Physics-Informed Gaussian Process Regression for Predicting Flow in an Urban Drainage System

Mohsen Rezaee [a,b], Peter Melville-Shreeve [a,b], Hussein Rappel [a,*]

[a] Department of Engineering, Faculty of Environment, Science and Economy, University of Exeter, Exeter EX44QF, UK

[b] Centre for Water Systems, University of Exeter, Exeter EX44QF, UK.

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

Highlights

- Physics-informed GPs enable robust probabilistic urban drainage forecasting
- Designed kernels and physical constraints improve flow forecasting accuracy
- Sparse GP significantly reduces execution times for high-resolution datasets
- Novel stratified sparsification efficiently captures rare wet-weather flow spikes

Abstract

Accurate forecasting of urban drainage flows is critical for mitigating environmental pollution and optimising wastewater treatment. While purely data-driven models are computationally efficient, they often lack physical interpretation and can produce unrealistic predictions. This study proposes a probabilistic framework using physics-informed Gaussian Process Regression (GPR) to forecast Wastewater Treatment Plant inflows and Combined Sewer Overflows (CSOs). Using high-resolution sensor data from the UWO dataset in Switzerland, a naïve baseline GPR was systematically enhanced by integrating domain knowledge through composite kernel engineering, SWMM-derived prior mean functions, and strict physical output constraints. To overcome the significant computational bottleneck of exact GPs, sparsification was implemented. Furthermore, a novel stratified sparsification method was developed specifically for event-based CSO time-series to optimally allocate inducing points during rare wet-weather surges, as these events do not follow a predictable pattern that can be captured by the basic physics of the model. Results demonstrate that integrating physical constraints significantly reduces predictive error and uncertainty. Operational forecasting of upstream tank dynamics provided a robust, probabilistic early-warning trigger for CSOs that mitigates false alarms. Crucially, the stratified sparse GPR framework reduced execution times by an order of magnitude with minimal accuracy loss, offering a scalable, physics-informed tool for real-time urban drainage management.

Keywords: Flow Prediction, Gaussian Process Regression, Sparsification, Physics-Informed, Sewer Overflow

1. Introduction

Sewer systems are designed and built to safely transfer, treat wastewater and release the treated water into the environment. In urban drainage systems, the stormwater enters the system along with the domestic, commercial and industrial wastewater. The main goal of an urban drainage system is to protect public health by minimising the pollutants coming back to the environment from flooding and untreated discharges (Butler et al., 2024).

Combined sewer overflows (CSOs) are structures installed to spill excess flow into the environment to reduce the risk of backups and manhole overflows in an urban area (Perry et al., 2024). While the flow is diluted after the spill, it still contains a large amount of contaminants (Botturi et al., 2021).

In addition to the contaminants entering the environment from untreated CSOs and manhole releases, wastewater treatment plant (WWTP) effluent and possible bypasses may have a significant amount of pollutants. If the WWTP inflow increases beyond the capacity of the plant, bypasses may occur. Moreover, the increased discharge of wastewater can cause a reduction in the treatment efficiency and an increase in treatment costs (Rezaee and Tabesh, 2022; Suchowska-Kisielewicz and Nowogoński, 2021).

Moreover, water companies must follow regulations in terms of treatment quality and untreated releases (Tiwari et al., 2025; Environment Act, 2021). Therefore, they seek methods to minimise CSO discharges and improve the treatment efficiency by predicting the flow in the drainage system, particularly downstream of the sewer network leading to treatment plants and discharge points.

Models used for predicting flow and other characteristics of a hydraulic system (depth, temperature, concentration, etc.) have been divided into simulators (physical models) and data-driven surrogate models (emulators) (Rezaee et al., 2025; Troutman et al., 2017). The first group of models has been used for more than five decades in this field. They solve a set of partial differential equations (PDEs) for finding the exact answer based on the physical characteristics of the model (Donnelly et al., 2024a).

Taking all these characteristics and physical relations between them makes these models so complicated and computationally inefficient. Therefore, many simplifications should be used for making a physical model capable of solving a real-world problem (Palmitessa et al., 2022). Although these simplifications add to the speed and uncertainty as well, simulators still suffer from a high computational demand.

On the other hand, data-driven surrogate models are capable of making predictions with high speed (Ge et al., 2024). They approximate the response of a computationally demanding model with a faster version (Garzon et al., 2024). These models capture patterns without considering physical interactions between parameters (Aliashrafi et al., 2021). Hence, they lack a physical interpretation which may make them unreliable or challenging to interpret to some extent (Swiler et al., 2020).

Therefore, hybrid models have been developed to take physical constraints into account while using the speed of data-driven models. Aside from speed, enhancement of model accuracy, compensation for missing data and the supplementation of lacking physical relations and parameters are other motivations for using hybrid models (Zhou et al., 2020). In these models,

physics can enter the data-driven models as explicit constraints or loss terms (physics-informed models) (Raissi et al., 2019; Willard et al., 2020) or shape the architecture of the model (physics-guided models) (Karpatne et al., 2017; Palmitessa et al., 2022).

Data-driven methods for modelling physical problems can be divided into two categories: (1) based on variants of artificial neural networks (ANNs) and (2) based on Gaussian processes (GPs) and kernels (Chen et al., 2021). While the first is known for accuracy in recognising patterns and being computationally efficient, the latter is capable of handling complex and small datasets through a natural Bayesian interpretation and accounting for uncertainties in the system (Ding et al., 2023). Hence, it is favourable over ANN methods for solving nonlinear complex equations in a physical system like an urban drainage system with multiple sources of uncertainties such as sensor errors, precipitation, etc. (Thorndahl and Willems, 2008).

Many hybrid models have been implemented in water engineering research, mostly using physics-informed neural networks (PINN). For instance, Li et al. (2024) developed a PINN-based framework to simulate transient free-surface and pressurised flow in a sewer system. They showed high accuracy compared to the finite volume method and Storm Water Management Model (SWMM). Luo et al. (2024) proposed an enhanced PINN model for unsteady hydrodynamics in river networks. Another study was carried out considering a PINN surrogate with discrete mass conservation for simulating inland fluvial floods (Donnelly et al., 2024b). Cedillo et al. (2022) employed PINNs to predict water surface profiles in gradually and rapidly varied flows, including hydraulic jumps, with extension to real-world step-pool channel hydraulics.

While PINN method application in water engineering has been on the rise in recent years, there has been no research showing the implementation of physics-informed (or physics-guided) Gaussian process for predicting urban drainage qualitative and quantitative characteristics, to the best of the authors' knowledge. This shows a need for such models when PINN still lacks a mature uncertainty quantification framework (Di Bella et al., 2026). Without a robust uncertainty quantification, the decision-making and risk-assessment remain unreliable, even in the presence of physical constraints.

To address this limitation in uncertainty quantification, standard GPs have been used for predicting flow, CSO, and concentration in many studies. Mahmoodian et al. (2018) made a GP emulator for predicting CSO events based on the next 90 minutes of precipitation, while Samuelsson et al. (2017) used GP for fault detection in treatment plants. It has also been used in predicting the demand in a water distribution network (Wang et al., 2016) and anomaly detection in a sewer network (Rezaee et al., 2025). Specifically for forecasting applications, Rezaee et al. (2025) used GPR for predicting the inflow entering a WWTP based on time-series data, and Zhang et al. (2024) implemented GPR to predict concentration in shallow groundwater with rare data. Pastrana Cortés (2024) has also implemented a stochastic model for predicting streamflow, using sparse variational GPs.

To further enhance these probabilistic models by embedding physical laws, Raissi et al. (2017) introduced a probabilistic physics-informed model based on GPs. Following their framework, Ye et al. (2024) made a physics-informed GP model, considering the uncertainty of the inputs. Kohanpur et al. (2023) used a physics-informed Gaussian process regression (PIGPR) model to assess the uncertainties in an integrated urban flood model. Balla et al. (2022) applied a GP regression (GPR) model to a laboratory drainage network with constraints on tank water level

and pump flow as inputs. Moreover, this method has been applied in other fields such as mechanics of materials (Rappel et al., 2022), structural health monitoring (Cross et al., 2021), soil moisture estimation (He et al., 2023), fluid mechanics (Padilla-Segarra et al., 2025) and HVAC systems' performance (Zhang et al., 2022).

Although GP is known for its suitability in working with rare and costly data, and providing reliable outputs, its drawback is the $O(N^3)$ computational complexity[1] (Wang, 2023). To overcome the limitation of the computational effort, many methods have been introduced for GP applications for constructing an approximation based on a small set of inducing variables that allow the complexity to reduce from $O(N^3)$ to $O(NM^2)$, when M<<N (Titsias, 2009; Quinonero-Candela and Rasmussen, 2005).

This research aims to propose a framework that implements a GPR model for forecasting flow in an urban drainage system by applying different levels of physics on the data-driven model. The physics enters the model both in the architecture of the designed kernel and as constraints affecting the possible outputs, such as the spectrum introduced by Cross et al. (2024).

Finally, for the sake of computational efficiency, a sparsification method is applied to the proposed framework reducing the runtime and making the computations feasible on an ordinary computer. Afterwards, a novel method of sparsification has been implemented to efficiently handle the event-based time-series, as is common on CSO time-series. In this method, inducing points, which are a sparse set of representative data points, are distributed across the time-series in a smart way. Instead of distributing inducing points evenly across the entire period, this method explicitly allocates a dedicated number of inducing points to the short, critical event periods. The remaining inducing points are then allocated to the baseline periods, where values remain near the mean without sudden changes. This targeted distribution remarkably reduces the overall runtime of the models.

The remainder of this paper is organised as follows. Section 2 presents the case study and the data used for training and testing the model. Also, the method for implementing the GPR model and the incorporation of physical aspects to the data-driven model is explained. In Section 3, results from different models are presented and a comparison between models is made to demonstrate the strengths of different models across various aspects of prediction. Also, the novel sparsification method is tested and the results are discussed in this section. The discussion is presented alongside the results for better cohesion of the material. Finally, the conclusion of this research with suggestions for future work is discussed in the last section.

2. Materials and Methods

2.1. Case Study and Data

The case study focuses on the Urban Water Observatory (UWO), a full-scale field laboratory situated in the municipality of Fehraltorf, Switzerland. The monitoring infrastructure tracks urban hydrology at high spatiotemporal scales via a coherent network of 124 sensors, which includes 14 precipitation and meteorological sources, 70 hydraulic sensors (measuring flow, depth, and overflow activity), and 40 sewer atmosphere temperature probes. To ensure resilient data collection from harsh underground environments, 89 of these sensor nodes transmit signals

[1] Computational complexity describes how the runtime of an algorithm scales with the size of the input data (N). An $O(N^3)$ complexity means the computation time increases with the cube of the number of training points.

wirelessly via long-range, low-power LoRaWAN mesh technology. For the three consecutive years of continuous recordings analysed in this study (2019–2021), the core hydraulic and precipitation data channels demonstrate an exceptional completeness rate exceeding 99% after initial quality control, offering a highly reliable foundation for probabilistic modelling (Blumensaat et al., 2025).

Beyond physical data collection, the UWO serves as a robust foundation for advanced hydrodynamic modelling and applied research. The physical sewer network is modelled using EPA SWMM, which is continuously calibrated to reflect complex, real-world dynamics such as seasonal groundwater infiltration. Ultimately, this open-access field laboratory empowers the scientific community to tackle major urban drainage challenges, such as testing automated anomaly detection with machine learning, quantifying extraneous water, and accurately assessing the environmental impact of CSOs.

This research utilises data from the downstream section of the network leading to the WWTP and the CSO structure. The main flow enters this section through two chambers, one mainly responsible for dry-weather flow (DWF) and the other responsible for spike flows due to rainfall. The flow is then directed to the treatment plant through a controlled inlet with a throttle, limiting the flow to 180 L/s.

The excess flow is routed to the main storage facility and the in-sewer storage volume which acts as the CSO structure and spills into the adjacent river during overflow events. The presence of multiple interacting overflow weirs and retention structures in this terminal section significantly increases the complexity of the network's hydraulic modelling. The image of the aforementioned part of the system, taken from the hydraulic model, is shown in Figure 1.

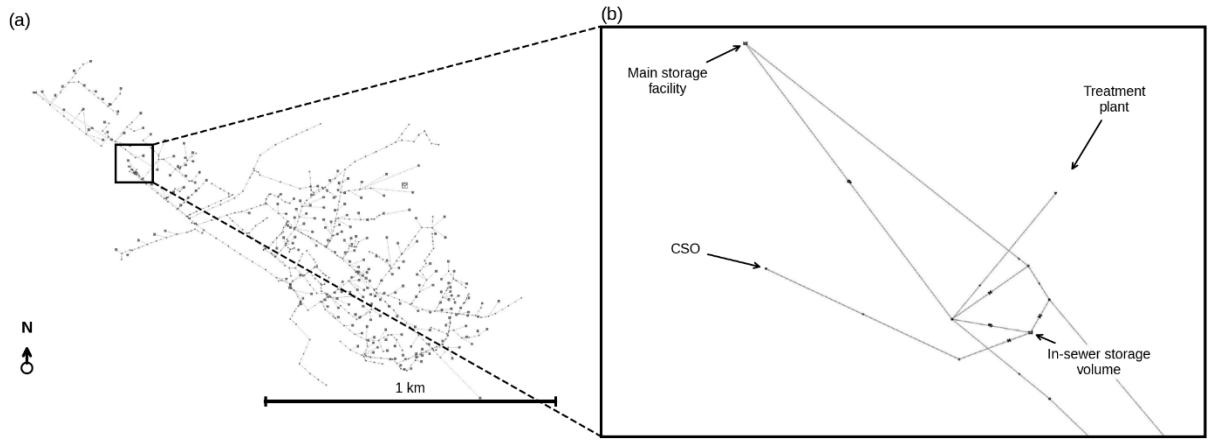

Figure 1. (a) skeleton drainage system of Fehraltorf, Switzerland and (b) downstream section of the system leading to the WWTP and the CSO structure.

High-resolution monitoring data of the flow entering the WWTP, the tank water level of the in-sewer storage structure and CSO flow are recorded at 1-minute intervals. Primary precipitation data were gathered from the highly consistent city centre school rain gauge. Minor data gaps accounting for less than 2% of the study window were compensated for using a secondary rain gauge situated at the treatment plant (1 km distance). For more information regarding the sensor details, readers are advised to consult the documentation of the UWO dataset.

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 (Rasmussen and Williams, 2006). In the GPR method, a GP is set as the prior for the latent function (***f(x)***) that maps the inputs into the output space and then updates the prior with the observations, using Bayes' rule (MacKay, 2003). Using the GP, we can define a distribution over functions ***f(x)***,

$$f(\boldsymbol{x}) \sim GP(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')) \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{x}'$ are two GP input vectors and $\boldsymbol{x} \in \mathbb{R}^N$ ($\mathbb{R}$ represents the set of real numbers). Also, m(.) and k(.,.) are mean and covariance functions that specify the GP (Deshpande et al., 2025).

Let $\boldsymbol{X}$ and $\boldsymbol{X}_*$ denote the training and test matrices and $K(\boldsymbol{X}_*, \boldsymbol{X})$ denotes the covariance matrix evaluated at all data points, the key predictive equations for GPR when the observations contain noise are as follows:

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

$$cov(\boldsymbol{f}_*) = K(\boldsymbol{X}_*, \boldsymbol{X}_*) - K(\boldsymbol{X}_*, \boldsymbol{X})[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 \mathrm{I}]^{-1}K(\boldsymbol{X}, \boldsymbol{X}_*) \tag{3}$$

where $\mathbf{f}_*$ is the predicted test output according to the prior and $\bar{\mathbf{f}}_*$ is the mean of the prediction on the test points, $\boldsymbol{X}_*$. In Equation 2, the observation values $\mathbf{y} = \mathbf{f}(\boldsymbol{x}) + \mathcal{E}$ where $\mathcal{E}$ is Gaussian noise with variance σ_n^2.

GPR parameters are often optimised by maximising the log marginal likelihood. Among parameters are hyperparameters which are free parameters controlling the shape and behaviour of the kernels.

$$log\ p(\boldsymbol{y} \mid \boldsymbol{X}) = -\frac{1}{2}\boldsymbol{y}^T(K + \sigma_n^2 \mathrm{I})^{-1}\boldsymbol{y} - \frac{1}{2}log|K + \sigma_n^2 \mathrm{I}| - \frac{n}{2}log\ 2\pi. \tag{4}$$

In the log marginal likelihood equation, $|K + \sigma_n^2 I|$ denotes the determinant of the covariance matrix containing noise.

2.3 Forecasting WWTP Inflow

As mentioned previously, forecasting the inflow of a WWTP is of importance. This subsection describes the forecasting framework. In each section, physics is added to the model to make it closer to the real-world problem. All coding is made in Python and the GPR model is constructed using the GPflow package in Python (Matthews et al., 2017).

The data are split into training and test sets, where the test data represent the forecasting period. The training dataset is entered into the model afterwards, and a kernel (single or composite) is applied to the data, and the hyperparameters of the model are obtained through an optimisation method (here a quasi-Newton method).

Moreover, as the data points are at 1-minute intervals, it is redundant to use all data points for training the model. Therefore, the mean value of a selected period is put into the model as an input. Then the predictions are made based on those time intervals.

2.3.1. naïve data-driven baseline

At first, a baseline model for predicting WWTP inflow is developed without any constraints or preference for the kernel design. This model is made with two primary inputs: time and precipitation. These inputs are believed to have the most impact on the change in flow entering the WWTP. The model can take more attributes as inputs (e.g., land use, geometry, etc.), but each attribute adds to the complexity and computational effort of the model.

As the precipitation has a lagged effect on the WWTP inflow because of the travel time through the network, it is better to consider it as an accumulated rainfall over a time period before the data point of interest. Finding this time frame is done in this research through a cross-correlation analysis between flow and rainfall. This analysis is performed on 15-minute intervals from zero to two hours. Then the best lagged time is selected and the accumulated rainfall in that period is selected as the input of the GPR model.

In this step, a simple kernel (such as Squared Exponential (SE)) is selected with no bounds on hyperparameters. Also, no mean function is assigned to the model which means the model automatically sets it on zero.

2.3.2. kernel design

From an overview of water usage, and consequently wastewater production patterns, it can be realised that the wastewater flow follows some basic patterns that can be captured within the GPR model by designing the kernel. Here, it is assumed that the daily fluctuations of the water usage will lead to a periodic wastewater flow pattern.

Also, groundwater infiltration can constitute a major part of the overall flow, directed to the WWTP across different seasons (Zeydalinejad et al., 2024). Besides groundwater infiltration, water usage habits change during different months of the year, leading to a long-term pattern of the overall flow (Rezaee and Tabesh, 2022).

Therefore, the designed kernel should account for both short- and long-term periods to reflect the wastewater pattern entering the WWTP. Moreover, it should take the spike flows caused by rainfall. These added flows have no long-term pattern as the rainfall intensity and duration are not normally distributed in a year with particular patterns. Consequently, a kernel should be added to the composite kernel to account for that event-based additive flow. To this end, the composite kernel formula comes as Equation 5.

$$k_{WWTP}(\boldsymbol{x}, \boldsymbol{x}') = [k_{Per}(\boldsymbol{x}, \boldsymbol{x}') \times k_{SE}(\boldsymbol{x}, \boldsymbol{x}')] + k_{Mat1/2}(\boldsymbol{x}, \boldsymbol{x}'), \quad (5)$$

where k_{WWTP} is the designed composite kernel, k_{Per} is a periodic kernel capturing short-term patterns, k_{SE} is an SE kernel capturing long-term changes and $k_{Mat1/2}$ is a Matérn1/2 kernel taking sudden wet-weather changes. $\boldsymbol{x}$ and $\boldsymbol{x}'$ are as defined in Equation 1. As these individual base kernels are standard, off-the-shelf covariance functions, their full mathematical formulations are omitted here for brevity. For comprehensive theoretical definitions, readers are referred to Wilson (2014), while their specific algorithmic implementations are detailed in the GPflow documentation (Matthews et al., 2017).

2.3.3. mean function design

Another physical information that can be added to the model and enhance its accuracy is a prior mean function. As the mean function will dominate the forecasts far from the training data period, it can have a major effect on the predictions, so it must be chosen with consideration (Roberts et al., 2013).

Finding a universal gold standard mean function for an urban drainage system across all seasons with many uncertain influencing parameters is almost impossible. However, the mean function can be defined for dry-weather conditions to mimic the DWF. The kernel handles the residuals caused by rainfall and other system uncertainties afterwards.

In this research, the calibrated hydraulic model made in EPA SWMM has been used to determine the baseline DWF. The precipitation is set to zero and the model is run for a 7-day period. Then the values entering the WWTP have been captured and averaged to form a 24-hour time-series. The interval between the data points is set to 3 minutes and based on the chosen intervals for the main GPR model, the data points are resampled by selecting the nearest available value.

To scale this pattern to match the observed data, the SWMM daily pattern was normalised by multiplying by the ratio of the average inflow of the training data to the pattern's own average. This ensures that the mean function maintains the SWMM's diurnal shape while matching the magnitude of observed flows.

2.3.4. adding constraints

The main constraints imposed on the output of the model are minimum and maximum possible values. As the output of the GPR model is a Gaussian distribution, it takes all values in (-∞, +∞). Therefore, if the output is not bounded by the constraints, it will have both negative values, which are impossible in a normal working system, and positive values beyond the 180 L/s flow limit of the throttle that sends the flow to the WWTP, which is impossible as well when the throttle is working properly.

Consequently, the [0, 180] L/s bound should be applied on the possible outputs of the model. This has been done through two major methods: Truncated Posterior function and Gaussian process Warping.

- Truncated Gaussian distribution:

Different techniques such as the Genz approximation (numerical multivariate integration) and Gibbs sampling (Markov chain Monte Carlo) have been studied for bounding the output of a normal distribution (Da Veiga and Marrel, 2012). However, for a computational speedup, here we apply the simple method of bound constraints to clip out-of-bound values and multiply the remaining probability density function (PDF) by the inverse value of the probability between the bounds.

For bound constraint vectors **a** and **b** in $\mathbb{R}^D$, the PDF of the truncated Gaussian vector **z** is given by:

$$\phi_{\mu,\Sigma,\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{z}) = \begin{cases} \dfrac{\phi_{\mu,\Sigma}(\boldsymbol{z})}{\mathbb{P}(\boldsymbol{a} \le \boldsymbol{z} \le \boldsymbol{b})}, & for\ \boldsymbol{a} \le \boldsymbol{z} \le \boldsymbol{b} \\ 0, & otherwise. \end{cases} \tag{6}$$

where $\phi_{\mu,\Sigma}(\mathbf{z})$ denotes the PDF of the multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ with μ as the mean vector and Σ the positive-definite covariate matrix. Also, $\mathbb{P}$ shows the probability function.

- Warping:

In probabilistic modelling, warping is a transformation that reshapes target distributions, allowing physically constrained or non-Gaussian data to satisfy standard Gaussian assumptions. Various warping processes have been studied and implemented for bounding GPs (Rios and Tobar, 2019; Snelson et al., 2003). In this work, a logit-sigmoid warping transformation is applied to map the bounded space to an unbounded latent space where Gaussian posterior assumptions are valid (Snelson et al., 2003). The warped latent variables are then standardised to be injected into the GPR model, where the kernel hyperparameters are optimised by maximising the log marginal likelihood.

Subsequently, predictions are obtained in the scaled latent space, and descaling should be performed to bring it back to the original scale. This process is completed by the sigmoid function; however, the results cannot be analytically brought back to the original space because of the nonlinearity of the sigmoid function. Therefore, numerical methods are required to approximate the intractable integral. In this work, two approaches are tested and compared: Monte Carlo sampling and Gauss-Hermite quadrature.

2.4. Forecasting CSO

Predicting CSOs is important for making proactive decisions and for warning the public (Rosin et al., 2021). Many physical and data-driven models have been developed for predicting CSOs, mostly using neural networks as the main architectures (Rosin et al., 2017; Mounce et al., 2014).

However, due to approximations made in the process of modelling and omitted physical relations, the output of such models will contain a significant amount of uncertainty (Kohanpour et al., 2023). This can be addressed by implementing a probabilistic method like GPR that is not only capable of predicting future values but also capable of finding the probability of CSO occurrences.

In this case study, the CSO is caused by the overflow of the in-sewer storage facility, shown in Figure 1. Consequently, two approaches have been studied for estimating CSO values and the possibility of happening. The first one takes CSO flow values as time-series and predicts future values based on the previous CSOs. The other predicts the occurrence based on the probability of overflow of the upstream storage tank. Each method has its advantages and disadvantages that make them suitable for different applications in the real world.

2.4.1. predicting CSO with CSO values

In this approach, the CSO values are set as the output of the model, and the inputs are time and rainfall. However, as the CSO is a direct result of tank overflow, there is a significant effect of previous rainfalls on the state of the tank at the moment of the new rainfall. Therefore, rainfall

is imported into the model as two different attributes. One as the current rainfall, taking the accumulated rainfall over a maximum of the last two hours (having the same cross-correlation approach for finding the optimum lag time). The other is the accumulated rainfall over the last 24 hours.

Below are the physics that are added step by step after constructing a naïve GPR model.

- Kernel design

As the CSO time-series show an event-based time-series, the kernel should be designed to take both dry-weather zero values and wet-weather surges. Therefore, a smooth time kernel like SE takes care of the base flow, then two Matérn kernels are designed for rainfall events to capture the instant spikes. Equation 7 can be used for the composite kernel used in CSO prediction.

$$k_{CSO}(\boldsymbol{x}, \boldsymbol{x}') = k_{SE}(\boldsymbol{x}, \boldsymbol{x}') + k_{Mat1/2}(\boldsymbol{x}, \boldsymbol{x}') + k_{Mat5/2}(\boldsymbol{x}, \boldsymbol{x}'), \quad (7)$$

where k_{CSO} shows the composite kernel designed for predicting CSO, k_{SE} is an SE kernel taking the base flow. Also, $k_{Mat1/2}$ and $k_{Mat5/2}$ are Matérn1/2 and Matérn5/2 kernels capturing wet-weather characteristics discussed above.

- Mean function

The mean function should be picked for dry-weather situations as the rain events are completely uncertain. And the CSO values in dry-weather periods are zero, unless some operational problems happen to the system. Therefore, a constant zero mean function is added to the model.

- Constraints

The minimum value of the CSO is zero like the WWTP inflow. However, the maximum value can go up without any limit. So, the Gaussian output of the model should be clipped on one side to show the bound.

2.4.2. predicting CSO with tank level values

To find the probability of CSO occurrence, we look at the probability of overflow in the upstream storage facility (in-sewer storage volume in Figure 1). The overflow happens when the water level exceeds the 3.2 m weir crest threshold in this facility. Therefore, by developing a model that predicts the water level in the storage tank, the probability of exceeding the 3.2 m threshold translates directly to the CSO probability.

During wet weather, this storage facility is fed by excess wastewater overflowing an in-sewer distribution chamber. The retained water is continuously emptied by a controlled structure. If the maximum storage capacity is exceeded, the resulting CSO follows the weir equation and is proportional to $H^{1.5}$ (where H is the water head above the weir's crest to the power of 1.5) (Butler et al., 2024).

After developing a GPR model that takes the rain and time as the main inputs and the water level as the output, the following physical aspects are inserted into the model:

- Kernel design

A time kernel is designed to capture the daily variations of the water level. A rain kernel is also added to account for surges in the water level after rainfall. As the daily variations are minimal due to the buffer effect of the main storage facility, a SE kernel works perfectly to this end. For the rain kernel, a Matérn1/2 kernel that deals with spikes better than other kernel types is utilised and the formula of the designed kernel comes as:

$$k_{Tank}(\boldsymbol{x}, \boldsymbol{x}') = k_{SE}(\boldsymbol{x}, \boldsymbol{x}') + k_{Mat1/2}(\boldsymbol{x}, \boldsymbol{x}'), \tag{8}$$

where k_{SE} takes the base level with an SE kernel, and $k_{Mat1/2}$ captures the wet-weather levels with a Matérn1/2 kernel.

- Mean function

A mean value of 100 mm is set as the constant mean function. This value represents the remaining water in the tank during dry periods.

- Constraints

The physical minimum and maximum values for the tank level are bounded between 0 and 3800 mm. Therefore, the Gaussian output of the model is clipped to prevent out-of-bound predictions.

2.5. Sparsification

The high computational demand of the standard GP is due to the covariance matrix (K_{NN}) inversion, with size N×N. Using M inducing points, a low rank approximation (Q_{NN}) is made, and the M×M matrix is inverted instead (Hensman et al., 2013).

$$K_{NN} \approx Q_{NN} = K_{NM} K_{MM}^{-1} K_{MN} \tag{9}$$

This reduces computational complexity from $O(N^3)$ to $O(NM^2)$, and storage from $O(N^2)$ to $O(NM)$ using variational inference.

Rather than optimising the exact marginal likelihood of the N input vectors, the Sparse GPR (SGPR) model optimises a variational lower bound on the marginal likelihood, known as the Evidence Lower Bound (ELBO). ELBO should be maximised in the model to ensure that the sparse approximation is as close as possible to the main GP model. This is performed by applying Kullback-Leibler divergence to the approximation (Q) and the GP posterior (P).

The main equations can be written as (Titsias, 2009):

$$\log p(y) = ELBO + KL[Q||P] \tag{10}$$

$$ELBO = \log[N(y|0,\ \sigma^2 I + Q_{NN})] - \frac{Tr(K_{NN} - Q_{NN})}{2\sigma^2}, \tag{11}$$

where σ^2 represents the observation noise variance and y is the set of observations. Also, Tr() is the trace function, and I is the identity matrix with size N×N.

During model training, the optimiser maximises the ELBO with respect to the kernel hyperparameters. If defined, it can also optimise the location of the inducing points.

Although the location of the inducing points can be optimised by the optimiser, the initial location of the inducing points has a significant effect on the efficiency of the optimisation. A naïve method for spreading inducing points is to uniformly select them from the training dataset (Burt et al., 2020). However, other more efficient methods have been applied for this purpose.

In this research, the classic method of distributing inducing points using k-means clustering has been applied, similar to Hensman et al. (2013). Moreover, a novel stratified sparsification method is implemented for event-based time-series (such as CSO and tank level time-series) to distribute inducing points to areas with more valuable data.

2.5.1. standard SGPR

k-means is one of the simplest clustering methods that classifies the data into a certain number of clusters by relying on data density to iteratively position centroids that represent the most common states of the system (Kodinariya and Makwana, 2013). Because the algorithm minimises the spatial variance between data points and their nearest centroid, it inherently allocates the majority of its clusters to regions where the data are most heavily concentrated.

In standard SGPR, the locations of the inducing points selected by the k-means algorithm serve as the initial coordinates, which are subsequently optimised during model training. This density-driven approach proves highly effective for modelling continuous, diurnal patterns, such as WWTP inflows. However, when applied to event-based time-series, such as rainfall-driven runoff, a critical computational flaw emerges: a vast majority of the inducing points are squandered on the flat, zero-variance regions under dry-weather conditions, leaving the complex wet-weather surges drastically underrepresented.

2.5.2. stratified SGPR

In event-based systems, such as CSO structures or retention facilities, the time-series typically exhibits a constant baseline for extended dry-weather periods, interrupted only by highly dynamic, short-duration rainfall events. Therefore, allocating the majority of the computational budget (inducing points) to these dry periods is highly inefficient.

To resolve this, the time-series is partitioned into two distinct physical regimes (dry and wet periods) by classifying the flow values into clusters below and above the dataset's average. Because the baseline dry-weather data constitutes more than 90% of the total dataset, a standard k-means algorithm would naturally force over 90% of the inducing points into this inactive region.

In the proposed stratified method, the total budget of inducing points is explicitly divided: a dedicated share is forced into the rare, active wet periods to capture the spike dynamics, while the remainder is allocated to the baseline dry periods. This targeted allocation drastically reduces the total number of inducing points required to achieve high accuracy. Furthermore, during optimisation, the sparse inducing points allocated to the dry periods naturally fall back on the prior mean function (which is explicitly set to the dry-weather condition, for instance, zero for CSO), ensuring stability without wasting computational resources.

2.6. Evaluation Metrics

To evaluate how close the predicted values are to the actual measured values, we need to use some metrics to assess the fit of the model. Different methods exist for showing how well the

model learns patterns from the training data and makes predictions. These methods include both statistical measures (Malde, 2018) and graphical checks (Gelman et al., 2021).

In this work, root mean square error (RMSE) and mean absolute error (MAE) are selected to show the difference between the predictions and the actual value. Moreover, Coverage and Entropy are two other metrics defined to show how the probabilistic output of the GPR model is acting. These metrics 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 (Rezaee et al., 2025).

2.6.1. RMSE

RMSE measures the accuracy of the predictions, using the following equation:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}\,, \tag{12}$$

where N is the number of data points, y_i is the actual value and $\hat{y}_i$ is the predicted value from the model. As it can be obtained from the equation, outliers penalise the model more (Hyndman and Koehler, 2006) which makes this measure the primary measure of this study to prevent missing spike flows or water levels that would cause more problems for the system.

2.6.2. MAE

While RMSE accounts for outliers more, MAE linearly considers errors of the predictions, using the equation below:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|\,. \tag{13}$$

MAE acts as a secondary error function in this work as recommended by Hodson (2022).

2.6.3. Coverage

The Coverage metric shows the percentage of the data that falls within the defined credible intervals (CIs). In this work, a 95% CI is used as the target interval, and the coverage is calculated based on that.

2.6.4. Entropy

If the model is chosen with only error metrics and Coverage, the best model will have very wide bounds that includes all actual values within the bounds and shows the maximum coverage. Therefore, another metric that makes a trade-off between the coverage and the uncertainty in the predictions is needed.

Entropy quantifies the uncertainty associated with the predicted distribution, with higher entropy showing greater uncertainty (Hasegawa and Nishiyama, 2025). The overall differential entropy (H) can be defined as:

$$H = \frac{1}{N}\sum_{i=1}^{N} 0.5\log(2\pi e\sigma^2)\,. \tag{14}$$

Here, σ denotes the standard deviation of each test point, indicating how narrow or wide the bounds are defined.

It is important to note that the differential entropy applied in this study differs slightly from the Shannon entropy defined for discrete distributions. In contrast to Shannon entropy, differential entropy may assume negative values when the data variance is very small, as it is sensitive to the units of measurement. Consequently, its values cannot be directly compared between models. Nevertheless, when the training and testing datasets are kept consistent, differential entropy remains a useful indicator for assessing which kernel configuration yields predictions with lower overall uncertainty (Rezaee et al., 2025).

3. Results

3.1. Physics-Informed WWTP Inflow Prediction

As mentioned in Section 2.1, the dataset covers three consecutive years starting in 2019. Therefore, to ensure a fair comparison among models, two different time-series from two different seasons of the year have been selected to show the various effects of wastewater production and groundwater infiltration.

The first prediction period starts at 30/10/2019 for five days of flow prediction. The other starts on 10/05/2021. Between these two periods (in August 2020) there has been a major upgrade in the upstream system of the WWTP particularly in the flow measurement system. Therefore, the data in 2019 may contain errors and anomalies due to the old measurement system.

The time-series of the two selected periods and their previous 30 days are plotted in Figure 2. The time interval between data points is one minute. However, to reduce the computational effort of the data-driven models, each 15-minute interval is used as a data point and the inflow data within each timestep are averaged. The precipitation is also summed over the timestep to show the overall rainfall.

Based on the described method in Section 2, a baseline model should be established on the time-series as the Naïve model. This model uses a GPR model and takes the hyperparameters of the kernel by optimising the log marginal likelihood. The kernel is a simple SE kernel that has no bounds on its lengthscale and variance.

In the next step, the kernel is designed based on the characteristics of the flow. It contains a daily kernel composed of a periodic kernel that captures the daily fluctuations. There is also a long-term kernel responsible for long-term changes in the flow which is defined in the model by a SE kernel type. These two kernels are multiplied to create a locally periodic kernel, allowing the model to capture the 24-hour cycle while permitting its amplitude to dynamically evolve over long-term seasonal shifts.

Another kernel is added to this product to add the characteristic of the wet-weather flow added to the DWF. It should take the spike characteristics of the wet-weather flow due to rain. So, a Matérn1/2 kernel is used to define this kernel. All these kernels receive bounds on their lengthscales to prevent unrealistic smoothness or wiggliness. Also, the variance value of each kernel is initially set close to the reality so that the optimiser does not search for unrealistic values.

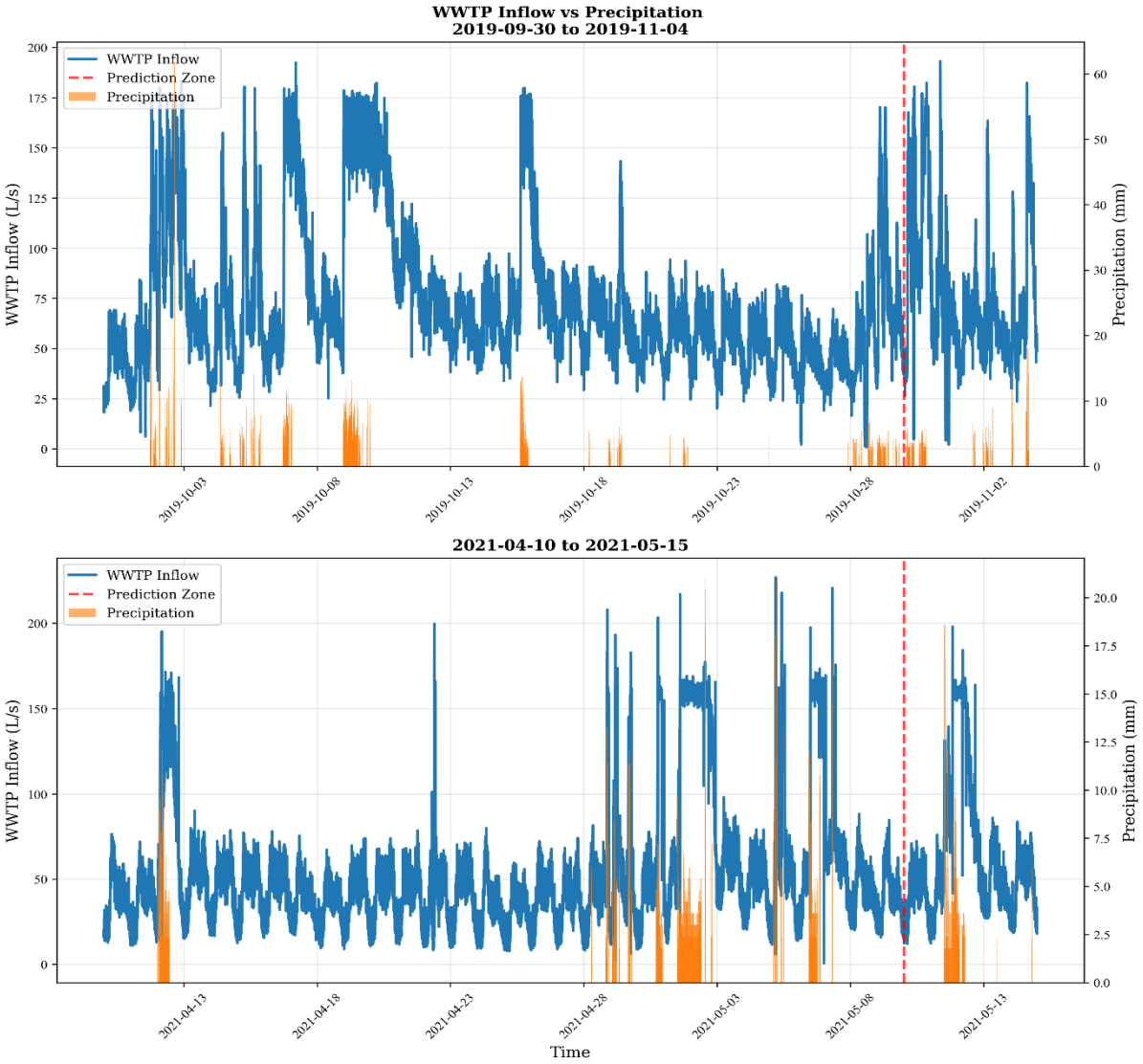

Figure 2. WWTP inflow time-series vs precipitation. The top and bottom plots respectively show the time-series for the year 2019 and 2021. The red dashed line shows the start date of the 5-day period of forecasting. The rest of the plot shows the 30 days period prior to the forecasting region.

After making the models based on the designed kernel, a mean function is added to the models to represent the DWF. The values obtained from the simulator should be transformed into a standardised form that the model is able to handle. Afterwards, the constraints mentioned in Section 2 are added to the output values.

Both Truncated GP and Warping have been implemented on three random time-series. Empirical testing indicated that the Truncated GP approach not only shows higher computational efficiency in comparison with the Warping method but also performs slightly better in reducing error and covering more points.

Figures 3 and 4 show the predictions made based on each four levels of physical aspects added to the GPR model. In these models, the prediction period starts exactly at the end of the training period and forecasts the WWTP inflow for the next five days. A 95% CI is used to show the possibility around the predictions.

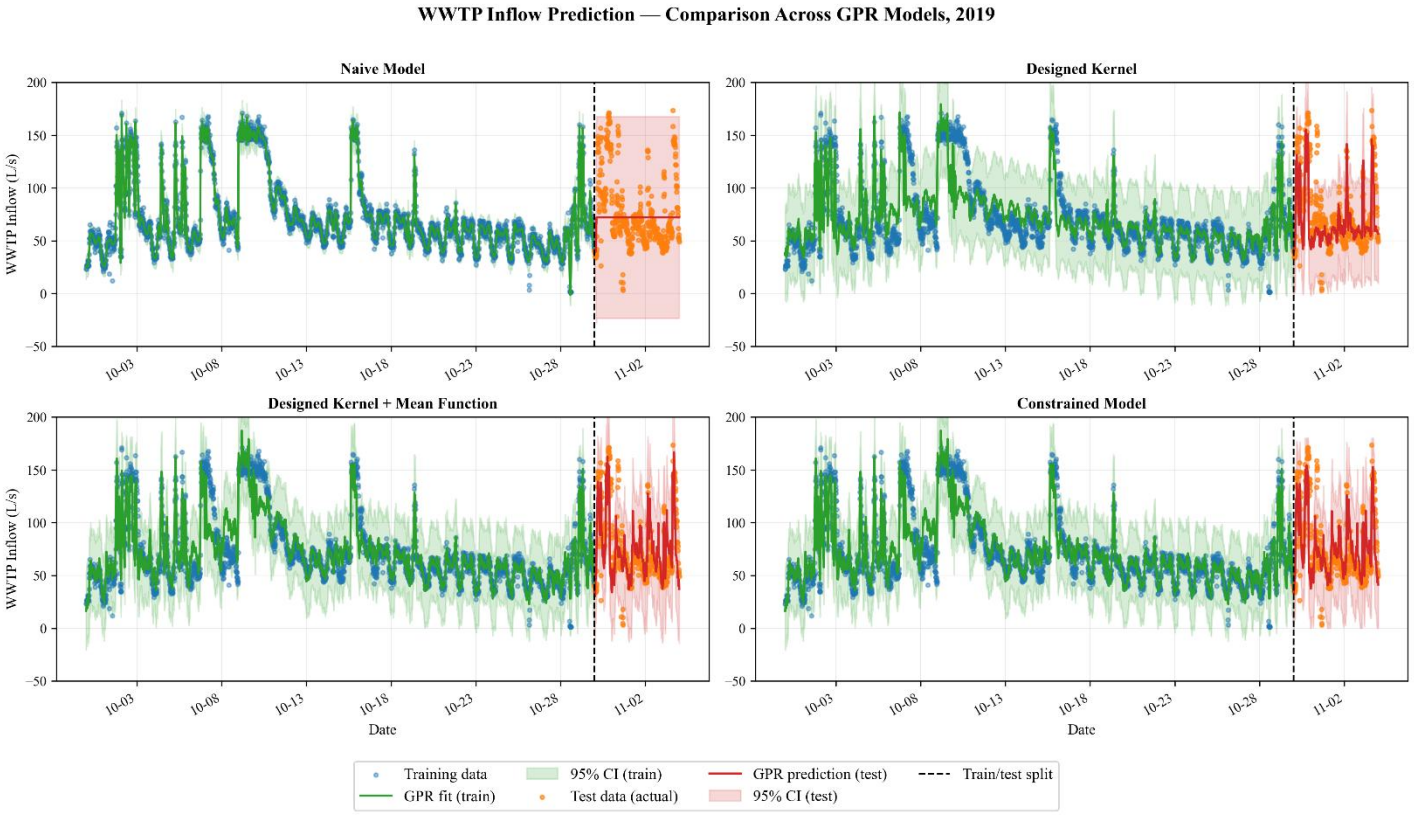

Figure 3. Comparison of GPR flow forecasting models with different levels of enforced physics: purely data-driven model (top left), model with a designed kernel based on the physical information (top right), SWMM-based mean function model (bottom left), and a fully physics-informed model enforcing output constraints (bottom right). The time-series are similar in the year 2019, and the predictions are made on the first 5 days after the training time ends (where the vertical dashed line is drawn). 95% credible intervals are shown around both training and test GPR lines.

The output of the model is a Gaussian distribution at each prediction point so the probability of having values above or below a certain flow can be calculated in each point. An example of the probability density at a point is shown in Figure 5. This point shows the predicted point 12 hours after the end of the training period produced by the constrained model.

As time goes by, predictions become less accurate and the model makes wider bounds around the prediction (test) points to cover the uncertainty caused by the distance from the training data points. Therefore, for longer prediction windows, physical information and constraints help the model avoid deviating from the actual values. Table 1 shows the effectiveness of the physics in the models using the evaluation metrics.

Table 1. Evaluation of different models for two time-series used in WWTP inflow prediction.

Model type	**Naïve**		**Designed Kernel**		**Designed Kernel + Mean function**		**Constrained**	
Year	**2019**	**2021**	**2019**	**2021**	**2019**	**2021**	**2019**	**2021**
RMSE (L/s)	34.299	41.749	25.813	32.120	25.654	33.388	25.568	29.975
MAE (L/s)	24.901	27.161	17.397	20.096	17.672	19.230	17.630	18.037
Coverage (%)	99.17	100.0	92.52	88.15	93.35	88.15	92.93	87.94
Entropy (nats)	7.6116	8.0631	6.5944	6.2887	6.6724	6.2664	6.6215	6.1559

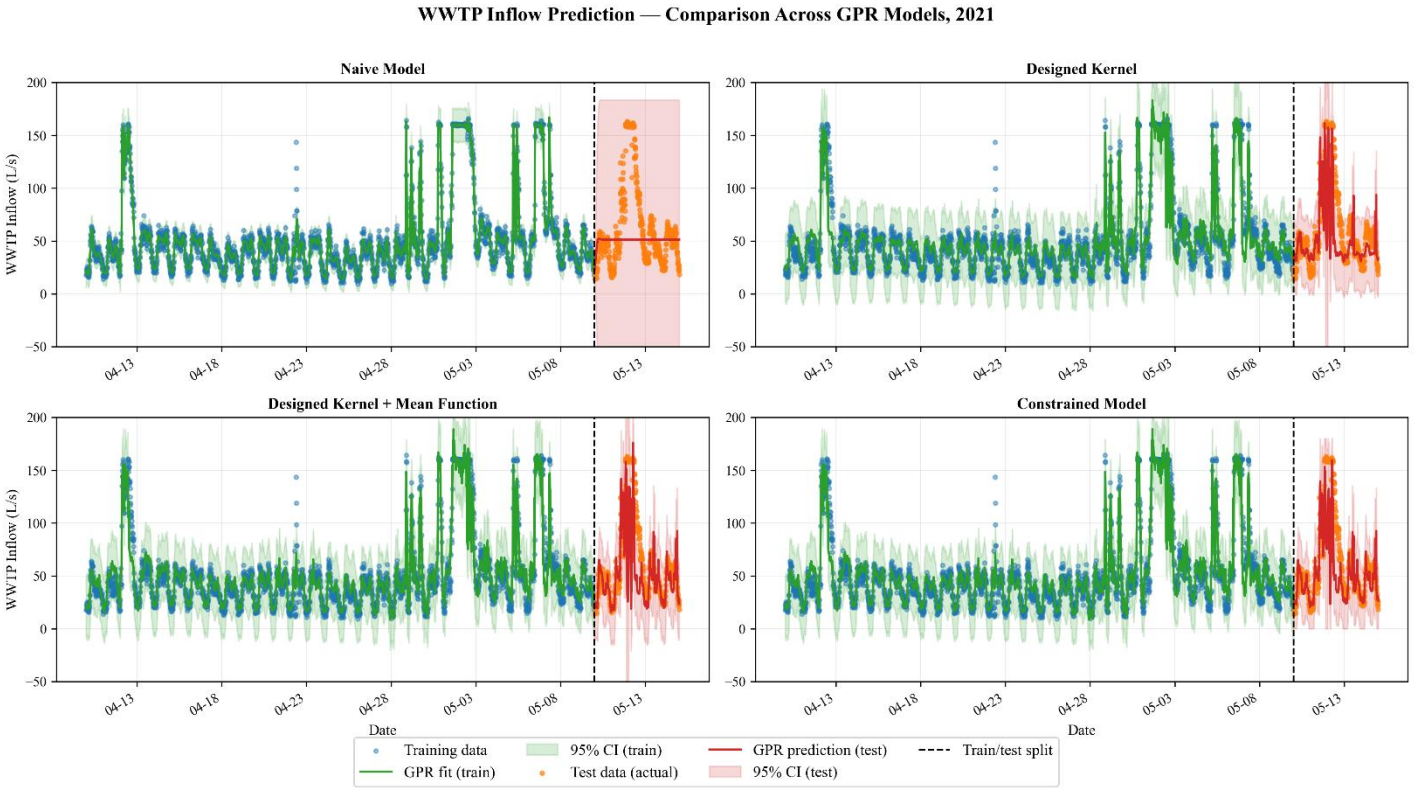

Figure 4. Comparison of GPR flow forecasting models with different levels of enforced physics: purely data-driven model (top left), model with a designed kernel based on the physical information (top right), SWMM-based mean function model (bottom left), and a fully physics-informed model enforcing output constraints (bottom right). The time-series are similar in the year 2021, and the predictions are made on the first 5 days after the training time ends (where the vertical dashed line is drawn). 95% credible intervals are shown around both training and test GPR lines.

As detailed in Table 1, the sequential addition of physics improved the accuracy of predictions across both testing periods. For example, in the 2019 time-series, RMSE dropped from 34.3 L/s to 25.6 L/s by designing the kernel, adding the mean function and bounding the output probability function. In addition to the error reduction, the uncertainty of the predictions is also reduced. Entropy in the Naïve model shows a high level of uncertainty, which has been reduced by 25% and 28% on the Constrained model.

However, coverage dropped when the pattern was captured by the model, as shown in Figures 3 and 4. After designing the kernel, it remained almost the same by adding to the physical knowledge of the models which shows that the models maintained acceptable coverage while reducing errors and uncertainty.

3.2. Physics-Informed Sparse GPR Prediction

Since adding physical knowledge and constraints to the data-driven model added to the computational effort, sparsification effectively mitigated this. In the time-series considered for prediction in this study, 2880 data points were used for training the model. In the sparse model, 100 inducing points were used to reduce the computational effort of the model.

The results are provided in Figure 6, showing the sparse version for the Constrained model. The 100 inducing points were distributed through the training period and beyond it using the

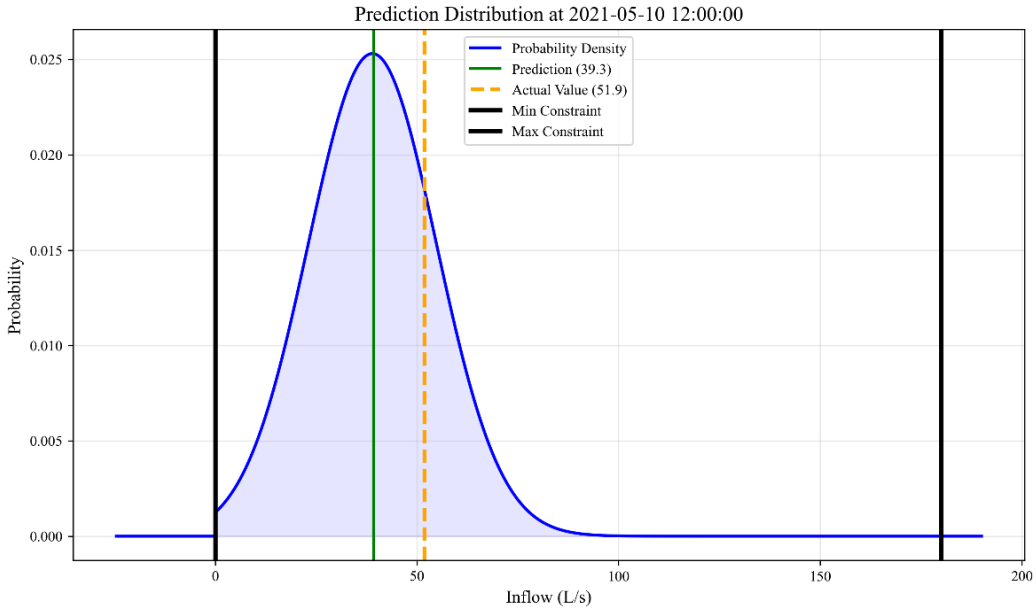

Figure 5. Probability density function of the predicted WWTP inflow at a 12-hour forecasting horizon. The Gaussian distribution is displayed within ±4 standard deviation for a better illustration. Enforced physical boundary constraints are indicated, with the probability strictly truncated to zero outside the feasible domain. The observed flow value successfully falls within the high-density region of the prediction.

k-means method. The locations of the points were later optimised with the hyperparameters of the model (via maximising ELBO) and final locations are shown on the plots.

The runtimes of the models have reduced significantly, with the total execution times reduced to less than 10% of the original model's runtime. Crucially, this efficiency was achieved with minimal degradation in predictive accuracy. The evaluation metrics show that the RMSE increased by 1.3% and 3.9% for the 2019 and 2021 test periods, respectively. All other metrics deviated by less than 3% compared to the Constrained model. As visually confirmed in Figure 6, the SGPR model successfully retains the physical fidelity of the exact model, accurately capturing both the diurnal dry-weather variations and the sudden spikes induced by rainfall.

3.3. Physics-Informed CSO Prediction

To evaluate how GPR models perform in CSO prediction with and without physical knowledge, the same two time-series used for WWTP inflow predictions have been used. In 2019 time-series, three CSO events occur in the 30-day training period and no CSO occurs in the prediction period. So, the model should be able to show the fewest false alarms in the forecasts. On the other hand, the 2021 period contains two small CSOs in the training period and a significant CSO in the five-day prediction period. Therefore, the model should predict the magnitude and timing of the CSO properly.

The results show that the Naïve model fails to capture the 2021 CSO event and produces many false predictions. However, with the addition of physical aspects to the model, the model starts to perform better in both time-series. The results of the evaluation metrics are presented in Table 2.

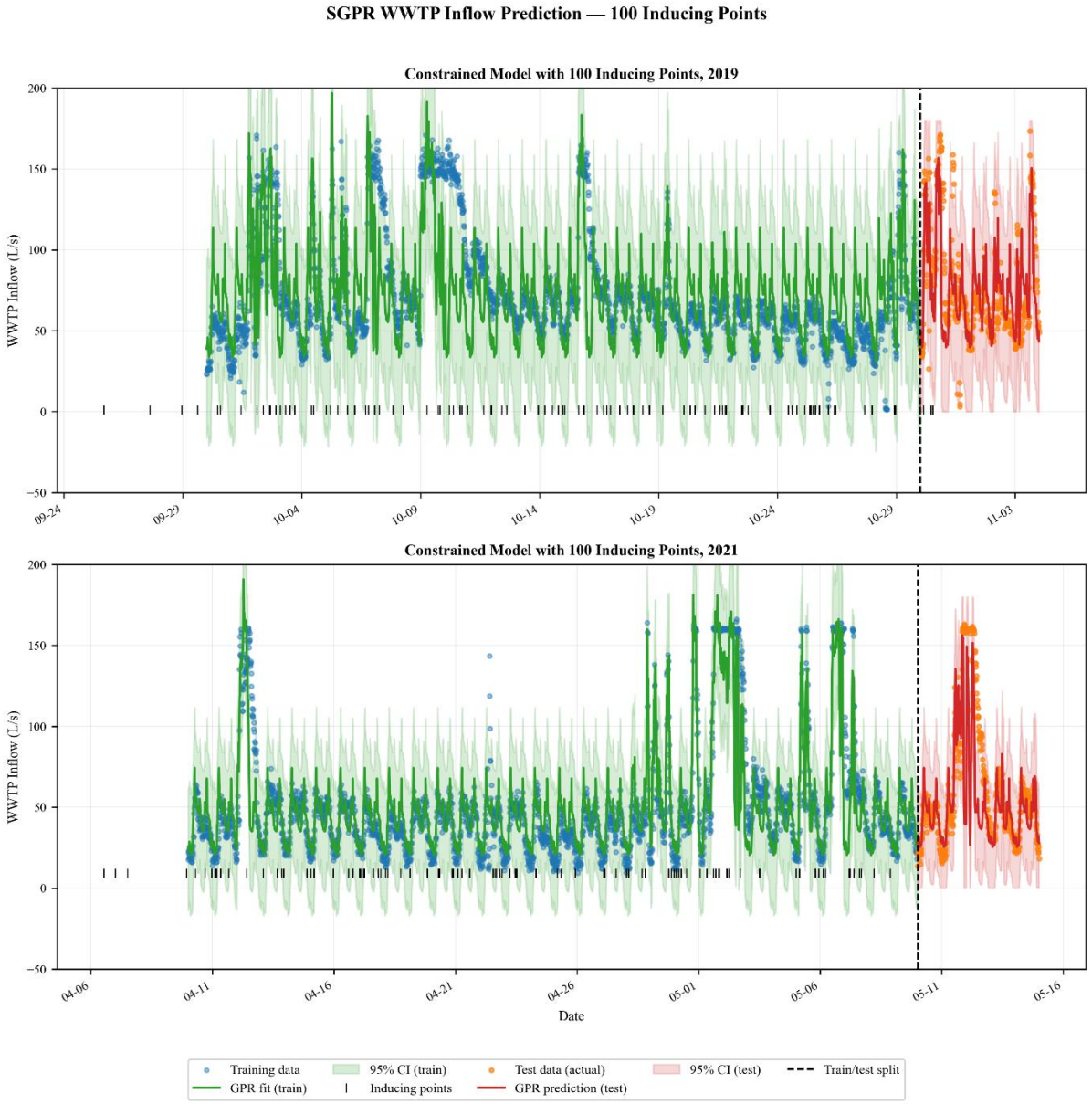

Figure 6. SGPR model with physical constraints for predicting WWTP inflow. The model implements 100 inducing points with the locations shown on the plots with black marks. Results over the 5-day forecasting period demonstrate performance and uncertainty bounds comparable to the full standard GPR model.

Table 2. Evaluation of different models for two time-series used in CSO prediction.

Model type	**Naïve**		**Designed Kernel**		**Designed Kernel + Mean function**		**Constrained**	
Year	**2019**	**2021**	**2019**	**2021**	**2019**	**2021**	**2019**	**2021**
RMSE (L/s)	127.08	102.62	24.81	127.67	24.85	104.25	40.37	95.79
MAE (L/s)	63.54	21.12	6.82	40.63	6.84	31.8	28.85	42.81
Coverage (%)	100	100	99.2	92.7	99.2	87.3	98.7	91.5
Entropy (nats)	12.535	9.77	6.936	6.675	6.937	6.013	6.203	5.423
Predicted CSO?	N/A	No	N/A	Yes	N/A	Yes	N/A	Yes
False Alarms?	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes

As shown in Table 2, the error and uncertainty were significantly reduced by designing the kernel for the 2019 time-series. However, it increased for the 2021 time-series. The reason behind this increase is that most prediction points have a value of zero because CSO rarely happens in a proper sewer system. Therefore, when the model fails to capture the CSO events, it lies at the mean value of the training dataset which is around zero. Consequently, the error values drop to near-zero values in non-CSO periods and only increase during short periods of overflow.

Moreover, by adding the constraint (minimum CSO value equal to zero) there is an increase in error values. This is due to the fact that in an unbounded distribution, the mean value of the distribution can easily be centred at zero, but when a strict zero-minimum physical constraint is applied, the predictive distribution is forced to remain positive, shifting the centre of mass slightly upwards and numerically increasing the error, despite being physically correct. However, this highlights a core advantage of the probabilistic framework over traditional deterministic evaluations. Rather than a modelling error, this upward shift reflects a physically correct reallocation of probability mass. Ultimately, this probabilistic approach ensures that unrealistic negative values are omitted, allowing the model to reliably capture actual CSO dynamics.

To enhance the ability of the models in capturing CSOs and avoiding false alarms, it is important to include more data into the training period. This increase in data points significantly increases the computational effort, therefore, a sparse model can help.

3.4. Stratified Sparse GPR for Predicting CSO

The novel method of distributing inducing points in selected periods significantly reduces the computational effort for running the models on even longer time-series. As the CSO occurs in a short period of time, more training information will enhance the quality of predictions.

As shown in Figure 7, the Naïve model failed to capture the CSO event, and the Constrained model made some false CSO predictions. However, the stratified sparse model could produce better results by gathering more information from the important periods with CSO events in the training data and prolonging the training period from 30 to 90 days. This extended training period not only enhanced the model's accuracy but also maintained an execution time less than one-quarter that of the full GPR model.

For both 30- and 90-day SGPR models, 250 inducing points were used which satisfies the minimum number of required inducing points for convergence. Since the composite kernel is a sum of one Matérn1/2, one Matérn5/2 and one SE kernel, a conservative lower bound is calculated based on the smoothest component (SE kernel) giving a theoretical minimum 38 needed inducing points.

From these 250 points, 50% were allocated to the CSO events and the rest 50% to the flat non-CSO periods. While these two models take narrower CI bounds around the predicted points and this causes a reduction in coverage rate, they are more precise in predicting the CSO values and making no false predictions. The evaluation metrics show that the RMSE was reduced from 95.8 L/s in the main Constrained model to 89.3 and 74.0 L/s in the sparse models (30-day and 90-day SGPR models, respectively). The error metrics remained almost the same for the classic SGPR model compared to the full GPR model but the coverage decreased and the entropy increased.

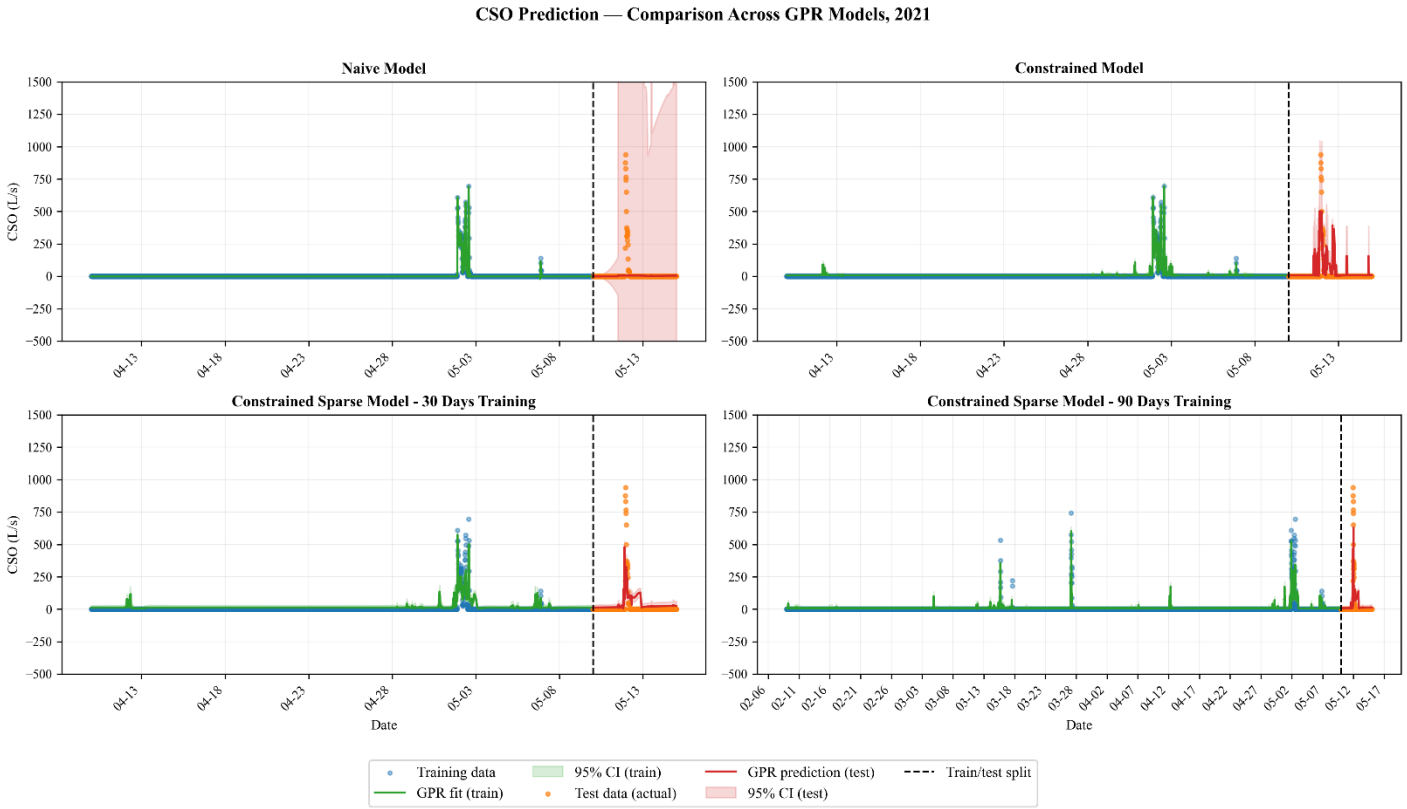

Figure 7. CSO prediction across different models with different levels of physical knowledge and training information. Top plots use the main GPR model for the prediction while the bottom plots use sparsification in the GPR model. All prediction periods are five days, and training periods are 30 days, except for the bottom right plot with 90 days of training period.

More importantly, the run time of sparse models was markedly lower than the run time of the GPR models without sparsification. This makes them feasible for predicting with longer training periods, finer data resolution and even with more input parameters.

3.5. Probabilistic Forecasting of CSOs via the Upstream Tank

Although direct prediction of CSOs using historical overflow time-series provides valuable insight for managing future events and warning the public, it still lacks a probabilistic view that predicts the likelihood of spills occurring at every timestep. This limitation can be effectively resolved by shifting the predictive focus to the upstream in-sewer storage facility. Because GPR produces a predictive posterior distribution, rather than deterministic outputs, the possibility of a CSO event can be calculated by finding the cumulative probability mass of the Gaussian distribution that exceeds the weir crest height.

Figure 8 shows the tank level prediction made by training the model for the same period in 2021 and predicting for the next five days which contain a CSO event. The model is constructed by designing a composite kernel and incorporating a mean function to make it smarter. The minimum/maximum constraints were not applied to allow the model to freely mimic the patterns, and the unrealistic values are automatically omitted from the prediction probability as it will show the likelihood of water levels exceeding the spill level.

The bottom plot in Figure 8 shows the overflow probability in the prediction period. In some timesteps, the CSO probability shows a considerable likelihood of spills that aligns with the actual CSO occurrences. This probabilistic crossing can be used as a warning window that is continuously updated with new data points which will be used as a great operational tool.

4. Discussion

4.1. Kernel Design and Addition of Physical Knowledge to the Model

As demonstrated in Tables 1 and 2, the transition from a naïve, unbounded GP to a physically constrained model is essential for accurately forecasting urban drainage systems. The results indicate that architectural kernel design had the most positive impact on predictive accuracy. However, this design is beyond simply choosing kernel types and summing or multiplying the covariance functions. It requires bounding the hyperparameters to reflect the physical characteristics of the system. For instance, by bounding the temporal lengthscales, the optimiser was prevented from falling into local minima or treating high-frequency sensor noise as physical flow. This is particularly critical in urban drainage datasets recorded at short intervals, where unconstrained lengthscales can approach the distance between data points, leading to near-singular covariance matrices and resulting in ill-conditioning.

Furthermore, the introduction of the SWMM-derived prior mean function shifted the mathematical role of the kernels in dry-weather periods. Rather than forcing the GP to learn daily DWF patterns from scratch, the mean function represented the baseline. Consequently, the temporal kernels were free to take the residuals between the baseline and the actual values. This was done by changing the lengthscale bounds to help the kernels handle the residuals more easily.

While adding hard physical constraints slightly improved the metrics in WWTP inflow prediction, it caused a decrease in the accuracy of the models used in CSO prediction. As mentioned earlier, the one-sided zero-minimum constraint on the distributions causes a shift to more positive values in the predictions, leading to an increase in errors, especially MAE. Nonetheless, it enables the model to capture the real CSO events which is the primary goal of a forecasting system in event-based time-series. A way to mitigate this type of error is to impose an upper bound for the flow, equal to the maximum capacity of the network.

A major advantage of enforcing these physical constraints is that the data-driven model becomes highly robust against severe, unrecorded system anomalies, such as sensor failures. For example, during the 2019 WWTP prediction period, the recorded sensor flow dropped to near-zero for over an hour. Because the GP is strictly governed by the physical diurnal patterns embedded in its architecture, it successfully resisted this non-physical deviation, maintaining a stable forecast within its 95% credible intervals. This highlights a critical strength of physics-informed GPs: they provide reliable, hydrologically sound responses that do not overfit to erroneous measurements. Consequently, the model maintains precise predictive bounds without artificially increasing its variance to account for non-physical system errors.

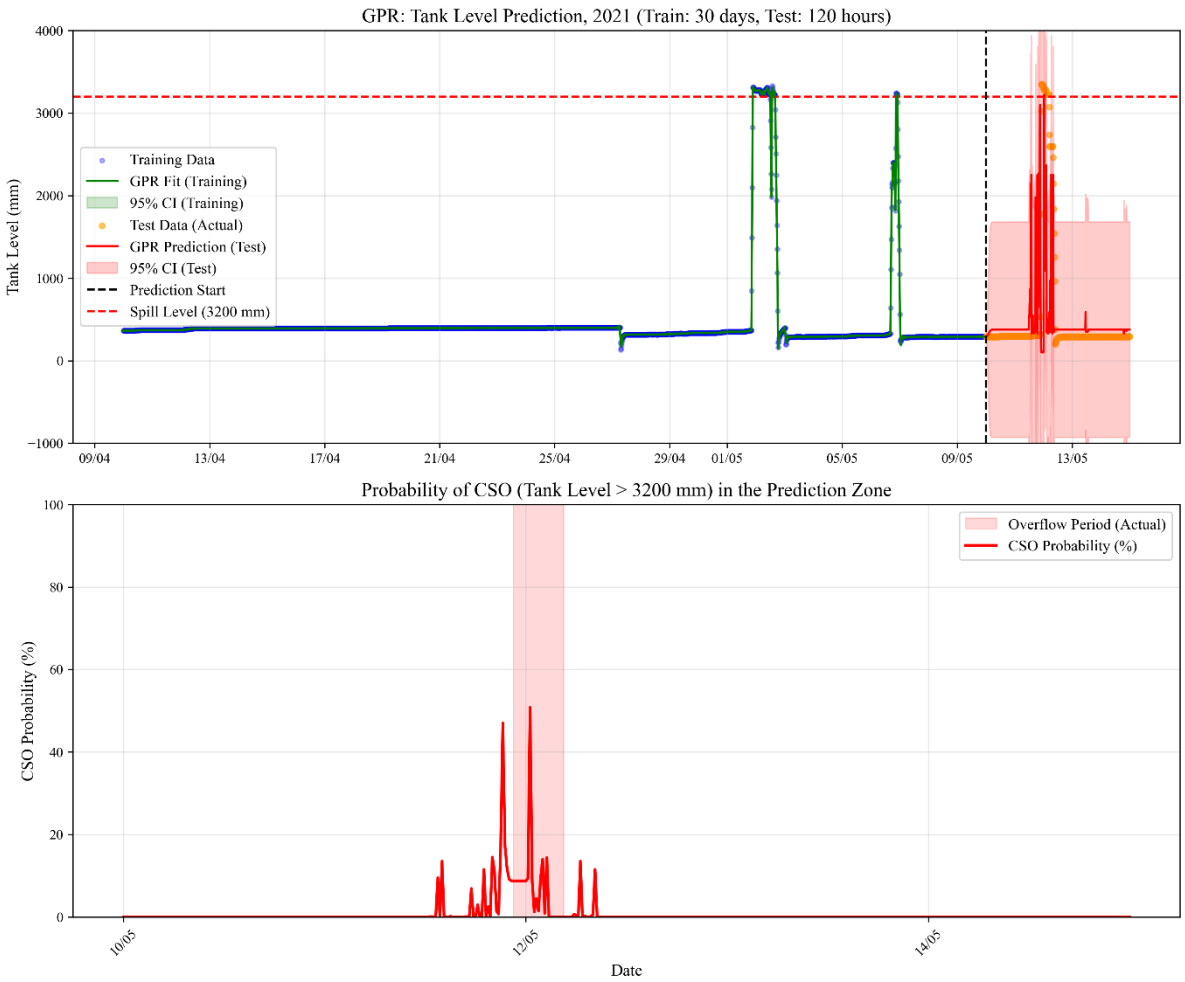

Figure 8. GPR model for predicting tank flow based on historic time-series. The top plot shows 30 days of training and 5 days of predicting. The bottom plot shows the probability of CSO occurrence in the prediction zone and the actual period of CSO.

4.2. Forecasting by CSO Data vs Tank Level Data

As the CSO prediction results show, while overflows can be directly predicted by a GPR model, the results contain errors in both overall accuracy and the generation of false alarms. On the other hand, the results of the tank level prediction showed that false alarms can completely disappear by using this technique, especially in systems where there is good sensor availability. However, relying only on tank level forecasts lacks the ability to show the exact magnitude of the overflow volume.

Therefore, in an urban drainage system like this case study with multiple sensors available, a combination of CSO prediction and tank level prediction can make a highly effective forecasting model. This combination enables water utility operators to take proactive actions and warn the public. Also, if one model is to be preferred over the other for proactive management, predicting the likelihood of overflow using the GPR model over the tank level data is a better choice. Because GPR provides a full probability distribution, it gives a clear, continuous insight into the possibility of a CSO happening before the threshold is breached, rather than just an estimated flow value with a credible interval.

4.3. Sparsification

As mentioned earlier, increasing the length of the training period can significantly improve the quality of the GPR model output in CSO prediction by taking more peak flows. It can also enhance the quality of WWTP inflow predictions by including more data points and better understanding the long-term trends in the data, caused by groundwater infiltration and water usage changes. However, this is only efficient when using an SGPR model that reduces the computational cost.

In selecting the number of inducing points, it should be noted that although more inducing points lead to a higher accuracy, the optimisation process for finding the location of these points is a computational burden for the model. Therefore, a trade-off between the accuracy and speed shows the optimal number of inducing points. For example, the overall runtime of the Constrained model with 2,880 training points was almost the same as the runtime of the SGPR model with 500 inducing points. Furthermore, the sparse model showed lower accuracy and more entropy in the predictions. A way to increase the speed of the model while having many inducing points is to exclude them from the optimisation process and spread them along the training dataset in constant intervals or using the k-means method. This way, the speed of the model increases, but the accuracy heavily drops.

As shown in Figure 6, some of the inducing points lie outside the training period. This means that the optimiser has placed them outside the training period to maximise the approximation to the full GP posterior. It also ensures the model does not fall back on the mean value on the edges of the training period.

Finally, the novel stratified SGPR model showed promising outputs in predicting CSOs with 30 and 90 days of training data with a significant decrease in computational effort. It also fixed the problem of false alarms in the full GP model and decreased the RMSE value, which is mainly caused by the CSO events. However, the overall MAE increased because of an increase in the baseline value. But as the main duty of such models is to spot overflows, the stratified SGPR model is believed to outperform the other models.

4.4. Limitations and Future Direction

It should be noted that in the predictions presented in this study, the rainfall data for the forecasted days is assumed to be exact. In an online operational deployment such as within industry-standard flood early-warning platforms like Delft-FEWS, these precipitation inputs must rely on numerical weather predictions (NWPs) and short-term radar casting, which inherently introduce an overriding layer of meteorological uncertainty into the system. However, the probabilistic nature of the GPR model is well-equipped to handle this limitation. As the uncertainty of the weather forecast increases further into the future, the GP naturally expands its credible intervals to reflect the lack of exact deterministic knowledge, ensuring that operators are provided with a realistic, bounded range of possible flow outcomes.

A valuable direction for future research would be integrating ensemble weather forecasts directly into the GP's input dimensions to explicitly quantify this meteorological uncertainty. Furthermore, future work could allow weather forecasts to enter the GP model as probabilistic distributions rather than deterministic data points. While this would rigorously account for

meteorological variance, propagating input uncertainty through nonlinear covariance functions introduces highly complex mathematical challenges for matrix inversion that remain outside the scope of the current study.

Furthermore, while the novel stratified SGPR method successfully resolved the computational bottlenecks associated with event-based time-series, the allocation strategy leaves room for further refinement. In this work, the separation of dry and wet periods was achieved through a simple, density-based clustering heuristic, and a dedicated percentage of the overall inducing points was explicitly allocated to each region. A promising future direction would be to integrate this allocation ratio directly into the optimisation process. By allowing the optimiser to dynamically adjust the percentage of inducing points assigned to each region to maximise the ELBO, the model could autonomously find the optimal balance between capturing the dry-weather baseline and the chaotic wet-weather surges. Additionally, the initial classification of dry and wet periods could be significantly enhanced by deploying advanced supervised machine learning algorithms, such as XGBoost, to more intelligently partition the hydrological regimes prior to sparsification.

5. Conclusion

This work introduced a physics-informed Gaussian process regression approach to deliver reliable, probabilistic flow forecasts across complex urban drainage networks. The framework systematically evolved a naïve, purely data-driven baseline into a rigorous, physics-informed model by integrating domain knowledge through composite kernel engineering, SWMM-derived prior mean functions, and strict physical output constraints. The methodology was validated using the high-resolution UWO dataset to predict WWTP inflow, downstream tank levels, and CSO discharges.

Based on the findings of this research, it is evident that the naïve GP model is not strong enough to capture flow patterns in forecasting both WWTP inflow and CSO. It can produce unrealistic results that stand on the mean value of the training data. These outputs also contain wide CIs that show high uncertainty in predictions.

Three levels of physical information were defined in this work. Among them, designing a domain-aware kernel had the most positive effect on improving the accuracy of models and reducing uncertainty in predictions.

Sparsification significantly reduces the computational effort of GP models which can be seen as the bottleneck of these models. The proposed sparse architecture reduced overall execution times to less than 10% of the exact full GP model's runtime. Through the sparsification process, the accuracy of the model is slightly reduced as a minor RMSE increase of 1.3% and 3.9% for the 2019 and 2021 periods, respectively; however, this minimal accuracy loss is acceptable when compared with the remarkable reduction in execution times.

CSOs can be predicted using GP models in two ways. Modelling the upstream storage tank provides a superior, probabilistic early-warning trigger that mitigates false alarms, whereas direct historical CSO prediction is required to quantify the exact volume of an anticipated spill. The use of each model depends on the availability of data and the goal of forecasting.

The novel stratified sparsification method helps increasing the accuracy of the GPR models in predicting event-based time-series such as CSO data and simultaneously reducing the

computational effort of the model. This approach enables GPR models to handle longer time-series and more data points, far more than conventional sparse GP methods.

A key advantage of GP-based models is their ability to produce reliable predictions from sparse and infrequent observations, reducing the need for dense sensor networks and high-frequency data collection. Unlike conventional models where finer temporal resolution typically improves performance, GP models extract sufficient physical information from averaged, wider time-interval data, making them equally effective at coarser resolutions. This property directly translates to reduced instrumentation costs, lower data transmission and storage requirements, and simplified monitoring infrastructure, while maintaining prediction quality. Consequently, GP-based frameworks offer a cost-effective solution for predicting urban drainage characteristics such as flow, particularly in resource-constrained catchments where sensor placement, maintenance, and data collection costs are significant.

In summary, this work offers a robust framework for predicting complex, uncertain parameters within a water system using physics-informed Gaussian processes. This methodology can be readily expanded to enhance probabilistic forecasting in other critical environmental domains, including water distribution networks, nature-based solutions, and water quality modelling in natural water bodies.

Acknowledgment

Hussein Rappel would like to thank the Leverhulme Trust for their support (RPG-2024-330).

Author contributions: CRediT

Conceptualisation, M.R., H.R., P.M; Data curation, M.R.; Formal analysis, M.R.; Investigation, M.R., H.R.; Methodology, M.R., H.R.; Project administration, H.R.; Resources, M.R., H.R., P.M.; Software, M.R.; Supervision, H.R., P.M.; Validation, M.R., H.R.; Visualisation, M.R.; writing—original draft, M.R.; writing—review and editing, M.R., H.R., P.M.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

All codes and generated results are available through the GitHub repository at https://github.com/MohsenRz/PI-GPR_UWOdataset.git

References

Aliashrafi, A., Zhang, Y., Groenewegen, H., & Peleato, N. M. (2021). A review of data-driven modelling in drinking water treatment. *Rev. Environ. Sci. Biotechnol.*, *20*(4), 985-1009. https://doi.org/10.1007/s11157-021-09592-y

Balla, K. M., Eringis, D., Ahdab, M. A., Bendtsen, J. D., Kallesøe, C., & Ocampo-Martinez, C. (2022). Learning- Based Predictive Control with Gaussian Processes: An Application to Urban Drainage Networks. In *2022 American Control Conference, ACC* (pp. 4627-4633). IEEE (Institute of Electrical and Electronics Engineers). https://doi.org/10.23919/ACC53348.2022.9867801

Blumensaat, F., Bloem, S., Ebi, C., Disch, A., Förster, C., Maurer, M., Rodriguez, M. & Rieckermann, J. (2025). The UWO dataset–long-term observations from a full-scale field laboratory to better understand urban hydrology at small spatio-temporal scales. *Earth Syst. Sci. Data Discuss.*, *2025*, 1-31.

Botturi, A., Ozbayram, E.G., Tondera, K., Gilbert, N.I., Rouault, P., Caradot, N., Gutierrez, O., Daneshgar, S., Frison, N., Akyol, Ç. and Foglia, A. (2021). Combined sewer overflows: A critical review on best practice and innovative solutions to mitigate impacts on environment and human health. *Crit. Rev. Environ. Sci. Technol.*, *51*(15), 1585-1618. https://doi.org/10.1080/10643389.2020.1757957

Burt, D., Rasmussen, C. E., & Van Der Wilk, M. (2019, May). Rates of convergence for sparse variational Gaussian process regression. In *International conference on machine learning* (pp. 862-871). PMLR.

Burt, D. R., Rasmussen, C. E., & Van Der Wilk, M. (2020). Convergence of sparse variational inference in Gaussian processes regression. *J. Mach. Learn. Res.*, *21*(131), 1-63.

Butler, D., Digman, C., Makropoulos, C., & Davies, J.W. (2024). Urban Drainage (5th ed.). CRC Press. https://doi.org/10.1201/9781003408635

Cedillo, S., Núñez, A. G., Sánchez-Cordero, E., Timbe, L., Samaniego, E., & Alvarado, A. (2022). Physics-Informed Neural Network water surface predictability for 1D steady-state open channel cases with different flow types and complex bed profile shapes. *Adv. Model. Simul. Eng. Sci.*, *9*(1), 10. https://doi.org/10.1186/s40323-022-00226-8

Chen, Y., Hosseini, B., Owhadi, H., & Stuart, A. M. (2021). Solving and learning nonlinear PDEs with Gaussian processes. *J. Comput. Phys.*, *447*, 110668. https://doi.org/10.1016/j.jcp.2021.110668

Cross, E. J., Gibson, S. J., Jones, M. R., Pitchforth, D. J., Zhang, S., & Rogers, T. J. (2021). Physics-informed machine learning for structural health monitoring. In *Structural health monitoring based on data science techniques* (pp. 347-367). Cham: Springer International Publishing.

Cross, E. J., Rogers, T. J., Pitchforth, D. J., Gibson, S. J., Zhang, S., & Jones, M. R. (2024). A spectrum of physics-informed Gaussian processes for regression in engineering. *Data-Centric Eng.*, *5*, e8. https://doi.org/10.1017/dce.2024.2

Da Veiga, S., & Marrel, A. (2012). Gaussian process modeling with inequality constraints. In *Annales de la Faculté des sciences de Toulouse: Mathématiques* (Vol. 21, No. 3, pp. 529-555). https://doi.org/10.5802/afst.1344

Daw, A., Karpatne, A., Watkins, W. D., Read, J. S., & Kumar, V. (2022). Physics-guided neural networks (PGNN): An application in lake temperature modeling. In *Knowledge guided machine learning* (pp. 353-372). Chapman and Hall/CRC.

Deshpande, S., Rappel, H., Hobbs, M., Bordas, S. P., & Lengiewicz, J. (2025). Gaussian process regression+ deep neural network autoencoder for probabilistic surrogate modeling in nonlinear mechanics of solids. *Comput. Methods Appl. Mech. Eng.*, *437*, 117790. https://doi.org/10.1016/j.cma.2025.117790

Di Bella, A., Raissi, M., Santoro, D., & Roccaro, P. (2026). Physics-informed neural networks in water and wastewater systems: a critical review. *Water Res.*, 125449. https://doi.org/10.1016/j.watres.2026.125449

Ding, C., Rappel, H., & Dodwell, T. (2023). Full-field order-reduced Gaussian Process emulators for nonlinear probabilistic mechanics. *Comput. Methods Appl. Mech. Eng*, *405*, 115855. https://doi.org/10.1016/j.cma.2022.115855

Donnelly, J., Daneshkhah, A., & Abolfathi, S. (2024a). Forecasting global climate drivers using Gaussian processes and convolutional autoencoders. *Eng. Appl. Artif. Intell.*, *128*, 107536. https://doi.org/10.1016/j.engappai.2023.107536

Donnelly, J., Daneshkhah, A., & Abolfathi, S. (2024b). Physics-informed neural networks as surrogate models of hydrodynamic simulators. *Sci. Total Environ.*, *912*, 168814. https://doi.org/10.1016/j.scitotenv.2023.168814

Garzón, A., Kapelan, Z., Langeveld, J., & Taormina, R. (2024). Transferable and data efficient metamodeling of storm water system nodal depths using auto-regressive graph neural networks. *Water Res.*, *266*, 122396. https://doi.org/10.1016/j.watres.2024.122396

Ge, J., Li, J., Qiu, R., Shi, T., Zhang, C., Huang, Z., & Yuan, Z. (2024). A data-driven method for estimating sewer inflow and infiltration based on temperature and conductivity monitoring. *Water Res.*, *261*, 122002. https://doi.org/10.1016/j.watres.2024.122002

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.

He, L., Zhang, Q., Shi, L., Wang, Y., Wang, L., Hu, X., Zha, Y. & Huang, K. (2023). Physics-constrained Gaussian process regression for soil moisture dynamics. *J. Hydrol.*, *616*, 128779. https://doi.org/10.1016/j.jhydrol.2022.128779

Hensman, J., Fusi, N., & Lawrence, N. D. (2013). Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*. https://doi.org/10.48550/arXiv.1309.6835

Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geosci. Model Dev. Discuss.*, *2022*, 1-10. https://doi.org/10.5194/gmd-15-5481-2022

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *Int. J. Forecast. 22*(4), 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Kohanpur, A. H., Saksena, S., Dey, S., Johnson, J. M., Riasi, M. S., Yeghiazarian, L., & Tartakovsky, A. M. (2023). Urban flood modeling: Uncertainty quantification and physics-informed Gaussian processes regression forecasting. *Water Resour. Res.*, *59*(3), e2022WR033939. https://doi.org/10.1029/2022WR033939

Li, S., Tian, W., Yan, H., Zeng, W., Tao, T., & Xin, K. (2024). Modeling transient mixed flows in sewer systems with data fusion via physics-informed machine learning. *Water Res. X*, *25*, 100266. https://doi.org/10.1016/j.wroa.2024.100266

Luo, X., Yuan, S., Tang, H., Xu, D., Ran, Q., Cen, Y., & Liang, D. (2024). Enhanced physics-informed neural networks for efficient modelling of hydrodynamics in river networks. *Hydrol. Process.*, *38*(4), e15143. https://doi.org/10.1002/hyp.15143

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

Mahmoodian, M., Torres-Matallana, J. A., Leopold, U., Schutz, G., & Clemens, F. H. (2018). A data-driven surrogate modelling approach for acceleration of short-term simulations of a dynamic urban drainage simulator. *Water*, *10*(12), 1849. https://doi.org/10.3390/w10121849

Malde, S. Gaussian Process Emulators in Coastal Wave Modelling. Ph.D. Thesis, University of Sheffield, Sheffield, UK, 2018.

Matthews, A.G.D.G., Van Der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z. & Hensman, J. (2017). GPflow: A Gaussian process library using TensorFlow. *J. Mach. Learn. Res.*, *18*(40), 1-6.

Mounce, S. R., Shepherd, W., Sailor, G., Shucksmith, J., & Saul, A. J. (2014). Predicting combined sewer overflows chamber depth using artificial neural networks with rainfall radar data. *Water Sci. Technol.*, *69*(6), 1326-1333. https://doi.org/10.2166/wst.2014.024

Padilla-Segarra, A., Noble, P., Roustant, O., & Savin, É. (2025). Physics-informed, boundary-constrained Gaussian process regression for the reconstruction of fluid flow fields. *arXiv preprint arXiv:2507.17582*. https://doi.org/10.48550/arXiv.2507.17582

Palmitessa, R., Grum, M., Engsig-Karup, A. P., & Löwe, R. (2022). Accelerating hydrodynamic simulations of urban drainage systems with physics-guided machine learning. *Water Res.*, *223*, 118972. https://doi.org/10.1016/j.watres.2022.118972

Pastrana Cortés, J. D. (2024). Stochastic modeling of multiple streamflow time series in Colombia based on Gaussian processes. PhD Thesis.

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

Quinonero-Candela, J., & Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.*, *6*(Dec), 1939-1959.

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.*, *378*, 686-707. https://doi.org/10.1016/j.jcp.2018.10.045

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2017). Machine learning of linear differential equations using Gaussian processes. *J. Comput. Phys.*, *348*, 683-693. https://doi.org/10.1016/j.jcp.2017.07.050

Rappel, H., Girolami, M., & Beex, L. A. (2022). Intercorrelated random fields with bounds and the Bayesian identification of their parameters: Application to linear elastic struts and fibers. *Int. J. Numer. Methods Eng.*, *123*(15), 3418-3463. https://doi.org/10.1002/nme.6974

Rasmussen, C.E.; Williams, C.K.I. Gaussian Processes for Machine Learning; MIT Press: Cambridge, MA, USA, 2006.

Reid, C. (2022). Environment Act 2021.

Rezaee, M., Melville-Shreeve, P., & Rappel, H. (2025). Probabilistic forecasting and anomaly detection in sewer systems using Gaussian processes. *Water*, *17*(16), 2357. https://doi.org/10.3390/w17162357

Rezaee, M., & Tabesh, M. (2022). Effects of inflow, infiltration, and exfiltration on water footprint increase of a sewer system: A case study of Tehran. *Sustain. Cities Soc.*, *79*, 103707. https://doi.org/10.1016/j.scs.2022.103707

Rios, G., & Tobar, F. (2019). Compositionally-warped Gaussian processes. *Neural Netw*, *118*, 235-246. https://doi.org/10.1016/j.neunet.2019.06.012

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., & Aigrain, S. (2013). Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*(1984). https://doi.org/10.1098/rsta.2011.0550

Rosin, T., Romano, M., Woodward, K., Keedwell, E., & Kapelan, Z. (2017). Prediction of CSO chamber water levels using rainfall forecasts. *Computing and Control for the Water Industry*.

Rosin, T. R., Romano, M., Keedwell, E., & Kapelan, Z. (2021). A committee evolutionary neural network for the prediction of combined sewer overflows. *Water Resour. Manage.*, *35*(4), 1273-1289. https://doi.org/10.1007/s11269-021-02780-z

Snelson, E., Ghahramani, Z., & Rasmussen, C. (2003). Warped gaussian processes. *Advances in neural information processing systems*, *16*.

Suchowska-Kisielewicz, M., & Nowogoński, I. (2021). Influence of storms on the emission of pollutants from sewage into waters. *Sci. Rep.*, *11*(1), 18788. https://doi.org/10.1038/s41598-021-97536-5

Swiler, L. P., Gulian, M., Frankel, A. L., Safta, C., & Jakeman, J. D. (2020). A survey of constrained Gaussian process regression: Approaches and implementation challenges. *J. Mach. Learn. Model. Comput.*, *1*(2).

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

Titsias, M. (2009, April). Variational learning of inducing variables in sparse Gaussian processes. In *Artificial intelligence and statistics* (pp. 567-574). PMLR.

Tiwari, A., Valkama, K., Krolicka, A., Miettinen, I. T., & Pitkänen, T. (2025). Coverage and treatment practices in nordic wastewater treatment plants under the recast EU urban wastewater treatment directive. *Discover Water*, *5*(1), 105. https://doi.org/10.1007/s43832-025-00297-9

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 Res.*, *126*, 88-100. https://doi.org/10.1016/j.watres.2017.08.065

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

Wang, J. (2023). An intuitive tutorial to Gaussian process regression. *Comput. Sci. Eng.*, *25*(4), 4-11. https://doi.org/10.1109/MCSE.2023.3342149

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2020). Integrating physics-based modeling with machine learning: A survey. *arXiv preprint arXiv:2003.04919*, *1*(1), 1-34.

Wilson, A.G. Covariance Kernels for Fast Automatic Pattern Discovery and Extrapolation with Gaussian Processes. Ph.D. Thesis, University of Cambridge, Cambridge, UK, 2014.

Ye, D., Yan, W., Brune, C., & Guo, M. (2025). PDE-constrained Gaussian process surrogate modeling with uncertain data locations. *Adv. Model. Simul. Eng. Sci.*, *12*(1), 33. https://doi.org/10.1186/s40323-025-00308-3

Zeydalinejad, N., Javadi, A. A., & Webber, J. L. (2024). Global perspectives on groundwater infiltration to sewer networks: A threat to urban sustainability. *Water Res.*, *262*, 122098. https://doi.org/10.1016/j.watres.2024.122098

Zhang, J., Liu, C., & Gao, R. X. (2022). Physics-guided Gaussian process for HVAC system performance prognosis. *Mech. Syst. Signal Process*, *179*, 109336. https://doi.org/10.1016/j.ymssp.2022.109336

Zhang, J., Xiao, C., Yang, W., Liang, X., Zhang, L., Wang, X., & Dai, R. (2024). Improving prediction of groundwater quality in situations of limited monitoring data based on virtual sample generation and Gaussian process regression. *Water Res.*, *267*, 122498. https://doi.org/10.1016/j.watres.2024.122498

Zhou, S., Yang, Y., Ng, S. T., Xu, J. F., & Li, D. (2020). Integrating data-driven and physics-based approaches to characterize failures of interdependent infrastructures. *International Int. J. Crit. Infrastruct. Prot.*, *31*, 100391. https://doi.org/10.1016/j.ijcip.2020.100391