https://doi.org/10.71573/vk078r78

© Authors. This work is licensed under a Creative Commons Attribution 4.0 International License

Development of a generic machine learning model for flowrate generation in catchments using a global database

Karim Claudio[1], Thibaud Maruejouls[1], Marcello Serrao[2]* https://orcid.org/0009-0001-9798-4223,
Abdelghani Zaid[1], Philippe Ginestet[2] & Wolfgang Rauch[3] https://orcid.org/0000-0002-6462-2832

[1] *SUEZ Innovation, Le LyRE R&D Center, 33600 Pessac, France*
[2] *SUEZ International, Engineering & Construction – Innovation & Technical Office, 92040 Paris La Défense, France*
[3] *Universität Innsbruck, Department of Urban Drainage Modelling, 6020 Innsbruck, Austria*
**Corresponding author email: karim.claudio@suez.com*

Abstract

This study introduces a generic data-driven model for urban hydraulic modelling, designed to estimate flow rates in catchment areas. The model employs a machine learning architecture trained on historical observational data from 30 Water Resource Recovery Facility (WRRF) sites across France, covering diverse hydraulic and environmental conditions. A generic data-driven model integrates principles of machine learning with extensive, varied datasets to create adaptable tools that generalize across different locations and conditions without requiring site-specific recalibration. In this case, the model predicts peak flow curves, enabling the estimation of peak and nominal flow rates over time intervals ranging from 1 hour to 6 months. The database was constructed from observational data measured in the sewage networks and at the inlets of wastewater treatment plants managed by Suez Eau France, supplemented with data on land cover, soil type and rainfall from Open Data sources. The model's effectiveness was evaluated on a pilot site, validating its versatility. This tool serves as a valuable decision-support resource for engineers and consultants in urban water management. By leveraging machine learning and a robust, diverse dataset, this approach enhances reliability, adaptability, and efficiency in addressing complex urban hydraulic challenges.

Key words:

urban drainage modelling, generic models, machine learning, flowrates prediction

Highlights

- Prediction of flowrate components using combined generic data-driven models.
- Machine Learning model trained and tested on datasets from 7 sites covering 3 years.
- Models fed from open data sources and GIS data.

Introduction

Hydraulics models are used for decades to better manage Urban Drainage (UD) systems to anticipate floodings, understand Combined Sewer Overflows (CSO) and estimate the inflow to treatment plants. As these models are constantly improved as phenomenon knowledge increases, they are now common in large cities. Their performance is strongly linked to their detail level, i.e. their degree of complexity which has a direct impact on the computational requirement. The recent infatuation in data-driven models have generated many studies exploring the capacity of these models to forecast hydraulics in urban drainage, highlighting the important simulation velocity capacities, even for wide UD systems.

A continuously increasing number of influent generators based on data-driven models is proposed in the academic research. These models were developed to reproduce either hydraulics during floodings, or at Combined Sewer Overflows (CSO), or WRRF inlet, but also pollutograms under dry, wet weather or even snow melting (Sitzenfrei et al., 2015; Li and Vanrolleghem, 2022). Data-driven modelling has been coupled with physical-based models to take advantage of complex processes understanding of first models and large dataset analysis capacity of the second one (Schneider et al., 2022). Despite the increasing interest in data-driven models, most of the studies build, train, and evaluate models on the same urban drainage systems, making by nature non-generic models for different sites.

This study presents the development of an algorithm aiming at generating UD models for almost any part of the world with or without GIS support. This algorithm relies mainly on open-source external datasets for rainfall, topography and soil occupancy. Models, based on machine learning approach, have been developed and the results generated were compared to observed data to assess their performance.

Methodology

The methodology applied in this project consists in a machine learning model trained from a global database, used to estimate 3 components of the flow at the entrance of the wastewater treatment plant: the infiltration flow, the wastewater flow (or dry weather flow) and the stormwater flow (or wet weather flow). Concretely, the generic model consists in three models, one for each component of the flow. Each model uses a specific set of explanatory variables (called parameters).

Data Collection & Quality

All data were collected at the catchment level. For each catchment, local data were acquired to enrich the database and train the model. Different data sources were involved in this process, among them:

- GIS data, specifically data on network length (including percentage of combined system),
- Terrain data, to collect terrain slope data, from US National Science Foundation,
- Soil imperviousness, from the European Space Agency,
- Hourly rainfall data, from the European Union's Space program,
- Population data, from the French National Institute of Statistics and Economic Studies (INSEE).

Finally, to train the generic models, flow data have been compiled for the pilot sites. In each case, the wastewater flow has been disaggregated into 3 components (EPA, 2008): the Infiltration water (IW), the Wastewater flow (WWF) and the Stormwater flow (SWF). This information is only required for model training and testing, not for its application.

Model identification & evaluation.

Once all data are collected the three models described previously must be calibrated. The models selected are regression random forest model (Breiman, 2001). A machine learning model has been preferred to stochastic approaches (e.g. time series models), because of its flexibility and easiness for generalization.

The calibration process of a regression random forest first consists in determining the hyperparameters of the model (i.e., its configuration) and then to design the different trees based on these hyperparameters and parameters (variables) defined. The hyperparameters are fine-tuned (Probst et al., 2019) by finding the optimal set that will maximize the Nash-Sutcliffe model efficiency coefficient (Gupta et al., 2009) where the model is designed on a training period and assessed on a test dataset. Each model is trained on a set of N catchments, and will be applied to a $N+1^{th}$, not use for training, to assess the replicability of the model on new site.

Application

The methodology previously described was designed and tested on pilot sites from 7 contracts in France, representing 30 catchments areas. In total, this database represents over 4 685km of wastewater network. Data from inflow water were collected between January 2022 and December 2024, however, as each network is operated independently, the history depth can differ from one contract to another. The different networks selected provide a diversity in configurations, with small/medium networks (< 500 km) and large networks (> 1000km). In consequence, the associated flows diverge in order or magnitude from one site to another, with some cases with flow centred around 30 m^3/h, while other cases have an average inflow around 400 or 800 m^3/h. As the model is meant to be applied on all kinds of network, it has been decided to normalize some outputs to predict: the infiltration water is divided by the total network length and the wastewater flow is divided by the population size.

The model is then tested on a totally new catchment, with a network of 1066 km of drainage system for almost 200 000 inhabitants. The results of the model are compared to the observed flow at the entrance of the wastewater plant, between July 2024 to February 2025.

Results and discussion

A random forest model has then been applied to model each one of the three flow components (Infiltration Water, Wastewater flow and Stormwater flow). The first model (infiltration water) relies on the following set of parameters: population size, network length, percentage of combined system, catchment area, catchment slope, catchment imperviousness, cumulative and maximal rainfall over the last one week, two weeks and one month. An example of results obtained for this model is displayed in Figure 1. As shown in the figure, the model was able to catch up the trend of the IW, despite not being trained on this pilot site.

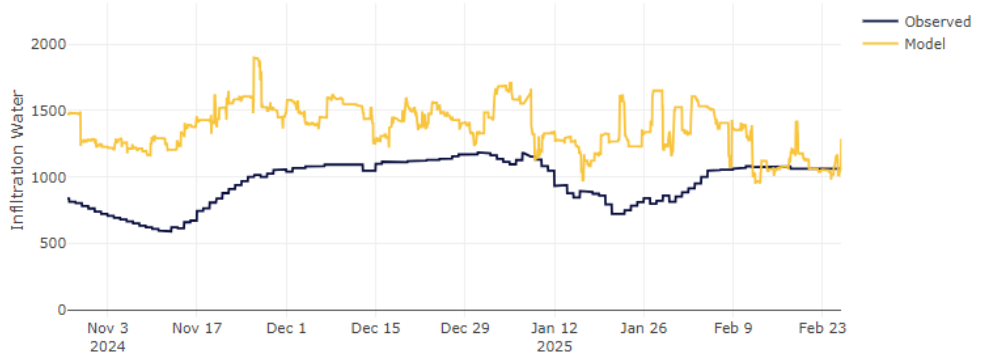

Figure 1. Observed and modelled infiltration water for one catchment of pilot site 1.

This process is repeated for the second model (wastewater flow) which includes the following features: population size, network length, percentage of combined system, catchment area, catchment slope, catchment imperviousness, month of the year, day of the week and hour of the day. Results for this model are displayed in Figure 2.

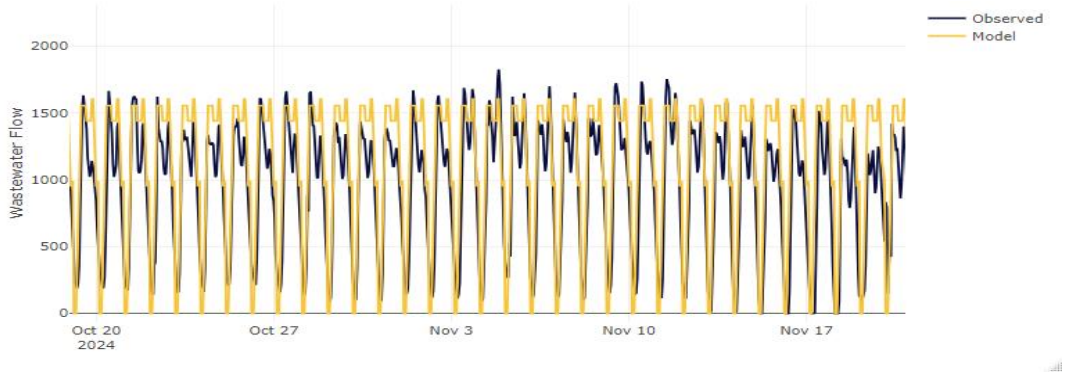

Figure 2. Observed and modelled wastewater flow for the validation catchment

Finally, for the last model (stormwater flow), the selected features are population size, network length, percentage of combined system, catchment area, catchment slope, catchment imperviousness, cumulative and maximal rainfall over the last 15 minutes, 30 minutes, 1 hour, 2 hours and 4 hours. Results for this model are displayed in Figure 3.

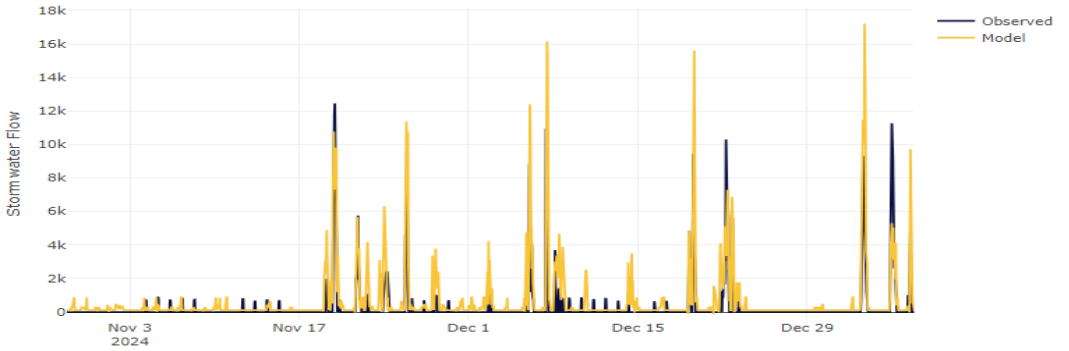

Figure 3. Observed and modelled stormwater flow for the validation catchment

Conclusions and future work

A new method aiming at generating urban drainage models in SWMM at global earth scale using poor a priori knowledge has been developed and presented. It relies on large open-data sources for weather, hydrology, and hydraulics that were selected for their global earth coverage and (temporal/spatial) resolution fitting with flowrate modelling purposes. Results show model performances good enough to be used for several use cases such as long-term scenario simulations or yearly average urban drainage outlet flowrate prediction.

Several challenges emerged during the study, including general poor GIS quality that needs robust algorithms to overcome the data gaps, inflows and infiltration processes that can vary strongly from one year to the other which depends on multiple factors. The authors are working on overcoming this limitation by providing a stochastic approach to cover certain parameter ranges for the one poorly known.

Finally, the results of the model will be used to simulate peak and nominal flow, by aggregating flow at different time scales (from hours to months) and extracting different indicators (median, 99th percentile).

Acknowledgement

We thank the operators of participating wastewater treatment plants for allowing us access to data.

References

Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5–32.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009) Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.* **377**, 80–91.

Li W. and P.A. Vanrolleghem (2022) An influent generator for WRRF design and operation based on a recurrent neural network with multi-objective optimization using a genetic algorithm. Wat. Sci. Tech., **85**(5), 1444-1452.

Probst, P., Wright, M.N and Boulesteix, A.L. (2019) Hyperparameters and tuning strategies for random forest. *WIREs Data Mining Knowl. Discov.* **9**(3).

Sitzenfrei R., Hillebrand S. and W.Rauch (2015) Development and application of a stochastic waste water production model for households to investigate heat recovery scenarios in sewers. *10th International Urban Drainage Modelling Conference*, 20-23 September, Mont Saint-Anne, Canada.

Schneider M.Y., Quaghebeur W., Borzooei S., Froemelt A., Li F., Saagi R., Wade M.J., Zhu J-J. and E.Torfs (2022) Hybrid modelling of water resource recovery facilities: status and opportunities. *Wat Sci Tech.*, **85**(9), 2503-2524.

U.S. Environmental Protection Agency (EPA). (2008). "Review of Sewer Design Criteria and RDII Prediction Methods." Report No. EPA/600/R-08/010, EPA, Washington, D.C.