more data is added and more monitoring areas are analyzed.

2.2 Model scale up

Building on the success of the pilot model, this project seeks to expand and operationalize the machine learning model Region-wide. The full-scale model will improve results visualization, as well as efficiency and accuracy in mapping priority areas. The model will also be able to forecast how the system will react to rainfall events based on historical data.

Deploying the full-scope model requires the following steps:

1. **Train the remaining five cities and towns with the current script** (completed) – this prepares for scaling up to a master script that covers the entire Region.
2. **Model data integration: bridge source data to data warehouse** (retained a consultant and initiated work - 80% complete) – includes upgrading the architecture to include the SQL Server and creating automated SSIS Pipelines for integration.

3. **Improve the current scenario planning model and upgrade the predicting functionality by building a time series model** (expected completion by Q4 2023)

   a. **Augment current proof of concept with a scenario planning algorithm**
   This comprehensive feature engineering process involves adding new features and assessing their significance with the aim of improving the model's performance and overall predictive capabilities.

   b. **Build out a time series-based model to predict system responses**
   Begin the buildout of the time-series based model to identify wastewater flows six to seven days in advance. Both the time series and scenario planning will be deployed into MLOps.

4. **Dashboard building and data visualization via automated mapping** (expected completion by Q1 2024) – for an accessible, intuitive and user-friendly experience. An interactive dashboard offers users the opportunity to explore data on a deeper level and make well-informed, data-driven decisions.

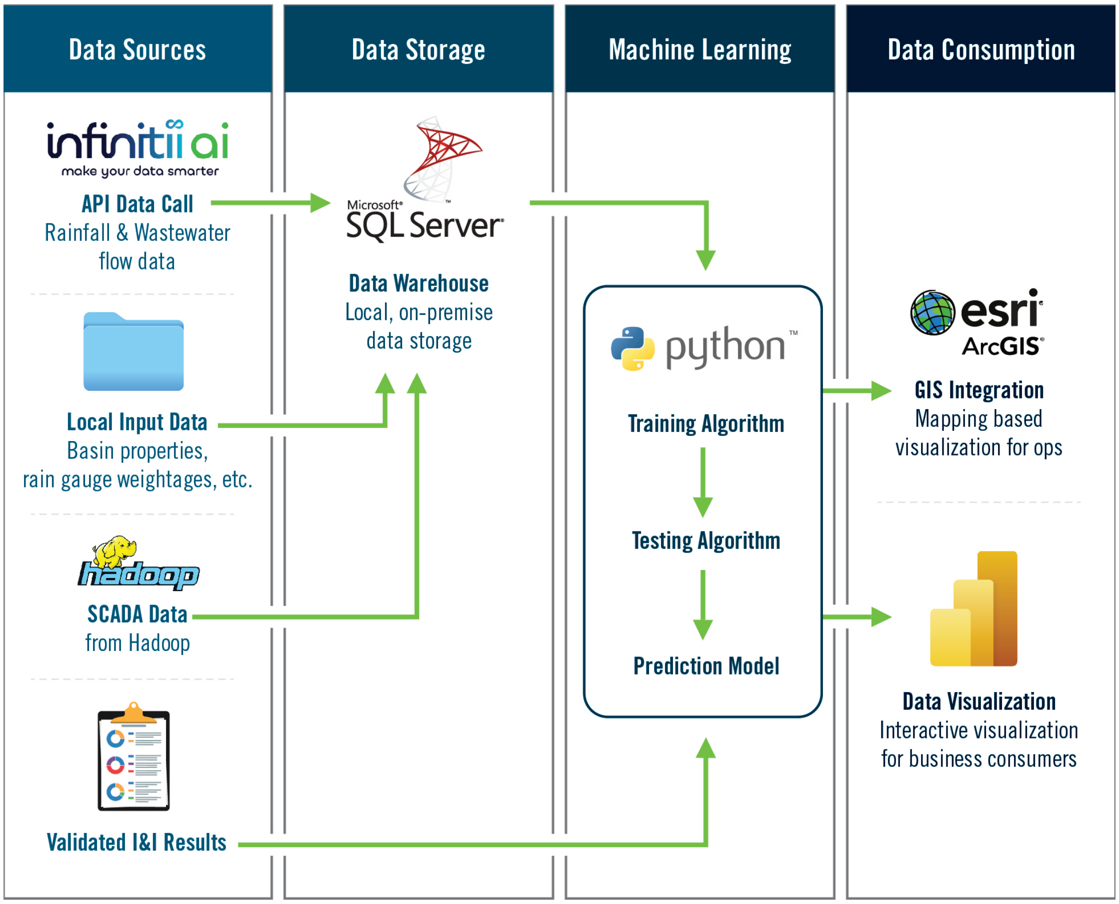


Figure 2. Intelligent I&I data flow

2.3 Model Development

The data streams and QA/QC considerations are shown in Figure 2. Additional details of model development, including hardware, software, algorithms and approaches to training, testing and updating the model are included in Appendices A.1 to A.4.

Basin properties and rain gauge distance weightage are updated annually, with other data updated regularly. Engineers review the data before feeding to the model.

The flow and rainfall monitoring data extracted from infinitii ai are inspected, reviewed and adjusted, if necessary, by the consultant before being uploaded to the final data channel. Then, data is extracted through infinitii ai's Application Programming Interface (API) and goes through a data ingestion and transformation stage. During this process, data is extracted, transformed and loaded in a format ready for querying.

2.4 Value Proposition

The Region's use of machine learning capitalizes on available big data collected over a decade of flow and rainfall monitoring data, leading to better forecasting, faster responses and adaptation to different scenarios. This will bring us the benefits shown in Table 1.

Social benefits	• Reduce the impact of storm events through faster identification of potential basement flooding and sewer overflows, allowing for quicker emergency response • Improve wastewater services by granting municipalities more time to plan mitigation efforts through timely delivery of I&I analysis results
Environmental benefits	• Minimize overflows to the environment by making informed decisions to divert flows to less-stressed systems • Increase system resiliency and adaption to climate change
Economic benefits	• Prioritize system repairs and rehabilitation, leading to more effective spending • Design infrastructure to better accommodate future growth • Save money by shifting from reactive to proactive I&I management

Table 1. Future benefits of the Intelligent I&I machine learning model

2.5 Challenges and mitigation

Figure 2 shows multiple input channels feeding into the model. In some situations, the model did not operate correctly due to a different data format being fed to the script. Although this error is minor, it may take a long time to locate and fix the bug. To ensure smooth model operation, the team has been conducting manual input data QA/QC. In the future, the team is looking into building an Extract, Transform and Load (ETL) process for data warehousing.

2.6 Performance evaluation

Table 2 below shows status compared to performance goals, while Table 3 shows the comparison between the Intelligent I&I machine learning tool and the existing analysis process. Additional acceptance criteria can be found in Appendix A.5.

Performance goals	**Status**
Reduce **processing time** by at least 50%	Achieved: 78% processing time reduction for the pilot
Maintain an 85% **accuracy** level	Achieved: Above 90% accuracy for the pilot
Achieve **other performance metrics** such as precision, recall, F1 score, or area under the receiver operating characteristic curve (AUC-ROC) should be reasonable	Achieved for the pilot and will be tested for the scaled-up model
Demonstrate **robustness** in handling variations in the input data	In progress
Demonstrate **scalability** by being able to process data efficiently and perform well under increased workloads.	In progress

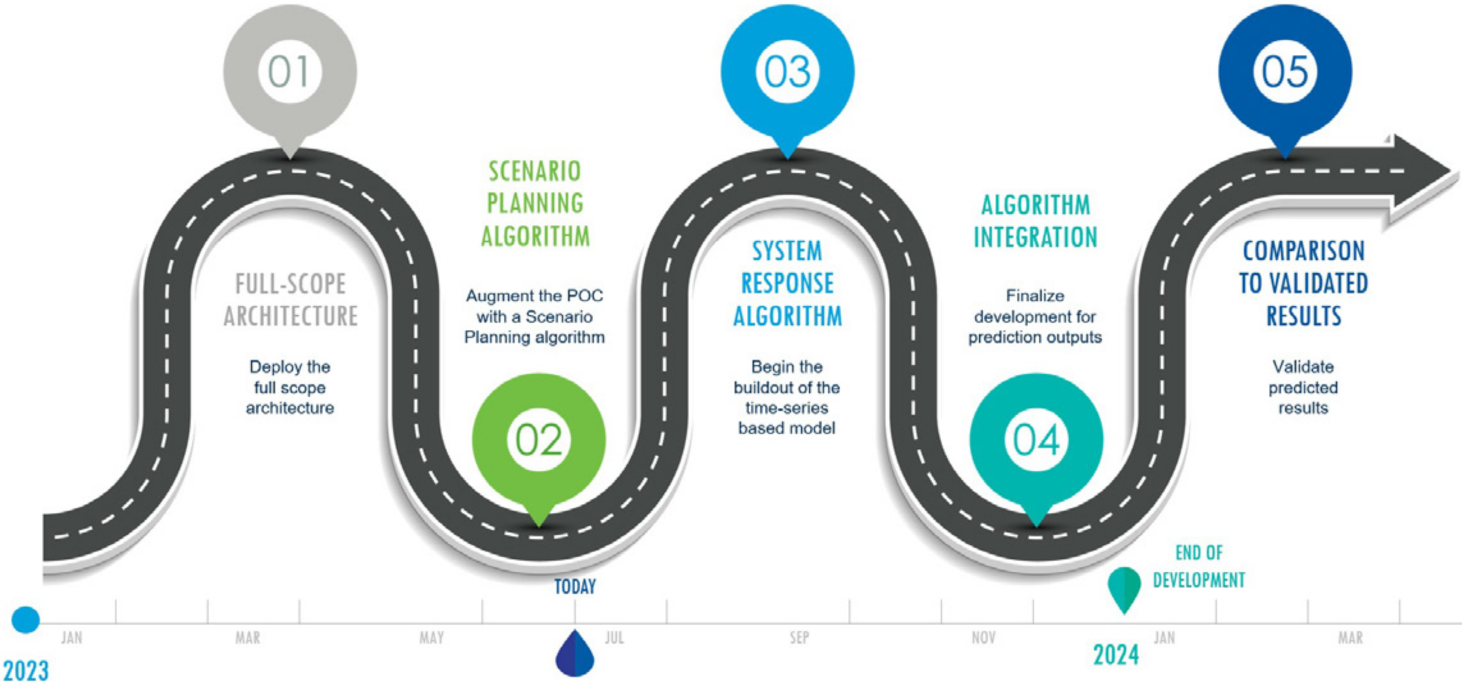
Table 2. Status towards meeting project performance goals

Traditional I&I analysis	**Intelligent I&I tool**
Reliance on third-party software	Accessible, programmed backend calculations developed in-house
Separate data extraction, analysis and mapping processes	Data integration in a consolidated data warehouse, Power BI dashboard, GIS and model scripts
Manually created priority maps	Automatic interactive dashboard with priority mapping
Heavy manual involvement requiring effort from engineers, data scientists and a GIS technologist	Fully automated process with minimal manual involvement required
18 weeks of analysis time	18 days of analysis time
Annual I&I analysis based on 25-year regression	Scenario-based analysis and prediction and near real-time simulation of potential events and sensitivity analysis

Table 3. Comparison between the Intelligent I&I tool and traditional I&I analysis

2.7 Implementation and next steps

As shown in Figure 3, the solution has passed the proof of concept phase and business validation. Implementation cost for consulting services is estimated to be approximately $150,000.



01
FULL-SCOPE ARCHITECTURE
Deploy the full scope architecture
SCENARIO PLANNING ALGORITHM
Augment the POC with a Scenario Planning algorithm
02
03
SYSTEM RESPONSE ALGORITHM
Begin the buildout of the time-series based model
ALGORITHM INTEGRATION
Finalize development for prediction outputs
04
05
COMPARISON TO VALIDATED RESULTS
Validate predicted results
TODAY
END OF DEVELOPMENT
2023
JAN
MAR
MAY
JUL
SEP
NOV
2024
JAN
MAR

Figure 3: Roadmap of the operational stage

Acknowledgements

We would like to acknowledge the partners in this project and the overall monitoring program.

- ADS for providing monitoring equipment and granting access to the I&I analysis software Sliicer
- AECOM for managing the flow monitoring program
- Infinitii ai for providing the data management platform and I&I analysis tools (e.g., Auto I&I)
- Converge for providing process improvement recommendations and assisting with the planning for moving into the full-scope operational stage

Appendix

Appendix A – Model development

A.1 Software required to implement the solution

A third-party software is required and will include a combination of platforms and technologies. They are closely linked to the Microsoft SQL Server suite of products, including, but not limited to, the various engines included with the product. Microsoft SQL Server products that will be leveraged for this solution are Database Engine, Machine Learning Services, and Integration Services.

A.2 Algorithm and training, testing and updating approaches

Current script: At the beginning of the proof-of-concept stage, a number of algorithms were tested to select the optimal one. Random forest was chosen based on its outstanding performance in three key performance measures: root mean squared error (RMSE), simple R-square (the square correlation between the predicted and observed values), and the mean absolute error (MAE).

During model resampling steps, several configurations were tested. The configurations include bootstrapping, 632 bootstrap, and 10-fold cross-validation. Since these three resampling configurations yield similar results, the most popular configuration 10-fold cross-validation was selected as the main resampling algorithm.

An additional factor that may impact model performance is events driven by climate change, such as the increased intensity and/or duration of severe storm events due to global warming. To overcome this challenge, we have set a retrain timeframe of three to five years to help the model keep up with the changes.

Future script outlook: The machine learning algorithm will have four strict phases:

1. **Data ingestion and transformation:** All the relevant data sources are brought in and merged into their relevant data type. They are cleaned and checked for quality.
2. **Feature engineering:** At this step, all non-numerical values are one-hot encoded, and the appropriate scaling is conducted to ensure that the values are accurately represented. The data is then checked for missing values. Meanwhile, various imputation methods (mean, median, mode, etc.) based on the data source are computed. The features then go through a balancing algorithm to identify if they are properly balanced. A balancing of the dataset is set to algorithmically take place if the data is not balanced. A principal component analysis is conducted to reduce the dimensionality of the features.

3. **Modelling:** Modelling consists of model training, and error correction with constant score checking. The modelling is done in an ensemble. It consists of a three-step approach:

    a. **Time-series:** Each flow meter is first run through a set of time-series algorithms, including sARIMA, KATS, Prophet. The future values of the meter are shown.

    b. **First layer shallow error regression:** We assume that the time-series model generated errors in future predictions, as normally all models do. A statistical regression model is then modelled based on the non-time-based features and generated to predict the errors for the future time values, based on how much error the time-series model has generated. The errors are then applied to the predicted future values to get closer to a real value. This is done through a set of regression algorithms, such as xgBoost.

    c. **Second layer deep error regression:** The third step utilizes a deep learning model for those errors that are outlying based on the predictions of the first two steps. These errors produce the largest amount of deviation in the score and are picked to be smoothed. This is done through a deep learning model, such as keras.

4. **Predictions:** Predictions are done in bulk for each flow meter for the near future (seven days) based on the model that has been generated for that meter. These predictions are then placed back into the database.

A.3 Hardware/device utilized

The hardware devices used for the monitoring program entail various flow and rainfall monitoring meters:

Flowmeters

ADS Triton+ and ADS RainAlert III

- Modem FCCID: R17ME910C1WW – Compatible with all 4G LTE-M Networks (Applies to newer models) with 2G fallback

ADS Triton IM / FlowShark

- Modem model – Quadband GSM/GPRS and HSPA (3G Networks) Wireless Networks with 2G fallback

Rain Gauges

ADS Triton+ and ADS RainAlert III

- Modem FCCID: R17ME910C1WW – Compatible with all 4G LTE-M Networks (Applies to newer models) with 2G fallback

Sutron – HSPALink Datalogger

- HSPA modem – 3G Network with 2G fallback

Sutron – XLink 500 Datalogger

- IRIDIUM, Cellular (3G, 4G, CAT-M1/LTE-M)

A.4 Application security and architecture

There are a few concepts around the security that are applied generally to the architecture. They include:

- **Physical security:** We strictly limit access to the physical server and hardware components. For example, we use locked rooms with restricted access for the database server hardware and networking devices. In addition, we limit access to backup media by storing it at a secure offsite location.
- **Operating system security:** Operating system service packs and upgrades include important security enhancements. We apply all updates and upgrades to the operating system after you test them with the database applications.
- **Network security:** Firewalls also provide effective ways to implement security. Our firewall is a separator of network traffic, which can be configured to enforce our organization’s data security policy.
- **SQL Server OS Files Security:** SQL Server uses operating system files for operation and data storage. We restrict access to these files.
- **Encryption:** Encryption enhances our security by limiting data loss even in the rare occurrence that access controls are bypassed.
- **User control access:** We apply controls so that only the required users have access to the various data sources, data views, data pipelines, models, and outputs.

A.5 Additional acceptance criteria

- **Interpretability:** In certain cases, interpretability or explain ability of the model’s decisions may be important. Interpretability will be provided where possible.
- **Integration:** The machine learning model will be integrated with the Region’s Data Platform that currently exists.
- **Documentation and Reporting:** The project requires comprehensive documentation of the model’s architecture, algorithms used, data sources, preprocessing steps, and any other relevant information.

LEVERAGING SANITARY SEWER FLOW AND RAINFALL MONITORING DATA FOR SYSTEM INTELLIGENCE | JULY 2023

Contacts

Ranin Nseir (Team Lead), Manager, System Sustainability Management (A) - **Ranin.Nseir@york.ca**

Faruque Mia, Infrastructure Engineer - **Faruque.Mia@york.ca**

Sophie Xu, Project Coordinator - **Sophie.Xu@york.ca**

Cassie Sining Liu, Project Manager, Inflow and Infiltration Strategy and Analysis (A) - **CassieSining.Liu@york.ca**

Peter Chu Su, Data Scientist - **Peter.ChuSu@york.ca**

Ibraheem Nuaaman, Data Analyst - **Ibraheem.Nuaaman@york.ca**

york.ca/landl