https://doi.org/10.71573/wsg46f90

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

A probabilistic framework for urban wastewater flow forecasting

Mohsen Rezaee [1, 2, *] https://orcid.org/0000-0001-6156-1520,
Peter Melville-Shreeve [1, 2] https://orcid.org/0000-0001-9583-8006
& Hussein Rappel [1] https://orcid.org/0000-0003-0982-6733

[1] *Department of engineering, Faculty of Environment, Science and Economy, University of Exeter, Exeter, UK.*
[2] *Centre for Water Systems, University of Exeter, Exeter, UK.*
**Corresponding author email: m.rezaee@exeter.ac.uk*

Abstract

Sewer flow forecasting is critical for managing the performance of sewer networks and their treatment plants. While simulators have been used in modelling the sewer flow for years, data-driven emulators recently have gained attention in making predictions with a higher computational speed and feasibility. In this research, a framework is proposed based on multi-input single-output Gaussian Processes for predicting sewer flow using time and rainfall as inputs. The predictions are presented as Gaussian distributions, showing the confidence levels. The results of the GPR on the data of a sewer system in this study demonstrated a robust performance of the model with 93.6% coverage of the predictions in the 95% credible interval, and 89.5 L/s of RMSE.

Highlights

- A probabilistic model has been developed for forecasting sewer flow
- Multi-input GPR can predict sewer flow for short- and long-term periods showing uncertainties
- A complex kernel should be implemented to take the changes in the flow

Introduction

Flow management in sewer systems is crucial as poor operation can lead to wastewater surcharges and unexpected high flows that significantly reduce the efficiency of the wastewater treatment plant (WWTP) and damage pumping stations (Karimi et al., 2019). The world is shifting from dumb and passive to smart and monitored wastewater management, and real-time control of wastewater systems is now on the horizon (Sweetapple et al., 2023). Two primary modelling approaches exist for predicting flow in a sewer system: physical simulators that numerically solve hydraulic equations and data-driven emulators that learn from complex datasets (Troutman et al., 2017).

Sewer flow is influenced by various factors, including water usage, groundwater level, precipitation, and sewer conditions (Zhang et al., 2019). These interconnected factors are sources of uncertainty when predicting the flow in a wastewater network. Therefore, deterministic sewer flow forecasting approaches often contain considerable errors due to the probabilistic nature of the input factors and inherent uncertainties arising from parameters and observations (Honti et al., 2013).

Among various AI models that can be used in flow forecasting, the Gaussian processes (GP) method provides a useful approach by containing uncertainty quantification in predictions and taking complex datasets through its natural Bayesian interpretation (Ding et al., 2023). This method has been implemented in some water-related studies like urban water demand forecasting (Wang et al., 2014) or streamflow predictions (Pastrana-Cortés et al., 2024).

In this research, we explore the capability of GP regression in forecasting flows in a sewer system. Then, the model will be evaluated by some model checking metrics.

Methodology

Data

The data is derived from the SWMM simulator engine using the Python interface PySWMM (McDonnell et al., 2020) with the input of a synthetic combined sewer network serving 11,300 inhabitants. The WWTP receives 18.31 L/s of wastewater and 4.58 L/s of groundwater infiltration on average in dry-weather conditions. Precipitation data, totalling 644.4 mm, was obtained from a meteorological station in the UK and has an hourly resolution. Both the simulation and precipitation datasets cover a one-year period.

The output from the SWMM simulation is used to construct the training and test datasets of the data-driven model. To make the study more realistic, a Gaussian noise value (maximum 10% of the average flow value in dry-weather days) is added to all synthetic data taken from the simulator.

Forecasting method

Forecasting flow as a time-series can be done by a Gaussian process regression (GPR) model. In GPR, a GP is set as a prior for the latent function ($f(\mathbf{x})$) that maps the inputs into the output space and then updates the prior with the observations, using Bayes' rule (MacKay, 2003). The key predictive equations for GPR are as follows (Rasmussen and Williams, 2006):

$$\bar{\mathbf{f}}_* = K(X_*, X)[K(X, X) + \sigma_n^2 I\,]^{-1} y \quad (1)$$

$$cov(\mathbf{f}_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*) \quad (2)$$

where, $\mathbf{f}_*$ is the predicted test output according to the prior and the $\bar{\mathbf{f}}_*$ is the mean of the prediction on the test points, X_*. In Eq. 1, **y** = **f**(x) + Ɛ where Ɛ is the Gaussian noise with variance σ_n^2. Furthermore, $K(X_*, X)$ denotes the covariance matrix evaluated at all training (X) and test (X_*) points, and the *cov*($\mathbf{f}_*$) shows the covariance of the prediction function. To optimize the GPR, the best hyperparameters are found by maximizing the log marginal likelihood.

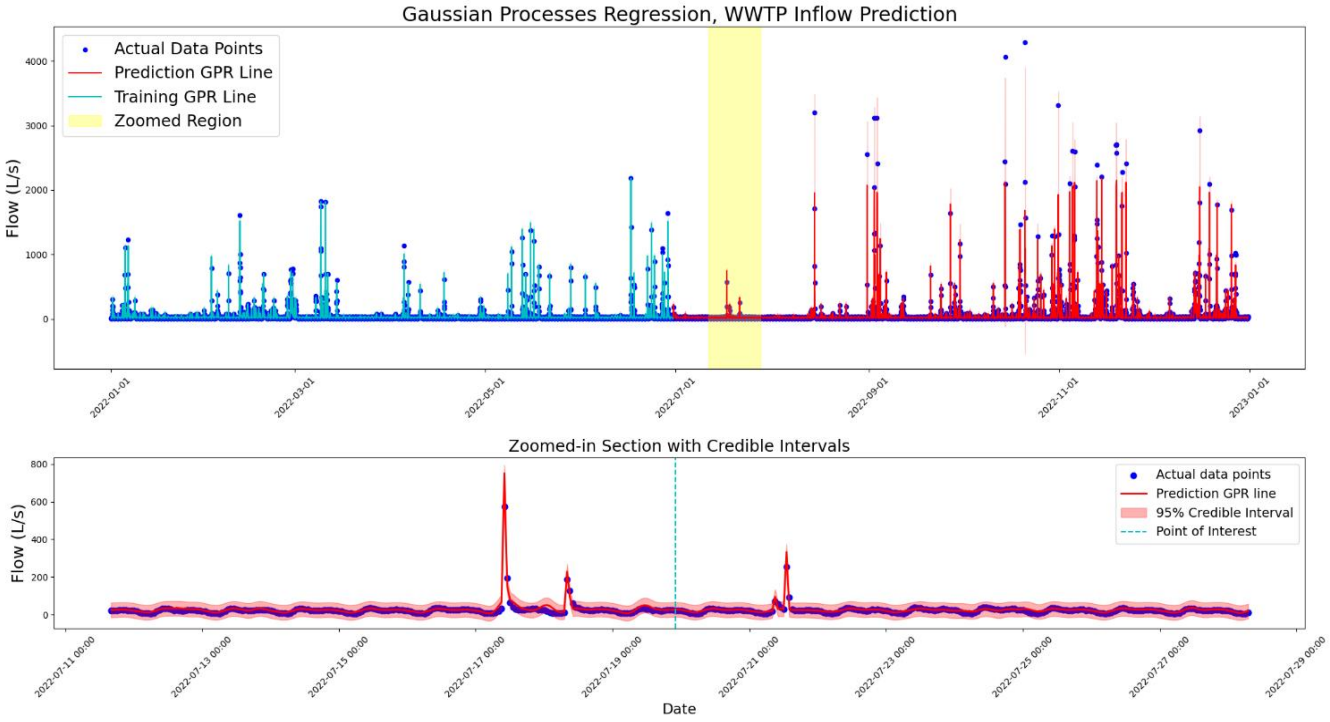
The GPR model takes time and rainfall depth as inputs and predicts the flow in a particular point as the output; therefore, we have a multi-input single-output model. With the independent inputs, different kernel functions should be used to capture the behaviours of the flow based on each input. Consequently, a complex kernel should be defined and used for the modelling.

Model checking

Various techniques exist for assessing the fit of a model including graphical and statistical tests. Among these checks, coverage measure and root mean square error (RMSE) are employed in this work. The first check calculates the percentage of the data falling within the defined credible interval. A good model is defined as a model having α% coverage for an α% credible interval. The latter, on the other hand, provides an absolute error metric in the same units as the data, allowing for a straightforward interpretation of prediction error.

Results and discussion

Based on the capabilities of the model, many parameters like flow, water depth and velocity can be used as the training dataset of the data-driven model. The inflow of the WWTP is a critical metric for sewer system management, as it significantly influences the performance of the treatment plant. Therefore, a 6-month dataset was used to train the model and capture medium-term flow patterns effectively.



Since time and rainfall depth have the most impact on the wastewater flow value, we used these two as inputs of the GPR model. These two parameters can be considered independent based on their characteristics, thus the GPR model can have different kernels acting on each input.

Figure 1. Gaussian process regression model on 6 months of training and 6 months of test data. The regression lines are shown in blue and red for training and test sets. A part of the test dataset is selected for a detailed output which is shown in the subplot with 95% credible intervals around the GPR line.

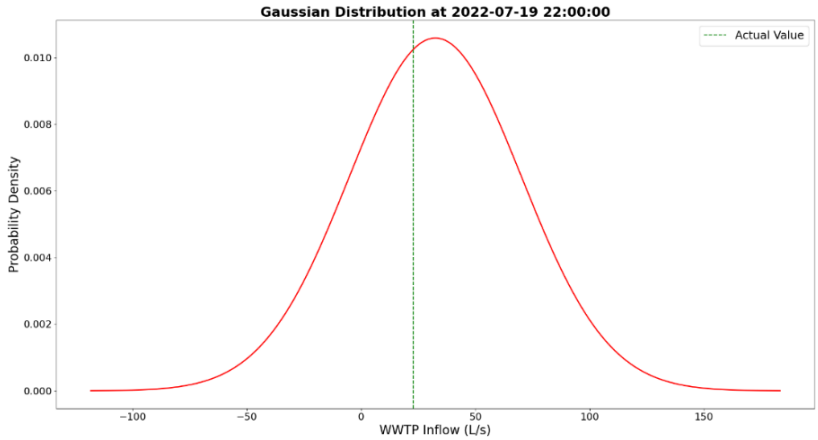


Figure 2. Gaussian distribution of the predicted WWTP inflow at 2022-07-19 at 10pm. The green dashed line shows the test point value. The distribution's domain is ±4 standard deviation of the distribution.

The covariance function of this GPR model is comprised of a time kernel that includes two Periodic kernels, taking the hourly and daily variations of the wastewater flow, and the rain kernel including a Matern12 kernel which is designed to take spike changes in the data, similar to what happens to the wastewater flow during rainfalls. For a detailed discussion on kernels, see Duvenaud (2014).

The modelling was implemented using the GPflow package in Python (Matthews et al., 2017). The results of 6 months of flow prediction and the specified area of the test set are shown in Figure 1. Moreover, to show the format of outputs of a GPR model in each prediction point, Figure 2 illustrates the predicted distribution of wastewater flow for a randomly selected point.

As the figures indicate, all the predictions are in the form of Gaussian distributions that show the level of uncertainty. The mean value of the predicted distribution shown in Figure 2 is near the actual value of that test point, so the proposed framework seems capable of predicting observations with high plausibility. However, the model checks mentioned before can talk more about the goodness-of-fit. The coverage of the test data is 93.6% for the 95% credible interval which is satisfactory. Also, the RMSE of this model is 89.5 L/s which is good considering WWTP inflow values of more than 2,000 L/s. Some outliers can be seen in Figure 1, particularly when there is a rainfall event and consequently a high flow. But it is important to note that the training set is placed in the first half of the year with less rainfall compared to the second half. Therefore, another model is built with the inversed training and test sets. The results from this model indicate a coverage of 97.2% and an RMSE of 37.4 L/s which are significantly better than the first modelling results. Consequently, it can be inferred that the GPR model performs better when trained on data including a wider range of conditions, such as more wet-weather days and higher peak flows.

In essence, the model's reliability improves when the training set covers values spanning from very small to very large. However, when the prediction point exceeds the range of the trained values, the model adopts a conservative approach, gravitating towards predictions closer to the mean line.

Given GPR's effectiveness with minimal datasets, a short-term training was conducted using 7 days of WWTP inflow data and tested over the following 3 days. The results, illustrated in Figure 3, show an RMSE of 2.5 L/s and 100% of coverage which are significantly better compared to the previous predictions. The reason behind this enhancement in metrics is the shorter length of test period and the strength of GPs on limited datasets.

The results of this modelling show potential for assisting decision-makers in planning the management of sewer systems more effectively. They can also be used to alert stakeholders about impending high flows and potential surcharges whilst providing actionable insights with a quantifiable level of confidence. For example, if an inflow threshold is defined for the treatment plant, the GPR model's predictions can indicate the probability of exceeding that threshold at each timestep.

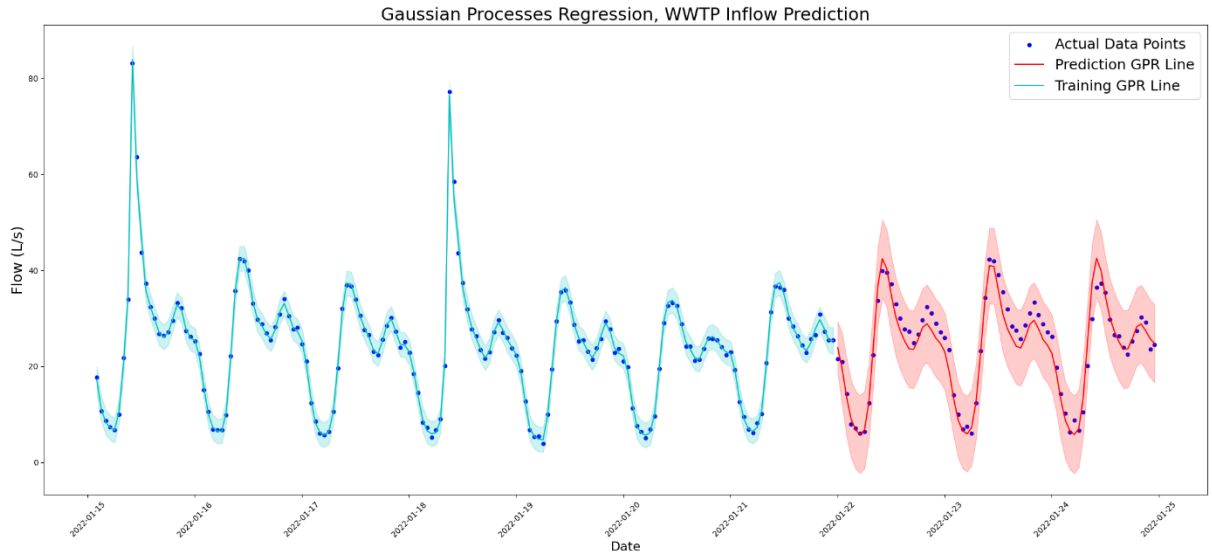


Figure 3. Gaussian process regression model on 7 days of training and 3 days of test dat. The GPR mean and the 95% credible interval in the training period are shown in blue. The data points are also shown as blue bullet points. The prediction and its corresponding credible interval are shown in red.

Conclusions and future work

This research demonstrated the potential of a Gaussian process regression (GPR) model to be used in forecasting flow in a sewer system. Since the flow prediction in sewers contains various uncertainties from data collection to modelling, the probabilistic approach of GPR delivers results with quantifiable

levels of confidence. This GPR model is one of the data-driven modelling methods that stand out for being computationally efficient compared to traditional simulators.

The results demonstrate that the GPR model performs well in short- and long-term predictions with limited training data. A six-month training period enables accurate predictions over the next six months, with satisfactory coverage and error metrics. Moreover, training the model on a more diverse dataset, including higher flows during wet-weather conditions, significantly improves its predictive accuracy and reliability.

Future developments could enhance the model by incorporating physical constraints from hydraulic equations, creating a physics-constrained emulation. This refinement would bridge the gap between data-driven and physics-based approaches providing more reliable forecasting in wastewater flow—a goal the authors aim to pursue in their future research.

References

Ding, C., Rappel, H., & Dodwell, T. (2023). Full-field order-reduced Gaussian Process emulators for nonlinear probabilistic mechanics. Computer Methods in Applied Mechanics and Engineering, 405, 115855. https://doi.org/10.1016/j.cma.2022.115855

Duvenaud, D. (2014). *Automatic model construction with Gaussian processes* (Doctoral dissertation). Available at: https://www.repository.cam.ac.uk/handle/1810/247281

Honti, M., Stamm, C., & Reichert, P. (2013). Integrated uncertainty assessment of discharge predictions with a statistical error model. Water Resources Research, 49(8), 4866-4884. https://doi.org/10.1002/wrcr.20374

Karimi, H. S., Natarajan, B., Ramsey, C. L., Henson, J., Tedder, J. L., & Kemper, E. (2019). Comparison of learning-based wastewater flow prediction methodologies for smart sewer management. *Journal of Hydrology*, *577*, 123977. https://doi.org/10.1016/j.jhydrol.2019.123977

Matthews, A. G. D. G., Van Der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., Le, P., Ghahramani, Z., & Hensman, J. (2017). GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, *18*(40), 1-6. Available at: https://www.jmlr.org/papers/v18/16-537.html

McDonnell, B. E., Ratliff, K., Tryby, M. E., Wu, J. J. X., & Mullapudi, A. (2020). PySWMM: the python interface to stormwater management model (SWMM). *Journal of open source software*, *5*(52), 1. https://doi.org/10.21105/joss.02292

MacKay, D. J. (2003). Information theory, inference and learning algorithms. Cambridge university press.

Pastrana-Cortés, J. D., Gil-Gonzalez, J., Álvarez-Meza, A. M., Cárdenas-Peña, D. A., & Orozco-Gutiérrez, Á. A. (2024). Scalable and Interpretable Forecasting of Hydrological Time Series Based on Variational Gaussian Processes. *Water*, *16*(14), 2006. https://doi.org/10.3390/w16142006

Rasmussen, C. E. & Williams, C. K., (2006). Gaussian processes for machine learning, Cambridge, MA: MIT press. ISBN 0-262-18253-X.

Sweetapple, C., Webber, J., Hastings, A., & Melville-Shreeve, P. (2023). Realising smarter stormwater management: A review of the barriers and a roadmap for real world application. *Water Research*, *244*, 120505. https://doi.org/10.1016/j.watres.2023.120505

Troutman, S. C., Schambach, N., Love, N. G., & Kerkez, B. (2017). An automated toolchain for the data-driven and dynamical modeling of combined sewer systems. Water Research, 126, 88-100. https://doi.org/10.1016/j.watres.2017.08.065

Wang, Y., Ocampo-Martínez, C., Puig, V., & Quevedo, J. (2014, October). Gaussian-process-based demand forecasting for predictive control of drinking water networks. In *International Conference on Critical Information Infrastructures Security* (pp. 69-80). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-31664-2_8

Zhang, Q., Li, Z., Snowling, S., Siam, A., & El-Dakhakhni, W. (2019). Predictive models for wastewater flow forecasting based on time series analysis and artificial neural network. *Water Science and Technology*, *80*(2), 243-253. https://doi.org/10.2166/wst.2019.263