Probabilistic Forecasting and Anomaly Detection in Sewer Systems Using Gaussian Processes

Mohsen Rezaee , Peter Melville-Shreeve , Hussein Rappel *

Posted Date: 20 May 2025

doi: 10.20944/preprints202505.1621.v1

Keywords: combined sewer system; urban drainage; Gaussian processes regression; blockage detection; data-driven forecasting; water level prediction

 doi:10.20944/preprints202505.1621.v1

Article

Probabilistic Forecasting and Anomaly Detection in Sewer Systems Using Gaussian Processes

Mohsen Rezaee [1,2], Peter Melville-Shreeve [1,2] and Hussein Rappel [1,*]

[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
* Correspondence: h.rappel@exeter.ac.uk

Abstract: This study investigates the capability of Gaussian process regression (GPR) models in probabilistic forecasting of water flow and depth in a combined sewer system. Traditionally, deterministic methods have been implemented in sewer flow forecasting and anomaly detection which are two crucial techniques for a good wastewater network and treatment plant management. However, with the uncertain nature of the factors impacting on sewer flow and depth, a probabilistic approach which takes uncertainties into account is preferred. Using hydraulic simulation data generated in this study, a composite kernel is designed to take the flow and depth patterns based on the flow characteristics, and the hyperparameters are optimised through maximisation of the log-likelihood function. The GPR model is a multi-input, single-output model taking time and precipitation as inputs. Prediction results show reliable model outputs and are evaluated by three metrics: root mean square error, coverage and differential entropy. The model effectively predicts treatment plant inflow and manhole water levels with different training periods. Finally, the model is used for anomaly detection by identifying deviations from expected ranges, enabling the estimation of surcharge and overflow probabilities under various conditions.

Keywords: combined sewer system; urban drainage; Gaussian processes regression; blockage detection; data-driven forecasting; water level prediction

1. Introduction

Sewer systems convey domestic, industrial and commercial wastewater (and sometimes stormwater and groundwater) from their sources to a wastewater treatment plant (WWTP) or a discharge point. These systems can be mainly divided into combined and separate sewer systems [1]. Combined sewer systems are built to take both wastewater and stormwater from different sources and convey them to the final point of the system. Such systems are prevalent in the UK with Victorian era infrastructure deployed in the form of combined drainage systems in many urban areas.

Wastewater flow contains various types of pollutants which should be effectively treated to provide a safe environment. However, for several reasons, there are untreated releases in almost every sewer system which are detrimental to the environment and health [2]. The most important case in untreated wastewater releases is the combined sewer overflow (CSO) which is a structure, installed to spill wastewater into the environment to reduce the risk of backup discharges and manhole overflows [3]. Even though CSOs minimise the risk of manhole overflows, blockages or pump failures may still lead to unintentional discharges [4,5].

Given the adverse effects that CSOs cause, sewer systems should be designed and maintained in a way that overflows are reduced to a minimum. Moreover, water companies risk paying penalties due to regulatory requirements set by the Water Service Regulation Authority (OFWAT) in the UK (and similar authorities in other countries). Therefore, water companies seek methods and models to predict the flow and be able to forecast overflows in the system because proactive responses are more reliable than reactive responses in a sewer system [6].

 doi:10.20944/preprints202505.1621.v1

There are three types of models that can be used in forecasting overflows and other characteristics of a hydraulic or hydrologic environment: simulators [7], data-driven surrogate models [8] and hybrid models [9]. The first models, also known as physical models, have been used for more than five decades; however, considered computationally expensive for years [10,11]. These models simulate the underlying physics of the environment and represent it as a set of partial differential equations (PDEs) [12].

The mathematical modelling of every physical system often requires tremendous simplifications that lead to significant uncertainties in the computation. Additionally, high computational cost, demand for updating model parameters in response to urban developments, and lack of knowledge in every physical behaviour of the system are known as drawbacks of these models [13,14]. Confidence in input parameters like the network topology, invert levels and current levels of service (for example root ingress, sediment build up, etc.) creates uncertainty in such models which can typically be improved through calibration approaches using short term flow surveys at key nodes in the network to some extent.

On the other hand, data-driven models are capable of making predictions with greater computational efficiency. These models present an opportunity to capture relationships between parameters without considering interactions [15]. Nonetheless, data-driven modelling suffers from a lack of physical interpretation which may make it unreliable or challenging to interpret [16]. Therefore, some researchers have come to a mixture of data-driven and physical models, called physics-informed surrogate models, taking advantage of both physical and data-driven models [17]. While these models benefit from physical constraints, development and integrity challenges along with added complexity make them challenging to implement in this study and similar works.

Data-driven models range from simple linear regressions to complex deep-learning architectures, each varying in computational complexity. These models can be classified into deterministic and probabilistic approaches based on the way they handle uncertainty in the modelling process. While deterministic models offer an explicit solution to a water-related issue, probabilistic models take the inherent uncertainties of the data and the model parameters into account.

In sewer system modelling, uncertainties arise from multiple sources, including variability in precipitation and water usage, simplifications in physical equations, and sensor inaccuracies [18]—all of which can significantly impact the reliability of predictions. While most research in uncertainty quantification has focused on water quality and treatment plant operations, some researchers have addressed flow-related uncertainty. For example, Breinholt [19] assessed the uncertainty in flow prediction and modelling in a sewer system with a frequentist stochastic approach. Raimondi et al. [20] adopted a deterministic framework to quantify uncertainty in flow rate and temperature, while Sriwastava et al. [21] employed stochastic sampling to evaluate CSO volume variability.

While many data-driven models, like artificial neural networks (ANNs) and support vector machines (SVMs), require large datasets and lack uncertainty quantification in predictions, the Gaussian processes (GPs) method provides an approach with an ability to take complex and small datasets through a natural Bayesian interpretation [22]. This method makes predictions incorporating expert's prior knowledge with a model (called likelihood i.e. a function showing the probability of data happening given a parameter) and provides uncertainty measures in its predictions [23]. Moreover, GP is known as a suitable method for working with rare and costly data, while it can provide good results with larger datasets with an $O(N^3)$ computational complexity [16].

Gaussian process regression (GPR) has been implemented in many different fields of study due to its capability in uncertainty quantification and flexibility (i.e., the ability of emulating different function shapes). This capability has led GPR models to be used in engineering and scientific fields as surrogate models, such as in computational engineering [24], climate science [12], finance [25] and many time-series related forecasting models [26].

In water engineering, GP-based frameworks have been used in forecasting streamflow [27,28], fault detection in wastewater treatment plants [29], evaluation of the flooded nodes in a sewer system

 doi:10.20944/preprints202505.1621.v1

[30], drinking water demand prediction [31], streamflow temperature forecasting [32], water level fluctuations forecasting [33] and approximating parameters of a storm water management model (SWMM) [13].

This research aims to develop a framework for forecasting flow and depth in different parts of sewer systems. To this end, the framework deploys a GPR model with a custom-designed kernel to capture various patterns affecting flow and depth changes. This probabilistic model aids water companies make short- and mid-term asset management plans and respond proactively to potential issues in the network and treatment facilities.

In addition to what data-driven, and particularly GPR, models can forecast in wastewater systems, they offer valuable applications in real-time control and nowcasting. With abundant sensors placed in sewer systems, smart and monitored wastewater management is on the horizon [34]. One critical challenge in sewer systems is anomaly detection [35]. When a pump failure occurs or a blockage develops in a pipe, the water flow and depth suddenly decrease downstream while increasing upstream. The timely prediction of these anomalies plays an important role in urban water management [36] and reduces costs associated with pollution events driven by such issues [37].

Data-driven methods provide a more cost-effective alternative to hardware-based techniques (e.g. CCTV, acoustic sensing, etc.) [36,38]. Among them, probabilistic methods are credited for their ability to take complexities into account [39]. Nonetheless, there is no systematic research using GPR for detecting anomalies in a sewer system.

Accordingly, this research proposes a novel probabilistic method for anomaly detection using a GPR model. The method estimates the likelihood of blockage occurrences in real time, and raises alarms when discrepancies exceed predefined thresholds.

The remainder of this paper is organized as follows. Section 2 discusses the data and the hydraulic model used to prepare the training and test datasets. It also outlines the GPR model structure, its implementation within the framework, and the evaluation metrics applied to assess its performance.

Section 3 presents the results taken from the GPR model along with a discussion. It shows how the GPR model works with long- and short-term training periods and how it can be used in presenting probabilistic results for the stakeholders. Finally, the paper concludes with remarks and suggestions for future work.

2. Materials and Methods

2.1. Data and the Case Study

The case study of this research is a hypothetical combined sewer system built by EPA SWMM v5.2 software [40]. The system is a skeleton sewer network serving 11,300 inhabitants of a city, assumed to be placed in the UK. The average water consumption is assumed to be 140 LPCD with a 100% wastewater generation rate. A snapshot of the system and the details are presented in Figure 1 and Table 1.

A CSO is set in the network to release excess flows via a weir. There is also a storage tank and a pumping station set to regulate the flow in the network. The mains and trunk sewers have been designed based on the IDF curves of London [41] and have sizes between 500 mm to 1200 mm.

Groundwater infiltration can have a major contribution to the flow of a sewer system which makes it important when designing a sewer network and the treatment plant [42]. The seasonal effect of such can be complicated and driven by geological conditions such as soil porosity and bedrock characteristics. In this case study, the groundwater infiltration is set as 25 per cent of the dry-weather flow (DWF) which can be considered a normal value for a mid-life network based on recent research papers [43]. The values of the infiltration are presented in Table 1. The sewer exfiltration is assumed to have negligible amounts in the network and has not been included in the calculations.

 doi:10.20944/preprints202505.1621.v1

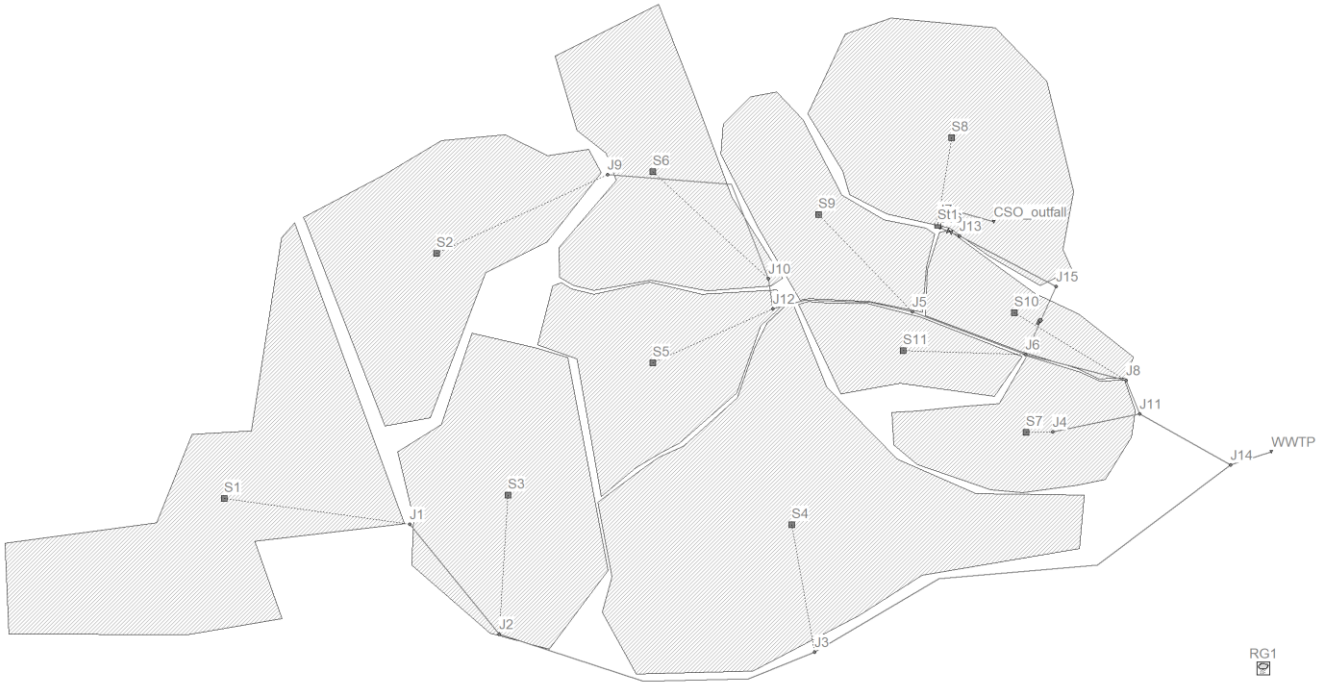

Figure 1. Skeleton sewer system terminating at the WWTP on the right side of the figure. The hatched areas show the sub-catchments, the links show the main sewers, and the square points are the Nodes labelled in the figure. The rain gauge (RG1) is shown at the right bottom of the figure.

Table 1. Details of the wastewater production values and infiltration in each sub-catchment.

Sub-catchment	Terminates at Joint	Population	Wastewater production (L/s)	Infiltration (L/s)
S1	J1	1200	1.94	0.49
S2	J9	700	1.13	0.28
S3	J2	700	1.13	0.28
S4	J3	1200	1.94	0.49
S5	J12	1300	2.11	0.53
S6	J10	1200	1.94	0.49
S7	J4	900	1.46	0.36
S8	J7	1200	1.94	0.49
S9	J5	1000	1.62	0.41
S10	J8	1000	1.62	0.41
S11	J6	900	1.46	0.36
Total	**WWTP**	**11300**	**18.31**	**4.58**

Different time patterns have been synthetised to approximate the model to a real-world case. These include daily, weekday, weekend and monthly water usage patterns, all constructed to reflect household wastewater production patterns. Moreover, a pattern is made for one of the sub-catchments showing that the water usage in the area is affected by a nearby university, and a groundwater infiltration pattern is constructed based on the UK groundwater fluctuation maps.

Finally, the hydraulic model is built in EPA SWMM, and the hydraulic simulation data are generated using PySWMM (the Python interface to SWMM) [44], allowing the authors to integrate the simulator and data-driven model within a single platform.

While rainfall data could have been synthetically generated at a specified resolution, this study uses open-access precipitation data from the Sheffield meteorological station [45]. With a 5-minute

 doi:10.20944/preprints202505.1621.v1

resolution and 0.2 mm accuracy collected with an ARG-100 tipping bucket rain gauge, this dataset is ideally suited for the objectives of the research.

2.2. Gaussian Process Regression

A GP is a collection of random variables, any finite number of which have a joint multivariate Gaussian density function [46]. GP is completely specified by its mean and covariance functions [47]. Mean (m($\boldsymbol{x}$) i.e. the central value of the set $\boldsymbol{x}$) and covariance (k($\boldsymbol{x},\boldsymbol{x}'$) i.e. the measure indicating how two random variables depend on each other) functions of a real process, $f(\boldsymbol{x})$ (i.e. a function from real inputs to random outputs), are defined as:

$$m(\boldsymbol{x}) = \mathbb{E}\,(f(\boldsymbol{x})) \tag{1}$$

$$k(\boldsymbol{x},\boldsymbol{x}') = \mathbb{E}\,((f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x}') - m(\boldsymbol{x}'))). \tag{2}$$

Using the GP, we can define a distribution over functions $f(\boldsymbol{x})$,

$$f(\boldsymbol{x}) \sim GP(m(\boldsymbol{x}), k(\boldsymbol{x},\boldsymbol{x}')), \tag{3}$$

where $\boldsymbol{x}$ and $\boldsymbol{x}'$ are two GP input vectors and $\boldsymbol{x} \in \mathbb{R}^N$.

Forecasting flow as a time-series can be done by a GPR. In GPR, a GP is set as a prior for the latent function ($f(\boldsymbol{x})$) that maps the inputs into the output space and then updates the prior with the observations, using the Bayes' rule [48]. In this work, the inputs are time and rainfall depth, and the output can be water flow or depth.

Let $\boldsymbol{X}$ and $\boldsymbol{X}_*$ denote training and test matrixes and $K(\boldsymbol{X}_*,\boldsymbol{X})$ shows the covariance matrix evaluated at all data points, the key predictive equations for GPR when there is noise in observations are as follows [46]:

$$\bar{\boldsymbol{f}}_* = K(\boldsymbol{X}_*,\boldsymbol{X})[K(\boldsymbol{X},\boldsymbol{X}) + \sigma_n^2 \mathrm{I}]^{-1}\boldsymbol{y} \tag{4}$$

$$cov(\boldsymbol{f}_*) = K(\boldsymbol{X}_*,\boldsymbol{X}_*) - K(\boldsymbol{X}_*,\boldsymbol{X})[K(\boldsymbol{X},\boldsymbol{X}) + \sigma_n^2 \mathrm{I}]^{-1}K(\boldsymbol{X},\boldsymbol{X}_*), \tag{5}$$

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, $\boldsymbol{X}_*$. In Equation 4, observation values $\mathbf{y} = \mathbf{f}(\boldsymbol{x}) + \varepsilon$ where ε is the Gaussian noise with variance σ_n^2.

In GPR, the prior mean function is often assumed to be zero, and the posterior mean is updated based on the observed data. To optimise the GPR, best hyperparameters are selected by maximising the log marginal likelihood. Hyperparameters are some free parameters in the covariance function that control how outputs vary with inputs.

$$log\, p(\boldsymbol{y} \mid \boldsymbol{X}) = -\frac{1}{2}\boldsymbol{y}^T(K + \sigma_n^2 \mathrm{I})^{-1}\boldsymbol{y} - \frac{1}{2}log|K + \sigma_n^2 \mathrm{I}| - \frac{n}{2}log\, 2\pi. \tag{6}$$

$|K + \sigma_n^2 I|$ denotes the determinant of the covariance matrix containing noise. This optimisation process allows the model to best fit the observed data while avoiding high complexity of the model. The most common hyperparameters are length scale, output variance and periodicity, found in Squared exponential (SE), Rational quadratic (RQ) and periodic kernels [49]. For further details on advanced kernels, the readers are referred to Wilson [50].

2.3. Constructing Forecasting Model

The hydraulic model data is used as training and testing datasets. To make the model closer to real-world conditions, random Gaussian noise with zero mean is added to the simulated data. This noise is bounded within 3 standard deviations which is set to be equal to 10% of the average dry-weather values.

 doi:10.20944/preprints202505.1621.v1

The GPR model can be trained with a single input (only time) and forecast the flow or depth at future time steps (single-input, single-output model) or incorporate other influencing factors in forecasting (multi-input, single-output model). In this study, time as the primary input in the time-series and the precipitation as the key influencing factor are considered as the inputs of the model. The output of the model is the desired flow or depth.

Once the data from PySWMM are collected and processed, they are imported into a GPR model constructed by the GPflow package in Python [51]. Depending on the data pattern and the dominant features of the time-series, a single or composite kernel, i.e., a combination of multiple covariance functions, is employed to capture different characteristics of the data.

Furthermore, as the wet-weather flow (WWF) is not as predictable as the dry-weather flow (DWF) due to the intense flows entering the sewers after each rainfall in a catchment, ordinary kernels may not fully capture these changes. So, a composite kernel designed to take stormwater and wastewater characteristics into account is needed.

2.4. Kernel Design

The selection of an appropriate kernel, or covariance function, is fundamental to the performance of a GPR model, as it encodes assumptions about the function being modelled. Various methods have been proposed for automatically selecting the optimal kernel for a GP model [52,53]. However, these methods add to the overall complexity of the model significantly which is not desirable in this study.

Moreover, from an overview of the water usage and sewer flow patterns, and the impact of the rainfall magnitude on the flow, it can be understood that the wastewater flow will follow some basic patterns that can be effectively captured within the main GP model by tuning the hyperparameters.

The main kernels used in this research are Periodic and Matérn kernels. Periodic kernel can be formulated as follows [46]:

$$k_{Per}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{2Sin^2\left(\frac{x - x'}{2}\right)}{l^2}\right), \tag{7}$$

where l is the length scale, and $\mathbf{x}$ and $\mathbf{x}'$ are GP input vectors.

Matérn kernel family can be formulated as [46]:

$$k_{Mat}(\boldsymbol{x}, \boldsymbol{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}|\boldsymbol{x} - \boldsymbol{x}'|}{l}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}|\boldsymbol{x} - \boldsymbol{x}'|}{l}\right), \tag{8}$$

where K_{ν} is the modified Bessel function [54]. l is the length scale and ν is a free parameter which is typically selected as 0.5, 1.5, or 2.5. Matérn kernels with higher value of ν are smoother, compared to smaller ν values.

The effectiveness of the designed kernel is assessed by the model checking metrics discussed in Section 2.5.

2.5. Anomaly Detection

To apply the GPR method for anomaly detection in sewer systems, the same training process used in the forecasting model is adopted, but with real-time depth measurements from sensors installed in manholes. These depth values serve as the training and testing datasets for the GPR model.

A GPR model is trained using real-time depth measurements from the sensors and precipitation data obtained from the nearest online rain gauges. The trained model predicts expected depth values for each manhole in real-time. By comparing these predicted values with the actual sensor readings, discrepancies can be identified. If deviations exceed predefined thresholds set by the GPR model, an anomaly is detected. Anomalies typically manifest as unusually low or high depth values, indicating

 doi:10.20944/preprints202505.1621.v1

potential blockages—either downstream (causing increased water levels) or upstream (leading to sudden drops in depth).

In many cases, blockages only reduce the capacity of the pipe rather than fully obstructing the flow. Consequently, they do not lead to a surcharge, or a spike change in the depth enough to be flagged as an anomaly. However, GPR model can detect these changes as small deviations that can be considered as probable blockages and be put into inspection plans for early interventions.

The anomaly detection mechanism also considers data resolution. A single outlier may not always indicate an issue, so a predefined number of consecutive data points exceeding the threshold is required to trigger an alarm. The exact criteria can be adjusted based on case-specific conditions.

2.6. Model Checking

The role of predictive model checking is to assess the practical fit of a model [55]. Various techniques exist for assessing the fit of a model including statistic measures [56] and graphical checks [57]. Among them, root mean square error (RMSE), coverage and entropy are selected as three key metrics assessing the GPR model in this study. RMSE, coverage, and entropy together offer a balanced evaluation of the model by capturing its accuracy, the effectiveness of its uncertainty bounds, and the overall reliability of its probabilistic predictions.

2.6.1. Root Mean Square Error

RMSE measures the accuracy of the predictions by calculating the standard deviation of the residuals, providing a straightforward interpretation of the prediction error [58].

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (7)$$

where N is the number of test points, y_i is the actual value (flow, depth, etc.), and $\hat{y}_i$ is the predicted value.

2.6.2. Coverage

Coverage represents the percentage of data points that fall within the defined credible interval of the model. A good model is defined as a model having an α% coverage for an α% credible interval [56]. The typical value for α is 95, which is used in this research as well.

2.6.3. Entropy

Entropy quantifies the uncertainty associated with the predicted distribution, with higher entropy showing greater uncertainty [59]. For a univariate Gaussian probability density function with a variance of σ^2, the differential entropy $h(t)$ is given by the formula [60]:

$$h(t) = 0.5\log(2\pi e\sigma^2) \quad (8)$$

The overall differential entropy of the test points can be calculated as:

$$H = \frac{1}{N}\sum_{i=1}^{N} 0.5\log(2\pi e\sigma^2) \quad (9)$$

This value quantitatively reflects how narrow or wide the bounds are defined. A wider bound can increase the coverage but will result in higher entropy. Thus, a trade-off between these three measures—RMSE, coverage and entropy—is necessary in model evaluation.

It is noteworthy that the differential entropy used in this study is slightly different with the Shannon entropy used for discrete distributions. Unlike Shannon entropy, differential entropy can

 doi:10.20944/preprints202505.1621.v1

take negative values when the variance of the data is very small which completely depends on the unit of the data. Therefore, differential entropy values from different models are not directly comparable.

3. Results and Discussion

3.1. Forecasting the WWTP Influent

The inflow of the WWTP is a critical metric for sewer system management, as it significantly influences the performance of the treatment plant. Accurate forecasting of WWTP influent enables proactive decision-making, allowing operators to optimise treatment processes, manage resources effectively, and mitigate potential issues such as overflows. Using a GPR model, flow predictions can be interpreted probabilistically, providing not only point predictions but also a measure of the associated uncertainty.

The WWTP influent values have been derived from the hydraulic simulator (PySWMM). To simulate real-world noise and variability, a zero-mean normal noise with a standard deviation of 0.8 L/s was added to the simulated values. The train and test datasets were then prepared and processed to construct the GPR model. The model's ability to capture the temporal dynamics of the influent, including both DWF patterns and the response to rainfall events, is crucial for effective wastewater management.

Figure 2 illustrates the WWTP influent and precipitation over the study period. The graph highlights the variability in influent flow, demonstrating the influence of rainfall events on peak flow rates. The total precipitation for the year is 875.6 mm.

To build and test the GPR model's predictions, 6 months of the study period were selected as the training dataset, and the subsequent 2 months for testing. The 6-month period allow the model capture daily, weekly and monthly patterns effectively. The interval between the data points is one hour and the simulation timesteps are 180 seconds. The results of the GPR model are presented in Figure 3.

The bounds illustrate a 95% credible interval of the Gaussian distributions, corresponding to 1.96 standard deviations. Figure 4 illustrates the predicted distribution at a randomly selected point, from the zoomed-in section in Figure 3, showing how the Gaussian distributions work with the data.

As shown in Figure 3, the predictive uncertainty in the GPR model increases as the predictions move further from the training data points, leading to a wider bound around the GPR mean. This occurs because the farther a prediction is from the training set, the less information the model has to constrain it. The mathematical reason behind this fact can be seen in Equation 5. As the distance between $\boldsymbol{X}$ and $\boldsymbol{X}_*$ increases, the $K(\boldsymbol{X}_*, \boldsymbol{X})$ becomes smaller and the covariance of the predictions gets higher values. This complies with the fundamental principle that predictions become more uncertain as they extend beyond the observed data points.

The designed kernel for this model consists of four distinct covariance functions. The first two are periodic kernels, selected to capture the hourly and daily variations of the wastewater production pattern. These two kernels target the time input in the training data. The other kernel which is particularly designed to capture rainfall effect on the flow is a Matérn12 kernel (from the gpflow.GPR library in Python). This kernel belongs to the Matérn covariance family with $\nu = 1/2$ that makes the covariance function more rough compared to other ν values [46]. As a result, it can take the spike changes in flow caused by rainfalls.

The final kernel is a noise kernel used to handle the noise in observed data. The noise kernel is a two-dimensional function that is combined with other kernels by a simple addition.

 doi:10.20944/preprints202505.1621.v1

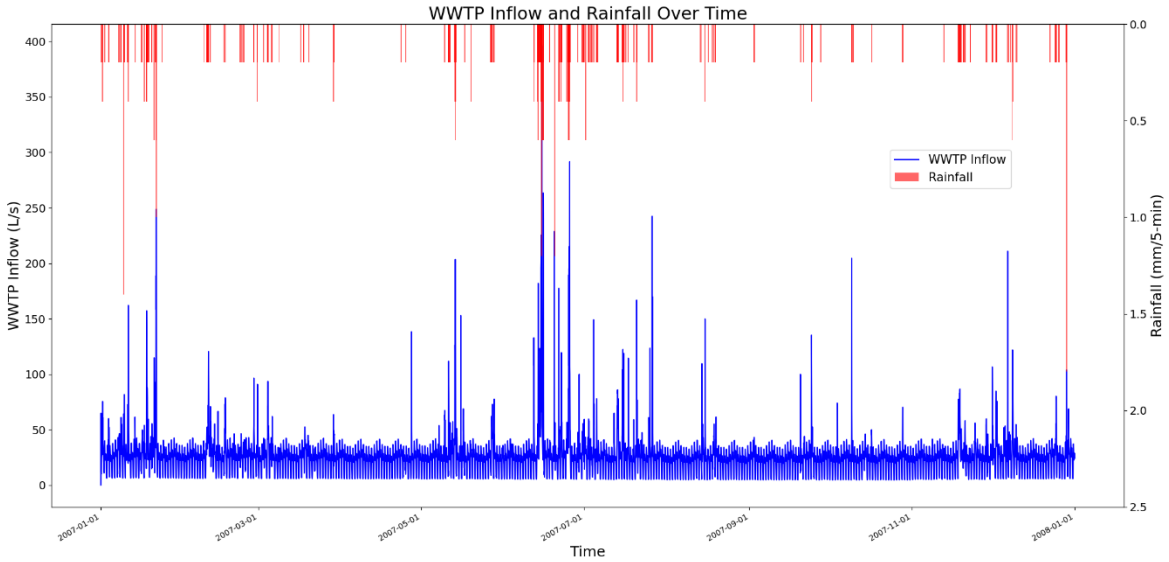

Figure 2. WWTP inflow and precipitation in the case study. The blue line shows the inflow change overtime and the red line shows the precipitation depth in 5-minute intervals.

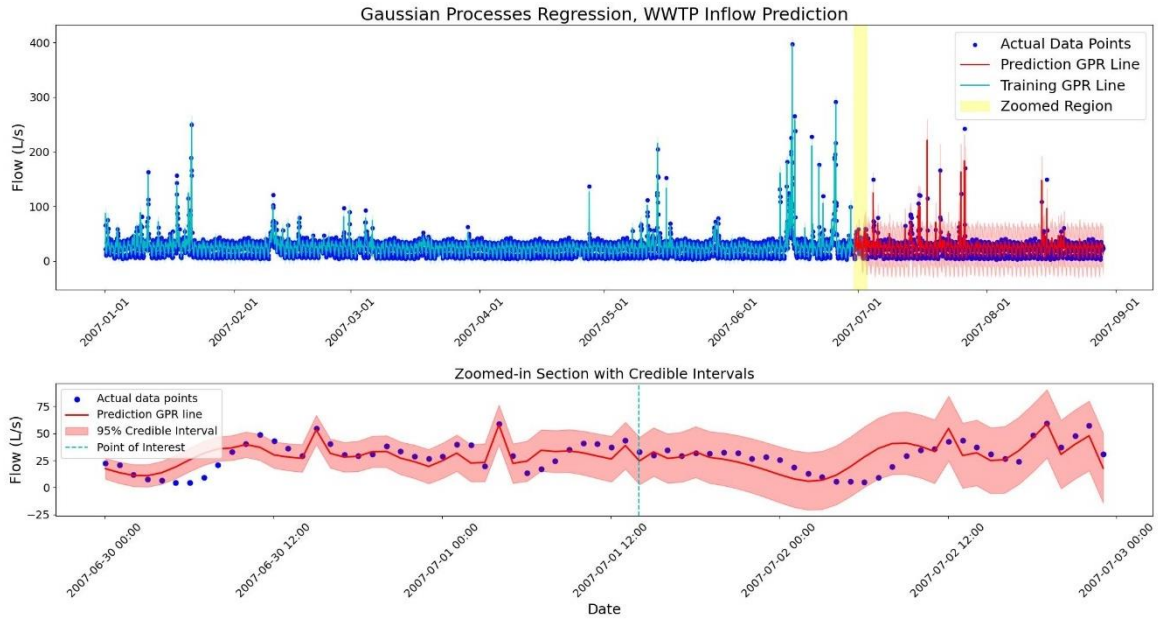

Figure 3. Gaussian process regression model on 6 months of training and 2 months of test data (top). The regression lines are shown in blue and red for training and test sets, respectively. The first 3 days of the test set is selected for a detailed illustration (bottom).

The choice of this kernel structure is supported by model evaluation results and its ability to reproduce the observed data patterns. The model checking shows that over the 2-month test period, Coverage of the model is 99.4% for the 95% credible interval which indicates a perfect coverage. The RMSE is 10.87 L/s, which is reasonable given the high variability in the data and a mean value of 26.16 L/s of the data points.

Additionally, the average differential entropy across the test points is 6.19, providing a measure of predictive uncertainty as defined in Section 2.5.3. While differential entropy's dependence on data scaling means its values are not directly comparable across different datasets and may not translate to an easily interpreted practical metric, it remains useful for model selection within a specific analysis. For instance, when evaluating different kernel designs using the same training and testing

 doi:10.20944/preprints202505.1621.v1

data, differential entropy helps identify which configuration provides predictions with lower overall uncertainty.

The model's reliability highly depends on the range of values used in the training step. For this model, having rainfall events with greater depth values can train the model better, particularly for predictions at wet-weather periods. As shown in Figures 2 and 3, most of the large rainfalls happen at the 6th month of the year where it is placed in the training zone of the GPR model. To show the importance of the wide span of training data values, the training period is shifted back by one month and the model uses the first five months of the period for training. Hence, the sixth month with its heavy rainfalls is put into the test zone.

The results of this model show that some of the high flow values at the testing zone fall outside predicted bounds which indicates the importance of a representative value range in the training data for achieving accurate predictions. The model checking measures with the same kernel setting indicate 13.82 L/s of RMSE, 98.2% of coverage, and 5.98 of entropy. The increased value of RMSE shows a weaker fit of data which also can be proved by the visual GPR line in the predictions. The entropy and coverage values have not changed significantly; however, outliers exist in the high flow times.

This GP regression, shown in Figure 3, is built with hourly data, resulting in 4,320 training points. For a more detailed prediction, data of the same period with a 5-minute resolution can be used. However, this increases the number of training points to 51,840 which significantly increases the computational cost of the model due to the $O(N^3)$ complexity of the GP. In such cases, methods such as Sparse Variational Gaussian Processes (SVGP) can be beneficial in handling the large dataset [61].

As it is evident in Figure 3, a few outliers exist in the predicted data. These outliers mostly happen in the timesteps when there was an intense rainfall and consequently, a spike change has occurred in the flow. Also, some of the data fall outside the bounds at the starting hours of the prediction when the model is adapting itself with the data and the bounds are narrower than the other predicted timesteps.

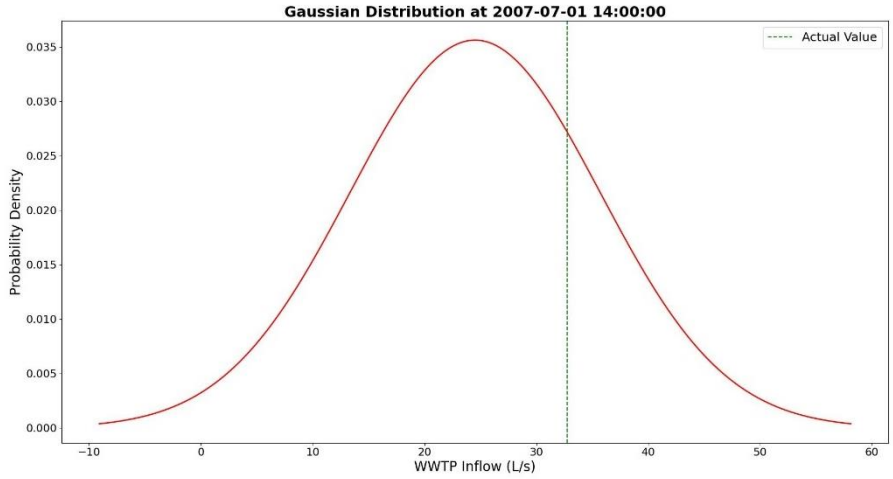

Figure 4. Gaussian distribution of the predicted WWTP inflow at 2007-07-01 at 14:00pm. The green dashed line shows the actual test point value. The distribution's domain is selected as ±4 standard deviation of the distribution for a better illustration.

3.2. Forecasting Node Surcharges and Overflows

Using the same method, the water level in manholes—where the sensors are usually placed—can also be predicted. These predictions are useful in managing surcharges and overflows. Moreover, since the CSOs are spotlights of the sewer systems and water companies monitor them intently, forecasting CSOs would be another helpful application of GPR.

 doi:10.20944/preprints202505.1621.v1

In this case study, the manhole with the most surcharge and overflow issues has been selected for the GPR forecasting. This selection ensures a sufficient number of data points are available for model evaluation. Also, it aligns with operational priorities, as manholes with persistent issues demand more attention over time. The results of the water depth prediction at the J15 node (the selected manhole) are presented in Figures 5 and 6.

Given that the manhole depth is 2.5 m, the maximum observed values reach 250 cm in Figure 5, and the GPR model accounts for this range of variation during training. The results show that although significant rainfall events occurred during the test period, the GPR model does not predict excessively high water depths. This prediction can be restricted to certain values, but the model appears to do well without such restrictions.

Since the water level pattern closely resembles the flow pattern, the same composite kernel with tuned length scales and variances has been implemented in the GPR model. The model checking results show that the RMSE of the predictions is 10.9 cm, which is somewhat high compared to the 17.6 cm average water depth in the manhole over time. The coverage of the prediction is 99.4% that shows a perfect performance. Also, the differential entropy is calculated as 6.22.

In Figure 6, a predicted point is randomly selected for illustration. The pipe diameter is 50 cm, and water levels exceeding this threshold cause surcharges in the manhole. This value is marked with a dashed line, and the area under the predictive distribution exceeding this threshold is coloured to indicate the probability of surcharge. In this figure, the surcharge probability is 1.75%, which can be assumed to be negligible. The actual value at this point also confirms that no surcharge happened at this timestep.

This type of probabilistic output can be beneficial for developing a decision-making system which can be used in the water industry. With the same approach employed in finding the surcharge probability, manhole overflow probability can be predicted over time. These probabilistic forecasts allow operators to prioritise maintenance and allocate resources more efficiently by focusing on locations and times with the highest risk. Figure 7 indicates the probability of surcharges and overflows in the 2-month test period of this case study.

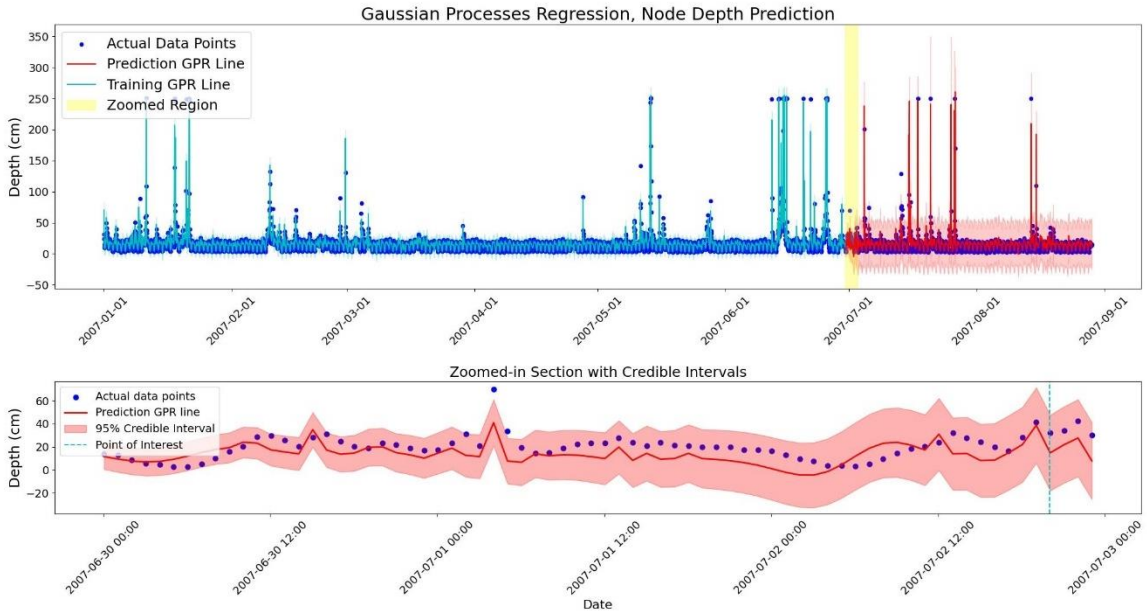

Figure 5. Water depth forecasting at node J15 for 2 months (top). GPR line at training and test datasets are illustrated in blue and red, respectively. The first 3 days of the test set is selected for a detailed illustration (bottom).

Figure 7 shows one overflow prediction with probability greater than 50% and six predictions with 25 to 50% probability of happening. Among these seven predictions, four test points show that

 doi:10.20944/preprints202505.1621.v1

there is an actual overflow and the other three show no overflow. For predictions in the low-risk range, no overflow occurred in the simulation and the forecasts are made accurately.

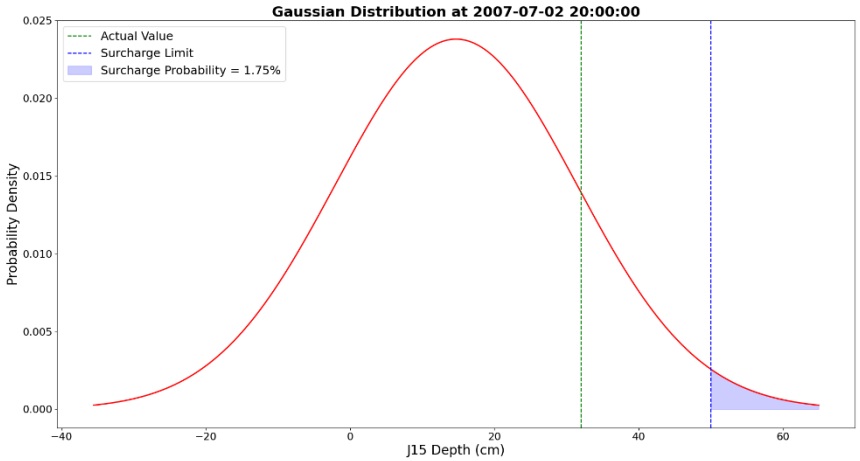

Figure 6. Gaussian distribution of the predicted node depth on 2007-07-02 at 8:00 pm. The green dashed line shows the actual test point value. The purple dashed line also indicates the surcharge level which equals the pipe diameter. The area under the distribution for values greater than the surcharge limit is coloured in purple and shows the surcharge probability. The distribution's domain is selected as ±4 standard deviation of the distribution for a better illustration.

During the 2-month period of the test period, surcharges occurred in 26 timesteps. The surcharge probability derived from the GPR model assigned different probabilities for each event, as presented in Table 2. Among the 26 surcharges, 5 events were not flagged as high- or medium-risk events while the remaining 21 timesteps were correctly identified with significant probability of surcharge. These surcharges occurred across 11 rainfall events and the model could predict 8 as high-risk and 2 as medium-risk events.

The same approach can be applied to forecast CSOs in the system. In this case study, the CSO is located downstream of a weir, which receives flow from an upstream manhole with an inlet offset. By predicting the water level at the upstream manhole, it becomes possible to estimate CSO occurrences more effectively. This approach allows the model to run more confidently using a continuous timeseries from the upstream manhole, which spans all training points and contains meaningful variation. In contrast, the CSO flow data contains only a few non-zero values (eight, in the one-year simulation period of this study) during heavy rainfall events, with the majority of timesteps showing zero flow.

 doi:10.20944/preprints202505.1621.v1

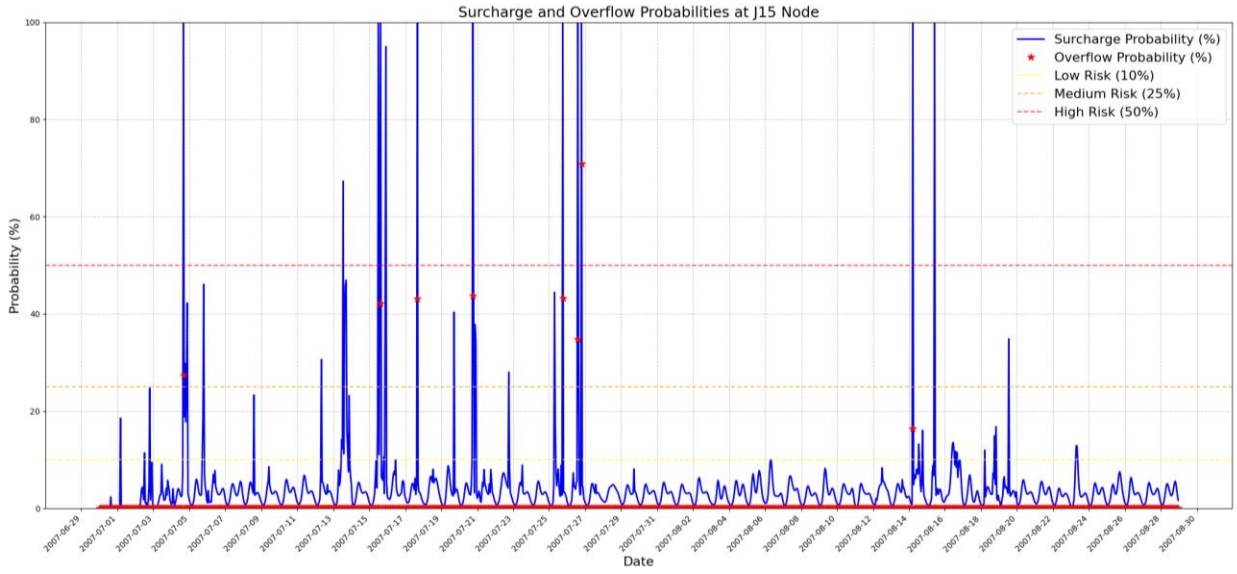

Figure 7. The probability of surcharge and overflow in the selected manhole. Blue line shows the probability of surcharge over time, and the red stars show the probability of an overflow from the manhole in each timestep. The dashed lines are drawn to divide the space into low, medium and high-risk zones.

Table 2. Details of the surcharge events and the corresponding probability values.

Surcharge event time	Node depth (cm)	Probability of surcharge (%)
2007-07-01 04:03	69.87	18.57
2007-07-04 16:03	200.96	100.0
2007-07-04 17:03	53.41	18.87
2007-07-04 18:03	57.83	29.83
2007-07-04 20:03	57.31	26.03
2007-07-05 19:03	74.39	46.10
2007-07-13 13:03	73.24	67.287
2007-07-13 14:03	56.46	11.23
2007-07-13 15:03	128.68	41.43
2007-07-13 16:03	74.78	45.86
2007-07-13 17:03	76.77	46.95
2007-07-13 21:03	71.39	23.23
2007-07-13 22:03	61.70	9.03
2007-07-15 12:03	95.31	100
2007-07-15 14:03	79.16	55.31
2007-07-15 15:03	75.72	100
2007-07-20 18:03	249.72	99.97
2007-07-20 19:03	59.88	47.38
2007-07-25 18:03	68.676	99.97
2007-07-26 14:03	250	99.99
2007-07-26 15:03	53.65	34.08
2007-07-26 19:03	169.6	100
2007-08-14 05:03	249.65	99.99
2007-08-15 10:03	110.05	100

3.3. Anomaly Detection

The primary cause of anomaly in sewer systems is the blockage. In this study, an artificial blockage is introduced in the C6 conduit of the hydraulic model, located between nodes J2 and J3. This blockage is simulated by increasing the roughness of the pipe instantly. This increase in

 doi:10.20944/preprints202505.1621.v1

roughness lowers the ability of the pipe to transfer water; therefore, the flow reduces, upstream depth increases, and downstream depth decreases.

These sudden changes in depth and flow can be considered as anomalies by comparing the observed values with GPR prediction results taken from the previous timesteps. When a sequence of observed values consistently falls outside the credible interval bounds of the predictions, it can serve as a reliable indicator of a potential blockage.

The results of the depth prediction for the J2 node (the upstream manhole) are illustrated in Figure 8. For clarity, only the last two days of the training data from the 2-month training period are shown.

As shown in Figure 8, after 3 hours of consequent data points falling outside the bounds, an alarm has been raised. This threshold is defined by the authors and can be adjusted based on the data points interval. Lowering this threshold would cause false alarms and lead to an increased inspection cost. However, false alarms are generally preferred over missed blockages [62].

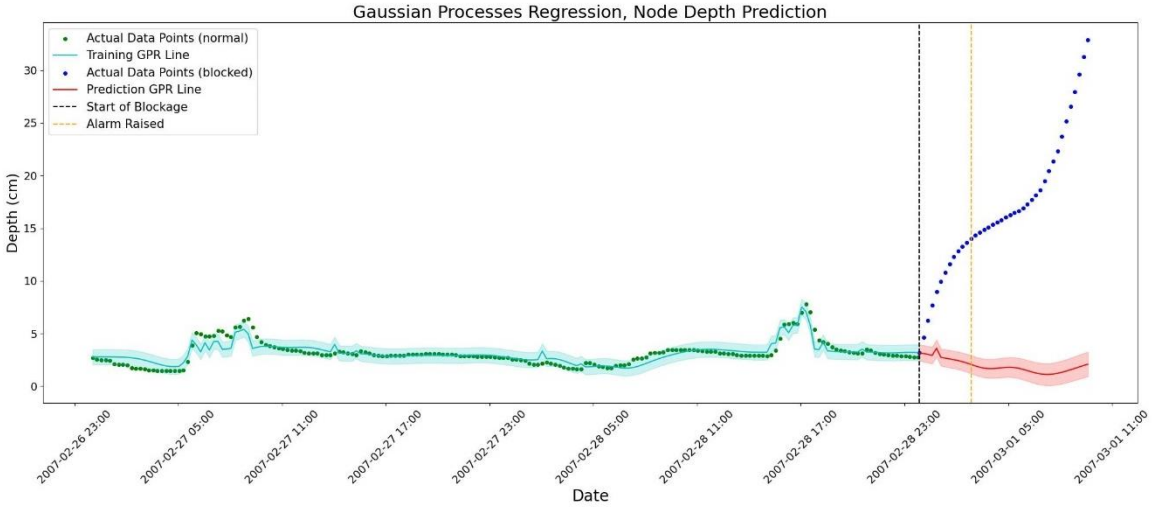

Figure 8. Blockage detection by water depth forecasting at J2 node. The prediction for 10 hours has been presented by red line and the observed values have been presented by blue bullets. After three hours of consequent data points falling outside the bounds, an alarm is raised showing a probable blockage. The time interval between points is 15 minutes.

The same approach can be applied to the downstream manhole (J3). Nonetheless, for the nodes receiving flow from multiple sources, blockage in one feeding pipe cannot be detected easily. A promising direction for future research is to use data-driven predictions from pairs of nodes that are directly connected or linked via the shortest paths. This can improve the ability of anomaly detection in the sewer system.

3.4. Prediction with Limited Data

As previously stated, GP models are capable of making predictions with minimal data for the training process and are more reliable in this case compared to other regression and classification methods among data-driven models. To demonstrate this, a flow prediction with one week of data has been made for a 3-day forecasting period. The test period is deliberately selected from the wet-weather periods to capture significant changes in the flow. The results are presented in Figure 9.

As the results show, the GPR line is fitted well on the data and the predictions seem reasonable. Model checking results also show a coverage of 87.5%, RMSE of 7.48 L/s, and entropy of 4.28, which outperform the long-term prediction made on the 6-month training data.

 doi:10.20944/preprints202505.1621.v1

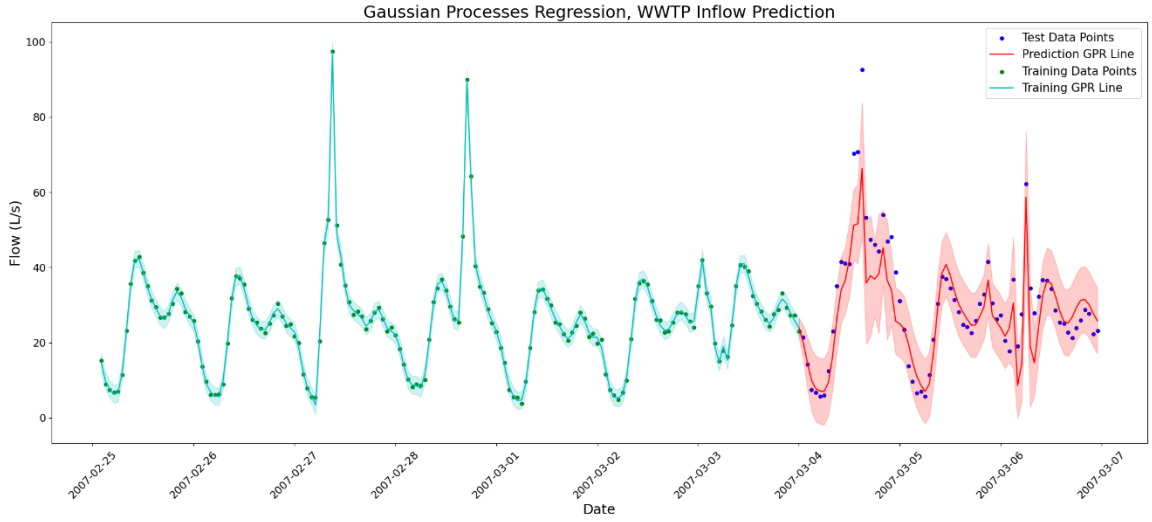

Figure 9. WWTP inflow prediction with seven days of training data using Gaussian process regression. The test period is three days, and the 95% credible interval is used to show the uncertainty around the predictions.

The primary limitation of using short training periods is the insufficient number of rainfall events available to train the model effectively. So, if a significant rain event occurs in the test points, the model is incapable of predicting the correct corresponding flow.

3.5. Further Discussion

The rainfall depths that have been used in this research were treated as deterministic inputs. However, in a predictive model, rainfalls will appear as probabilistic values with a likelihood of happening. Therefore, for an improved prediction, rainfall values can be set as Gaussian probability density function, each representing a level of uncertainty.

Moreover, the predictions made in this study were all in a range of short- and medium-term periods because the main goal of this study was to detect flaws and anomalies, and to warn for high or low flows in a sewer system. The long-term predictions can be conducted with the aim of GPs, but the computational costs and high values of uncertainties may make other data-driven models more suitable for asset management applications.

It is important to note that rain gauge data is typically not available in real time and often arrives with a delay of 15 minutes or more. However, this latency does not significantly affect the performance of the anomaly detection model, as the alarm mechanism is designed to trigger only after a defined number of consecutive data points fall outside the credible bounds.

Finally, incorporating physical constraints into the model can significantly enhance its realism and robustness. For example, the mean function which is commonly assumed to be zero in standard GPR can be replaced with a physics-informed prior that reflects expected system behaviour and is updated during posterior inference. Additionally, physical constraints, such as the continuity equation or operational limits like minimum flow rates and maximum water depth in each pipe or manhole, can be embedded into the model. These constraints help ensure that predictions remain within realistic and operationally acceptable bounds.

5. Conclusions

In this study, Gaussian processes regression (GPR) was implemented to predict flow and depth in a sewer system. The case study was a hypothetical skeleton sewer system with different characteristics of a real-world system serving 11,300 inhabitants of a city in UK.

Based on the results taken from different analyses, the following conclusions can be obtained:

 doi:10.20944/preprints202505.1621.v1

I. A probabilistic forecasting model can take different uncertainties into account and provides a likelihood of happening which seems to be more realistic than presenting a single value as a definite answer.
II. In sewer flow prediction, taking time and precipitation as the main inputs of the model makes the model reliable. In contrast, having a single-input model cannot effectively reflect flow changes during wet-weather periods. Also, adding new inputs to the model only adds to the dimension of the covariance matrix which is directly related to the computational cost of the GPs.
III. Designing a proper kernel is crucial for making a good forecast. The model can do it with maximising the log marginal likelihood; however, defining some metrics like RMSE, coverage and differential entropy can help in finding the best kernel setting.
IV. GPR can respond well to minimal training datasets and make reliable predictions. On the other hand, high number of data points can reduce the model's speed and make it computationally inefficient. In such cases, sparsification methods like SVGP can come helpful.
V. Outliers may appear in every prediction, often due to a lack of wide range in training datasets, especially when dealing with precipitation data. In such cases, increasing the length of the training period or defining physical constraints help.
VI. Real-time control has been traditionally done with deterministic models. Now, a probabilistic method is showing a reliable real-time prediction which can be applied by water companies in many real-world cases with natural uncertainties.
VII. The surcharge, overflow and CSO forecasting significantly contribute to an enhanced environment and public health—key objectives for every water-related research.
VIII. This probabilistic approach equips decision-makers to act on sewer risks with greater confidence, supporting resilient infrastructure planning and reducing threats to public health and the environment.

In summary, this research offers a valuable framework for probabilistic forecasting and anomaly detection in sewer systems. Future work can enhance this approach by incorporating physical constraints into the model and representing input variables, such as rainfall, as probability distributions to capture input uncertainty more accurately. Furthermore, this probabilistic platform has the potential to serve as an alternative to commercial products currently used by water companies, offering a transparent and adaptable decision-support tool.

Author Contributions: Conceptualisation, M.R., H.R., P.M; methodology, M.R., H.R.; software, M.R.; validation, M.R.; formal analysis, M.R.; investigation, M.R., H.R.; resources, M.R., H.R., P.M.; data curation, M.R.; writing—original draft preparation, M.R.; writing—review and editing, M.R., H.R., P.M.; visualisation, M.R.; supervision, H.R., P.M.; project administration, H.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The hydraulic model and the source code for forecasting flow are available through the GitHub repository https://github.com/MohsenRz/Sewer_System_Forecasting_Emulator.git .

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CSO	Combines Sewer Overflow
OFWAT	Water Service Regulation Authority
ANN	Artificial Neural Network
SVM	Support Vector Machine
GP	Gaussian Process

 doi:10.20944/preprints202505.1621.v1

GPR	Gaussian Processes Regression
LPCD	Litre per Capita per Day
SWMM	Storm Water Management Model
DWF	Dry-weather Flow
WWF	Wet-weather Flow
CCTV	Closed-circuit Television
RMSE	Root Mean Square Error
SVGP	Sparse Variational Gaussian Processes

References

1. Walski, T.M.; Barnard, T.E. *Wastewater Collection System Modeling and Design*; Haestad Press: Waterbury, CT, USA, 2004.
2. Owolabi, T.A.; Mohandes, S.R.; Zayed, T. Investigating the impact of sewer overflow on the environment: A comprehensive literature review paper. *J. Environ. Manag.* **2022**, *301*, 113810. https://doi.org/10.1016/j.jenvman.2021.113810
3. Perry, W.B.; Ahmadian, R.; Munday, M.; Jones, O.; Ormerod, S.J.; Durance, I. Addressing the challenges of combined sewer overflows. *Environ. Pollut.* **2024**, *343*, 123225. https://doi.org/10.1016/j.envpol.2023.123225
4. Faris, N.; Zayed, T.; Aghdam, E.; Fares, A.; Alshami, A. Real-time sanitary sewer blockage detection system using IoT. *Measurement* **2024**, *226*, 114146. https://doi.org/10.1016/j.measurement.2024.114146
5. Arthur, S.; Crow, H.; Pedezert, L. Understanding blockage formation in combined sewer networks. *Proc. Inst. Civ. Eng.-Water Manag.* **2008**, *161*, 215–221. https://doi.org/10.1680/wama.2008.161.4.215
6. Balla, K.M.; Bendtsen, J.D.; Schou, C.; Kallesøe, C.S.; Ocampo-Martinez, C. A learning-based approach towards the data-driven predictive control of combined wastewater networks–An experimental study. *Water Res.* **2022**, *221*, 118782. https://doi.org/10.1016/j.watres.2022.118782
7. Perez, G.; Gomez-Velez, J.D.; Grant, S.B. The sanitary sewer unit hydrograph model: A comprehensive tool for wastewater flow modeling and inflow-infiltration simulations. *Water Res.* **2024**, *249*, 120997. https://doi.org/10.1016/j.watres.2023.120997
8. Zhang, Q.; Li, Z.; Snowling, S.; Siam, A.; El-Dakhakhni, W. Predictive models for wastewater flow forecasting based on time series analysis and artificial neural network. *Water Sci. Technol.* **2019**, *80*, 243–253. https://doi.org/10.2166/wst.2019.263
9. Li, S.; Tian, W.; Yan, H.; Zeng, W.; Tao, T.; Xin, K. Modeling transient mixed flows in sewer systems with data fusion via physics-informed machine learning. *Water Res. X* **2024**, *25*, 100266. https://doi.org/10.1016/j.wroa.2024.100266
10. Stieglitz, M.; Hobbie, J.; Giblin, A.; Kling, G. Hydrologic modeling of an arctic tundra watershed: Toward Pan-Arctic predictions. *J. Geophys. Res.-Atmos.* **1999**, *104*, 27507–27518. https://doi.org/10.1029/1999JD900845
11. Ge, J.; Li, J.; Qiu, R.; Shi, T.; Zhang, C.; Huang, Z.; Yuan, Z. A data-driven method for estimating sewer inflow and infiltration based on temperature and conductivity monitoring. *Water Res.* **2024**, *261*, 122002. https://doi.org/10.1016/j.watres.2024.122002
12. Donnelly, J.; Daneshkhah, A.; Abolfathi, S. Forecasting global climate drivers using Gaussian processes and convolutional autoencoders. *Eng. Appl. Artif. Intell.* **2024**, *128*, 107536. https://doi.org/10.1016/j.engappai.2023.107536
13. Machac, D.; Reichert, P.; Rieckermann, J.; Albert, C. Fast mechanism-based emulator of a slow urban hydrodynamic drainage simulator. *Environ. Model. Softw.* **2016**, *78*, 54–67. https://doi.org/10.1016/j.envsoft.2015.12.007
14. Troutman, S.C.; Schambach, N.; Love, N.G.; Kerkez, B. An automated toolchain for the data-driven and dynamical modeling of combined sewer systems. *Water Res.* **2017**, *126*, 88–100. https://doi.org/10.1016/j.watres.2017.08.065
15. Aliashrafi, A.; Zhang, Y.; Groenewegen, H.; Vanrolleghem, P.A. A Review of Data-Driven Modelling in Drinking Water Treatment. *Rev. Environ. Sci. Biotechnol.* **2021**, *20*, 985–1009. https://doi.org/10.1007/s11157-021-09592-y

 doi:10.20944/preprints202505.1621.v1

16. Swiler, L.P.; Gulian, M.; Frankel, A.L.; Safta, C.; Jakeman, J.D. A survey of constrained Gaussian process regression: Approaches and implementation challenges. *J. Mach. Learn. Model. Comput.* **2020**, *1*, 119-156. DOI: 10.1615/JMachLearnModelComput.2020035155
17. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045
18. Thorndahl, S.; Willems, P. Probabilistic modelling of overflow, surcharge and flooding in urban drainage using the first-order reliability method and parameterization of local rain series. *Water Res.* **2008**, *42*, 455–466. https://doi.org/10.1016/j.watres.2007.07.038
19. Breinholt, A. *Uncertainty in Prediction and Simulation of Flow in Sewer Systems*. Ph.D. Thesis, Technical University of Denmark, Kongens Lyngby, Denmark, 2012.
20. Raimondi, A.; Sanfilippo, U.; Becciu, G. Uncertainty on flow rate and temperature measurement for the detection of illicit flows in sewers. *J. Hydrol.* **2024**, *632*, 130891. https://doi.org/10.1016/j.jhydrol.2024.130891
21. Sriwastava, A.K.; Tait, S.; Schellart, A.; Kroll, S.; Van Dorpe, M.; Van Assel, J.; Shucksmith, J. Quantifying uncertainty in simulation of sewer overflow volume. *J. Environ. Eng.* **2018**, *144*, 04018050. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001392
22. Ding, C.; Rappel, H.; Dodwell, T. Full-field order-reduced Gaussian Process emulators for nonlinear probabilistic mechanics. *Comput. Methods Appl. Mech. Eng.* **2023**, *405*, 115855. https://doi.org/10.1016/j.cma.2022.115855
23. Wang, J. An intuitive tutorial to Gaussian process regression. *Comput. Sci. Eng.* **2023**, *25*, 4–11. DOI: https://doi.org/10.1109/MCSE.2023.3342149
24. Ding, C.; Chen, Y.; Rappel, H.; Dodwell, T. Functional order-reduced Gaussian Processes based machine-learning emulators for probabilistic constitutive modelling. *Compos. Part A Appl. Sci. Manuf.* **2023**, *173*, 107695. https://doi.org/10.1016/j.compositesa.2023.107695
25. Gonzalvez, J.; Lezmi, E.; Roncalli, T.; Xu, J. Financial applications of Gaussian processes and Bayesian optimization. *arXiv* **2019**, arXiv:1903.04841. https://doi.org/10.48550/arXiv.1903.04841
26. Roberts, S.; Osborne, M.; Ebden, M.; Reece, S.; Gibson, N.; Aigrain, S. Gaussian processes for time-series modelling. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2013**, *371*, 20110550. https://doi.org/10.1098/rsta.2011.0550
27. Sun, A.Y.; Wang, D.; Xu, X. Monthly streamflow forecasting using Gaussian process regression. *J. Hydrol.* **2014**, *511*, 72–81. https://doi.org/10.1016/j.jhydrol.2014.01.023
28. Pastrana-Cortés, J.D.; Gil-Gonzalez, J.; Álvarez-Meza, A.M.; Cárdenas-Peña, D.A.; Orozco-Gutiérrez, Á.A. Scalable and interpretable forecasting of hydrological time series based on variational Gaussian processes. *Water* **2024**, *16*, 2006. https://doi.org/10.3390/w16142006
29. Samuelsson, O.; Björk, A.; Zambrano, J.; Carlsson, B. Gaussian process regression for monitoring and fault detection of wastewater treatment processes. *Water Sci. Technol.* **2017**, *75*, 2952–2963. doi: https://doi.org/10.2166/wst.2017.162
30. Ng, J.Y.; Fazlollahi, S.; Dechesne, M.; Soyeux, E.; Galelli, S. Robust optimal design of urban drainage systems: A data-driven approach. *Adv. Water Resour.* **2023**, *171*, 104335. https://doi.org/10.1016/j.advwatres.2022.104335
31. Wang, Y.; Ocampo-Martinez, C.; Puig, V. Stochastic model predictive control based on Gaussian processes applied to drinking water networks. *IET Control Theory Appl.* **2016**, *10*, 947–955. https://doi.org/10.1049/iet-cta.2015.0657
32. Grbić, R.; Kurtagić, D.; Slišković, D. Stream water temperature prediction based on Gaussian process regression. *Expert Syst. Appl.* **2013**, *40*, 7407–7414. https://doi.org/10.1016/j.eswa.2013.06.077
33. Bonakdari, H.; Ebtehaj, I.; Samui, P.; Gharabaghi, B. Lake water-level fluctuations forecasting using minimax probability machine regression, relevance vector machine, Gaussian process regression, and extreme learning machine. *Water Resour. Manag.* **2019**, *33*, 3965–3984. https://doi.org/10.1007/s11269-019-02346-0

 doi:10.20944/preprints202505.1621.v1

34. Sweetapple, C.; Webber, J.; Hastings, A.; Melville-Shreeve, P. Realising smarter stormwater management: A review of the barriers and a roadmap for real world application. *Water Res.* **2023**, *120505*. https://doi.org/10.1016/j.watres.2023.120505
35. Patil, R.R.; Calay, R.K.; Mustafa, M.Y.; Ansari, S.M. AI-driven high-precision model for blockage detection in urban wastewater systems. *Electronics* **2023**, *12*, 3606. https://doi.org/10.3390/electronics12173606
36. Rosin, T.R.; Kapelan, Z.; Keedwell, E.; Romano, M. Near real-time detection of blockages in the proximity of combined sewer overflows using evolutionary ANNs and statistical process control. *J. Hydroinf.* **2022**, *24*, 259–273. https://doi.org/10.2166/hydro.2022.036
37. Jimoh, M.; Abolfathi, S. Modelling pollution transport dynamics and mixing in square manhole overflows. *J. Water Process Eng.* **2022**, *45*, 102491. https://doi.org/10.1016/j.jwpe.2021.102491
38. Li, N.; Wang, X.; Li, Z.; Zhao, F.; Nair, A.; Zhang, J.; Liu, C. Real-time identification and positioning of sewer blockage based on liquid level analysis in rural area. *Processes* **2023**, *11*, 161. https://doi.org/10.3390/pr11010161
39. Kargar, K.; Joksimovic, D. Analysis of sewer blockage causes using open data. *Water Pract. Technol.* **2024**, *19*, 3855–3866. https://doi.org/10.2166/wpt.2024.218
40. Rossman, L.A.; Simon, M.A. *Storm Water Management Model User's Manual Version 5.2*; US Environmental Protection Agency: Cincinnati, OH, USA, 2022.
41. Prodanovic, P.; Simonovic, S.P. *Development of Rainfall Intensity Duration Frequency Curves for the City of London under the Changing Climate*; Department of Civil and Environmental Engineering, The University of Western Ontario: London, ON, Canada, 2007.
42. Rezaee, M.; Tabesh, M. Effects of inflow, infiltration, and exfiltration on water footprint increase of a sewer system: A case study of Tehran. *Sustain. Cities Soc.* **2022**, *79*, 103707. https://doi.org/10.1016/j.scs.2022.103707
43. Zeydalinejad, N.; Javadi, A.A.; Webber, J.L. Global perspectives on groundwater infiltration to sewer networks: A threat to urban sustainability. *Water Res.* **2024**, *262*, 122098. https://doi.org/10.1016/j.watres.2024.122098
44. McDonnell, B.E.; Ratliff, K.; Tryby, M.E.; Wu, J.J.X.; Mullapudi, A. PySWMM: The python interface to stormwater management model (SWMM). *J. Open Source Softw.* **2020**, *5*, 1-3. doi: 10.21105/joss.02292
45. Stovin, V. *Mappin Green Roof Test Bed Rainfall and Runoff Data* **2007**; The University of Sheffield: Sheffield, UK, 2024. https://doi.org/10.15131/shef.data.27918345.v1
46. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.
47. Deshpande, S.; Rappel, H.; Hobbs, M.; Bordas, S.P.; Lengiewicz, J. Gaussian process regression + deep neural network autoencoder for probabilistic surrogate modeling in nonlinear mechanics of solids. *Comput. Methods Appl. Mech. Eng.* **2025**, *437*, 117790. https://doi.org/10.1016/j.cma.2025.117790
48. Rappel, H.; Beex, L.A.; Hale, J.S.; Noels, L.; Bordas, S.P. A tutorial on Bayesian inference to identify material parameters in solid mechanics. *Arch. Comput. Methods Eng.* **2020**, *27*, 361–385. https://doi.org/10.1007/s11831-018-09311-x
49. Duvenaud, D. *Automatic Model Construction with Gaussian Processes*. Ph.D. Thesis, University of Cambridge, Cambridge, UK, 2014.
50. Wilson, A.G. *Covariance Kernels for Fast Automatic Pattern Discovery and Extrapolation with Gaussian Processes*. Ph.D. Thesis, University of Cambridge, Cambridge, UK, 2014.
51. Matthews, A.G.; Van Der Wilk, M.; Nickson, T.; Fujii, K.; Boukouvalas, A.; Le, P.; Ghahramani, Z.; Hensman, J. GPflow: A Gaussian process library using TensorFlow. *J. Mach. Learn. Res.* **2017**, *18*, 1–6. https://www.jmlr.org/papers/v18/16-537.html
52. Duvenaud, D.; Lloyd, J.; Grosse, R.; Tenenbaum, J.; Ghahramani, Z. Structure Discovery in Nonparametric Regression through Compositional Kernel Search. In *Proceedings of the International Conference on Machine Learning*, Atlanta, GA, USA, 26 May 2013; pp. 1166–1174.
53. Kuok, S.C.; Yao, S.A.; Yuen, K.V.; Yan, W.J.; Girolami, M. Bayesian generative kernel Gaussian process regression. *Mech. Syst. Signal Process.* **2025**, *227*, 112395. https://doi.org/10.1016/j.ymssp.2025.112395
54. Abramowitz, M.; Stegun, I.A., Eds. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*; US Government Printing Office: Washington, DC, USA, 1948.

 doi:10.20944/preprints202505.1621.v1

55. Gilks, W.R.; Richardson, S.; Spiegelhalter, D., Eds. *Markov Chain Monte Carlo in Practice*; CRC Press: Boca Raton, FL, USA, 1995.
56. Malde, S. *Gaussian Process Emulators in Coastal Wave Modelling*. Ph.D. Thesis, University of Sheffield, Sheffield, UK, 2018.
57. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*, 3rd ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2021.
58. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001
59. Hasegawa, Y.; Nishiyama, T. Thermodynamic entropic uncertainty relation. *arXiv* **2025**, arXiv:2502.06174. https://doi.org/10.48550/arXiv.2502.06174
60. MacKay, D.J. *Information Theory, Inference and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
61. Liu, H.; Ong, Y.S.; Shen, X.; Cai, J. When Gaussian process meets big data: A review of scalable GPs. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 4405–4423. DOI: 10.1109/TNNLS.2019.2957109
62. Roghani, B.; Cherqui, F.; Ahmadi, M.; Le Gauffre, P.; Tabesh, M. Dealing with uncertainty in sewer condition assessment: Impact on inspection programs. *Autom. Constr.* **2019**, *103*, 117–126. https://doi.org/10.1016/j.autcon.2019.03.012