Sensor Drift-Robust GNN Surrogates for Urban Wastewater State Estimation

Imane El Ghabi*
Informatics Institute, University of Amsterdam
The Netherlands
i.elghabi@uva.nl

Revin Naufal Alief*
Bernoulli Institute, University of Groningen
The Netherlands
r.n.alief@rug.nl

Guy Henckens
Aveco de Bondt
The Netherlands
ghenckens@avecodebondt.nl

Alexander Lazovik
Bernoulli Institute, University of Groningen
The Netherlands
a.lazovik@rug.nl

Dilek Düştegör
Bernoulli Institute, University of Groningen
The Netherlands
d.dustegor@rug.nl

Victoria Degeler
Informatics Institute, University of Amsterdam
The Netherlands
v.o.degeler@uva.nl

ABSTRACT

Urban wastewater systems are critical infrastructures whose monitoring is complicated by sparse sensing, stochastic inflows, and complex hydraulic behavior. Physics-based hydraulic models provide accurate simulations but are computationally demanding and unsuitable for real-time operation. We adapt a graph neural network surrogate model GATRes to wastewater networks and train it on synthetic data generated from hydraulic simulations checked against known critical locations. Our main case study is the Heusden wastewater network in the Netherlands, developed in collaboration with the local municipality. The model achieves prediction errors of up to 50 cm for maximum water levels in non-flooding manholes across multiple rainfall scenarios, which aligns with the precision expected by water domain experts, though much tighter tolerances (0–10 cm) are expected at weir locations. To assess generalization, we further evaluate the approach on the open Shunqing stormwater dataset from China. However, sparse-sensing training alone does not make the model robust to biased sensor measurements. We evaluate this failure mode using synthetic additive offsets and random-walk perturbations applied to the observed sensor values. Across three random seeds, a baseline model's error grows erratically under increasing sensor drift, from +16% to +94% relative MAE as drift goes from none to total. Injecting synthetic drift perturbations into a subset of sensors during training instead yields small, consistent error growth across all three seeds (+6–8%), closing most of the gap. Results demonstrate strong overall accuracy under sparse sensing and across unseen rainfall scenarios, while highlighting challenges near hydraulic structures such as pumps, weirs, and outfalls, where errors are higher but remain within acceptable ranges. These findings show that GNN surrogates can provide scalable wastewater state estimation and, with corruption-aware training, improve robustness under the evaluated synthetic sensor-drift conditions.

*Equally contributed.

© 2026 Copyright held by the author(s). This work is based on an earlier work: SAC'26 Proceedings of the 2026 ACM Symposium on Applied Computing, 2026 ACM 979-8-4007-2294-3. http://dx.doi.org/10.1145/3748522.3779823

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; Modeling and simulation; • **Applied computing** → Environmental sciences; • **Information systems** → Decision support systems;

KEYWORDS

Sensor Drift, Robustness, Graph Neural Networks, Surrogate Modeling, Wastewater Networks, State Estimation, Smart Cities, Critical Infrastructures

1 INTRODUCTION

Urban wastewater systems are essential components of smart cities and critical infrastructures, where reliable monitoring is necessary to reduce risks of flooding, combined sewer overflows, and infrastructure failure. In smart city contexts, these systems must operate efficiently under varying demand patterns while maintaining service reliability and environmental compliance. Yet, despite their importance, monitoring and managing wastewater networks remain challenging because rainfall-derived infiltration and inflow cause substantial wet-weather flow variation and contribute to sewer overflows, while underground sensor deployment introduces practical communication and coverage constraints [20]. Achieving reliable monitoring remains difficult due to three main challenges: (i) sensor coverage is sparse because of cost and maintenance constraints, (ii) inflows are highly stochastic and vary with weather, and (iii) flow dynamics are shaped by complex hydraulic interactions involving pumping stations, weirs, and outfalls. These challenges make it difficult for operators to obtain a complete and timely picture of system states such as flood depths and flows, limiting their ability to make informed decisions for infrastructure management and emergency response.

Physics-based hydraulic models such as InfoWorks ICM (Integrated Catchment Modeling) [3] or SWMM [21] are widely used by municipalities to simulate and manage urban drainage. While they provide accurate predictions when carefully calibrated, they are computationally demanding and unsuitable for real-time decision support [9]. This creates a gap between the need for scalable monitoring of wastewater networks and the practical limitations of physics-based tools.

Digital Twins (DTs) have been proposed as a promising paradigm to address such challenges by coupling physical systems with data-driven models that enable continuous monitoring, prediction, and control [26]. However, developing DTs for urban wastewater systems requires reliable and efficient surrogate models that can estimate complete network states from sparse sensor observations in real time.

Data-driven surrogate models have recently emerged as an efficient alternative. By learning from synthetic or historical data, surrogates can provide rapid estimates of network-wide states from limited sensor inputs [14]. Graph neural networks (GNNs) are particularly suited for this task because wastewater networks are inherently graph-structured, with nodes representing locations or structures, such as manholes, storage, and outfalls, connected by hydraulic links such as pipes and channels. Prior studies have demonstrated the effectiveness of GNNs in water distribution systems, but wastewater networks present unique challenges: they are gravity-driven, subject to stochastic inflows, and prone to overflow events. These properties make wastewater modeling a distinct problem that requires dedicated solutions.

In this paper, we adapt the GATRes model to urban wastewater networks and evaluate it on data generated from hydraulic simulations checked against known critical locations. This represents the first application of GATRes to wastewater systems, extending its proven capabilities from water distribution networks to the distinct challenges of gravity-driven drainage systems. Our main case study is the Heusden network in the Netherlands, developed in collaboration with the local municipality. The model is further tested on the open Shunqing stormwater dataset from Sichuan, China [28], enabling validation across different infrastructures and contexts.

The masking-based training strategy underlying this surrogate assumes that the available sensors provide reliable measurements. In real deployments, this assumption may be violated in different ways. Measurements may be missing, gradually biased by sensor drift, or affected when a sensor measures an obstruction rather than the true water surface. Moreover, physical changes, such as sediment accumulation, may produce measurement patterns similar to sensor drift. Distinguishing these causes from sensor data alone is difficult. This study does not attempt to diagnose the cause of an unusual measurement; instead, it evaluates how synthetic bias in the available sensor inputs affects state estimation. Because message passing can propagate the influence of a biased input to other nodes, a model trained only with clean sensor inputs may be vulnerable to this noise data. Beyond coping with such corruption, identifying that a sensor is reporting an anomalous value is itself operationally valuable, since unusual readings often have a physical cause, such as a developing blockage, that operators need to investigate promptly; we return to this point in Section 8.

This article extends our SAC'26 conference paper [8]. Contributions 1–3 summarize the state-estimation framework and findings established in the conference paper, while Contributions 4–6 are new to this extended article.

(1) **Novel Application:** We present the first application of GATRes to wastewater networks, demonstrating its adaptation from water distribution systems to gravity-driven drainage systems with stochastic inflows and overflow dynamics.
(2) **Expert-Validated Accuracy:** We show that the GNN surrogate achieves prediction errors of up to 50 cm for maximum water levels in non-flooding manholes across multiple rainfall scenarios in Heusden, a range validated by water domain experts as operationally sufficient for decision-making in smart city infrastructures, though much tighter tolerances (0–10 cm) are expected at weir locations.
(3) **Robust Generalization:** We demonstrate that the approach generalizes to unseen rainfall events and to an independent stormwater network (Shunqing), highlighting robustness across scenarios and infrastructures for practical sparse sensing applications.
(4) **Sensor-Drift Corruption Models:** Physically-motivated corruption models for both networks, applied on top of the sparse-sensing masking protocol and correctly ordered with respect to feature normalization: an additive random-walk model for Heusden's continuous water-level sensors and a multiplicative model for Shunqing's event-driven flow variable (Section 4.3.1).
(5) **Masking Alone Does Not Confer Drift Robustness:** A masking-only GNN surrogate's relative error growth from no drift to total corruption ranges from +16% to +94% across three random seeds on Shunqing, and from +124% to +249% across three random sensor placements on Heusden. This range depends on incidental factors such as sensor placement rather than on drift severity alone (Sections 5.4, 5.5).
(6) **Denoising Training Closes the Gap:** Exposing the same architecture to drift during training, resampled every mini-batch, yields a consistently small relative error growth on both networks (+6–8% on Shunqing, +32–48% on Heusden) and is more accurate than the baseline at every drift level above zero (Sections 5.4, 5.5).

This work provides a validated surrogate modeling framework for state estimation in wastewater systems. By combining simulation data with graph-based learning, it contributes to the development of smart city infrastructures and digital twins for urban water management.

2 RELATED WORK

Research on digital technologies for smart cities has explored diverse domains such as transportation, energy grids, and water distribution networks, often highlighting the role of data-driven models in enabling resilient and adaptive monitoring. Recent approaches to improving wastewater systems as critical infrastructures can be grouped into five main directions: (1) simplifying physics-based simulations, (2) leveraging machine learning for predictive analytics, (3) applying graph-based deep learning to capture network-wide dynamics, (4) integration of DTs for decision making, and (5) improving robustness to sensor drift and noise. Together, these approaches aim to advance smart urban wastewater infrastructures that are both computationally feasible and operationally robust.

(1) Physics-Based Surrogates and Hybrid Models: Reducing Calibration Burden. To tackle the issue of slow and complex urban wastewater simulations, Meirlaen et al. [15] propose mechanistic

surrogate models – simplified yet physically consistent representations calibrated with virtual data from detailed simulations. Simplified physics-based models are already used in regions like Flanders, Belgium. Vaes et al. [25] proposed long-term simulations using nonlinear reservoir models, calibrated against detailed hydrodynamic models, to evaluate combined sewer overflow (CSO) impacts on river systems. They highlight that using simpler, well-calibrated models can be more effective than highly complex ones, especially when the input data (like rainfall) is uncertain.

(2) Machine Learning for Rapid Process Estimation: Bridging Data Gaps. Zhao et al. [27] propose a machine learning approach to estimate readily biodegradable chemical oxygen demand (RB-COD) and slowly biodegradable COD (SB-COD) in municipal wastewater using oxidation–reduction potential (ORP) data from anoxic batch experiments. Among six tested machine learning models, the SVR-based AdaBoost model achieved the highest accuracy for predicting both RB-COD and SB-COD. Likewise, Inbar et al. [12] developed a generalizable machine learning model to predict biochemical oxygen demand (BOD) from commonly measured water quality parameters at 30 wastewater treatment plants in Israel. Both the studies highlight the potential of machine learning for rapid and dynamic control of biological wastewater treatment processes. This approach enables automated, real-time estimation without explicit physical modeling. Nevertheless, it lacks physical interpretability and may underperform under system changes unless frequently retrained.

(3) Graph-Based Deep Learning: Capturing Network Topology and Nonlinear Dynamics. Wastewater systems, like other urban infrastructures, have a complex spatial structure that traditional models often overlook. Recent efforts incorporate Graph Neural Networks (GNNs) to model spatial and temporal dependencies across infrastructure components. Chen et al. [5] introduce a Dynamic Attention Graph Convolutional Network (DA-GCN) designed to enhance early detection of sludge bulking in activated sludge treatment processes. Similarly, Baskar et al. [4] propose a node-level capsule graph neural network (PWWT-PEQ-NLCGNN) with the hermit crab optimization algorithm to predict effluent chemical oxygen demand (COD) in papermaking wastewater treatment. While this approach captures hidden spatiotemporal patterns and enables intelligent fault detection, it focuses on isolated subsystems rather than full-network modeling. While these innovations improve single subsystems, adopting a holistic network perspective rather than focusing on individual components has proven beneficial in water distribution networks (WDNs). For example, Li et al. [13] present a novel gated GNN model for real-time water quality prediction in WDNs, addressing challenges posed by limited sensor data and complex system dynamics. Meanwhile, Truong et al. [24] address pressure estimation in WDNs by combining physics-based modeling with GNNs, introducing a robust training strategy based on random sensor placement to handle unexpected sensor changes. The approach was tested on a WDN DT in Oosterbeek, Netherlands [7]. More recently, graph-based spatio-temporal architectures have also been applied to forecasting problems in other infrastructure domains; Tchalla [23] combines dynamic graph learning with multi-scale spatio-temporal attention for traffic forecasting, showing the broader applicability of graph-based models.

The study highlights the promising integration of GNNs with DTs for improved system-wide state estimation. However, current wastewater system approaches rarely consider the overall condition of the network, revealing an important gap for future research.

(4) Digital Twins for Water Infrastructure: Toward Graph-Based Integration. The potential of graph-based learning methods with DT has been shown in research from Homaei et al. [11] and Ngo et al. [17]. Homaei et al. [11] focus on proposing an intelligent DT platform, named CAUCCES, designed to enhance the management of Water Distribution Systems. By integrating real-time data from IoT sensors with various AI models such as LSTM, Prophet, LightGBM, and XGBoost, the platform enables accurate forecasting of water consumption. On the other hand, Ngo et al. [17] highlight the role of GNNs in enabling DT functionality, particularly in capturing the structural and spatial relationships within water infrastructure. Similarly, Roudbari et al. [22] build a GNN-based model for river-based flood forecasting in a surface water system and develop a simulation prototype offering visual insights for decision-making in disaster prevention and policy-making.

(5) Robustness to Sensor Drift and Noise: Toward Reliable Deployment. A parallel line of work addresses a failure mode distinct from missing sensors: sensors that remain in place but report values that have drifted from ground truth due to bio-fouling, sediment accumulation, or gradual miscalibration. Hasan et al. [10] propose a Wasserstein GAN-based digital-twin model to detect early drift faults in wireless sensor networks before they escalate into outright failures. In the water-network setting specifically, Mücke et al. [16] use generative deep learning within a probabilistic digital twin to localize leaks from sparse, noisy pressure measurements, explicitly propagating measurement uncertainty rather than assuming clean sensor inputs; a follow-up study by Pandey et al. [18] extends this framework to anomaly detection and localization across a water distribution network digital twin. More broadly, deliberately injecting noise into training inputs, so that a model learns invariance to the corruption rather than memorizing clean data, has been shown to improve robustness in other domains, such as image classification [1]. These works establish that sensor degradation is a distinct failure mode and that exposure to corrupted inputs during training is an effective mitigation; however, none combine this with sparse-sensing (missing-sensor) training for wastewater or stormwater state estimation. We address this gap by jointly modeling sensor coverage and sensor drift within a single GNN surrogate.

State estimation under uncertainty is likewise an active concern beyond water infrastructure; Aslami et al. [2] model power system frequency dynamics and state estimation using neural ordinary differential equations combined with soft actor-critic control, reflecting a broader trend toward learned state-estimation and control methods across cyber-physical infrastructure domains.

Collectively, these studies demonstrate that no single method suffices for effective monitoring of wastewater infrastructures. Physics-based surrogates reduce simulation complexity while retaining interpretability, machine learning helps bridge gaps caused by sparse sensing, and graph neural networks capture the spatial dependencies critical for system-wide modeling. However, most existing works focus either on isolated processes (e.g., treatment plants, sludge bulking) or on clean water distribution networks, leaving

an open opportunity for approaches that extend GNN-based methods to wastewater networks. Most existing GNN work, moreover, assumes clean, undistorted sensor inputs, leaving a further opportunity for approaches that jointly handle sparse *and* noisy sensing, which we pursue in Section 4.3.1.

3 PROPOSED PIPELINE

The overall pipeline architecture and methodology used in this study, from dataset generation to model design and training can be seen in Figure 1. The topology of the system (e.g., pipes, manholes) alongside static and dynamic parameters such as roughness, water levels, and infiltration rates, serve as inputs to the pipeline. These elements are fed into InfoWorks ICM (Integrated Catchment Modeling), a comprehensive platform for urban drainage and water network modeling [29, 30], which generates a synthetic dataset simulating various hydraulic behaviors under different conditions. This dataset captures diverse system states over time, which are used as independent snapshots for spatial state estimation.

Following the simulation, the dataset is preprocessed to clean, normalize, and format the data into graph snapshots that represent the state of the network at different timesteps ($T_1, T_2, ..., T_n$). These snapshots are used as inputs to a GNN training module, where a masking strategy is applied to hide certain node values (e.g., water levels) during training. The GNN learns to infer or reconstruct the masked values based on the known parts of the graph. The model's predictions are then compared to the ground truth (clear target), and the loss is minimized iteratively. The performance is finally evaluated through model validation, completing the pipeline for robust and generalizable spatial state estimation in urban water networks.

4 IMPLEMENTATION ON A USE CASE

We describe the implementation of the proposed pipeline on Heusden area in the Netherlands. We begin with dataset generation, which involves hydraulic simulations using a physically realistic urban drainage model. And also add an additional dataset from [28] which is a Shunqing area in China generated from SWMM for additional experiment. The resulting synthetic data is then preprocessed and structured into graph snapshots suitable for GNN. Finally, we describe the GNN architecture used for water level estimation, discussing its input format, network structure, and training methodology, including the sensor-drift noise injection introduced in this paper.

4.1 Dataset

4.1.1 *Heusden*

For the first dataset, this paper tried to create a synthetic yet realistic dataset that captures the dynamic behavior of a wastewater system under varying conditions, enabling robust training and evaluation of GNN models for water level sensor data estimation using InfoWorks ICM. The wastewater network model is based on the infrastructure from the Heusden municipality in the Netherlands, including key elements such as outfalls, manholes, pipes, and pumping stations. The model captures the structural and spatial complexity of a real-world urban drainage system, making it a suitable basis for testing data-driven estimation methods. The structural details of the network are summarized in Table 1.

Table 1: Structural details of the network within Heusden, the Netherlands.

Nodes		***Links***	
Entities	***Amount***	***Entities***	***Amount***
Manholes	873	Pipes	928
		Channels	14
Storages	9	Orifices	1
		Pumps	13
Outfalls	12	Weirs	11
		Flap Valves	1
Total Nodes	894	***Total Links***	968

To ensure the dataset captured a wide range of system responses, we simulated multiple rainfall events following classifications from the Dutch technical guidelines, specifically Rainfall Events #01 to #10 [19]. These events vary in intensity and duration, allowing us to assess model performance under different hydraulic loading conditions. Additionally, we varied several parameters commonly used in hydraulic model calibration tasks, as listed in Table 2, following guidance from [6]. Other parameters, such as pipe diameters, slopes, and network connectivity, were configured in consultation with domain experts in water engineering.

The simulated outputs consist of water level values at various nodes, which serve as the target variables for our state estimation model. Data were recorded at 5-minute intervals to capture short-term variations in water levels. Through the variety of rainfall events, the dataset includes both normal and extreme system conditions, providing a comprehensive foundation for training and evaluation.

Although the simulations produce time series data, each timestep is treated as an independent system state for model training: the model receives no explicit information from previous timesteps, and each snapshot is presented to the GNN as a standalone graph state. Using different rainfall events and the parameter variations listed in Table 2, we generated a total of 255 unique InfoWorks ICM simulation scenarios, resulting in 19,320 different snapshots.

To evaluate generalization across hydraulic scenarios, we split the dataset by rainfall event ("bui") index, ensuring each event is assigned exclusively to either training (9,660 states), validation (3,864 states), or testing (5,796 states) without any snapshot overlap.

Each rainfall scenario (Bui#01–Bui#10) represents a design rain event characterized by a specific *return period*, which indicates how often an event of that intensity is expected to occur. Based on standard hydrological classifications [19], we assign the following intensity categories:

- **Low intensity:** Bui#01–Bui#04 (0.25–0.5 year return period)
- **Medium intensity:** Bui#05–Bui#06 (1 year)
- **High intensity:** Bui#07–Bui#08 (2 years)
- **Extreme intensity:** Bui#09–Bui#10 (5 and 10 years)

Using this classification, we define our data split as follows:

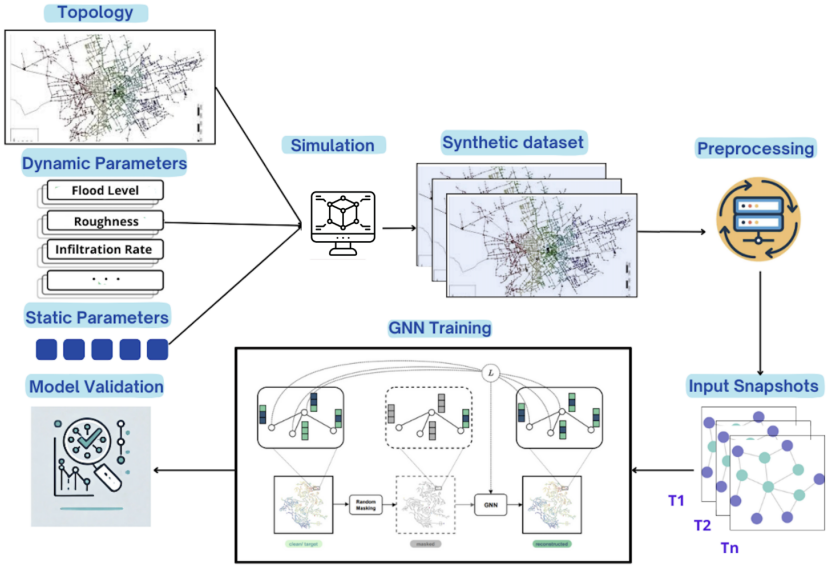


Figure 1: Proposed pipeline of GNN-based state estimation for urban wastewater digital twin.

Table 2: Possible parameter values.

Parameters	Values
Bottom Roughness	Ranged between 2.0 mm and 8.0 mm
Top Roughness	Ranged between 2.0 mm and 8.0 mm
Discharge Coefficient	Ranged between 0.2 and 2.0
Initial Infiltration	Ranged between 2 mm/h and 20 mm/h

- **Training:** Bui#01 (low), Bui#03 (low), Bui#05 (medium), Bui#07 (high), Bui#09 (extreme)
- **Validation:** Bui#04 (low), Bui#06 (medium)
- **Testing:** Bui#02 (low), Bui#08 (high), Bui#10 (extreme)

This split provides a balanced representation of rainfall intensities across all subsets while maintaining strict separation between scenarios, allowing us to test how well the model generalizes to unseen hydraulic conditions.

4.1.2 *Shunqing*

A stormwater network from Shunqing, Sichuan Province, China [28] is used as an additional case study to evaluate the hydraulic estimation performance of the surrogate models. The Shunqing model, developed in SWMM, serves as the physics-based reference and consists of 105 manholes, 8 outfalls, and 131 conduits. The dataset contains 148 rainfall-event records with durations ranging from 6 to 24 hours. These events were used as rainfall inputs to SWMM, and the resulting simulated hydraulic states serve as labels in this study.

4.2 Pre-processing

The raw simulation outputs from InfoWorks (Heusden) and SWMM (Shunqing) are converted into graph snapshots suitable for GNN training. Each row in the simulation output corresponds to the hydraulic state of a network element at a given timestep. Two preprocessing steps are applied before feeding the data to the model.

Outfall completion. For the Heusden dataset, InfoWorks does not compute valid values for the water level at outfall nodes, which appear as NaN or negative entries. To avoid disconnecting parts of the graph, missing values are completed using a k-nearest-neighbor ($k = 5$) imputation based on nearby nodes, and a binary mask is retained to mark imputed entries. The Shunqing dataset, in contrast, provides complete hydraulic states directly from SWMM simulations and does not require this step.

Feature normalization. For Heusden, the main target variable is **flood depth**, which accounts for terrain elevation and is therefore more meaningful than absolute water levels. It is computed as:

$$\text{FloodDepth}(t) = \text{Level}(t) - \text{FloodLevel}. \tag{1}$$

For Shunqing, the target variable is **cumulative flood volume** as provided in the dataset. For both datasets, values are standardized to zero mean and unit variance using statistics from the training split only, which are reused during validation and testing to avoid leakage.

4.3 Methodology

This study adapts the **GATRes** model [24], originally developed for water distribution networks, to wastewater and stormwater systems. We use GATRes as a surrogate learning framework without modifying its architecture, focusing on its applicability to heterogeneous urban drainage systems. For architectural details, we refer readers to the original paper.

Graph representation. Each network is represented as a directed graph $G = (V, E)$, where nodes correspond to manholes, storage, or outfalls, and edges to physical links such as pipes or channels. Node

values represent observed or unobserved hydraulic states (flood depth for Heusden, cumulative volume for Shunqing).

Learning setup and masking. To emulate sparse sensing, a fixed subset of nodes $\mathcal{V}_{\text{obs}} \subset V$ is designated as sensors, and their values are provided as input. The model is trained to predict the states of all other nodes $v_j \in V \setminus \mathcal{V}_{\text{obs}}$. A fixed masking strategy ensures consistency across timesteps within each scenario. Training minimizes the mean absolute error between predicted and simulated values.

This setup allows us to test the transferability of a GNN surrogate across different datasets, simulation frameworks, and performance targets, while evaluating its robustness under sparse sensing conditions. In this paper, we extend it to also evaluate and improve robustness under *noisy* sensing, described next.

4.3.1 Sensor-Drift Noise Injection

The masking protocol above emulates sparse *coverage*: some nodes have no sensor and their values must therefore be inferred from the available observations and graph structure. This setup does not represent a distinct failure mode in which a sensor remains available but gradually reports biased measurements. Such drift can arise from sensor ageing, fouling, sediment accumulation, or calibration changes. Since the hydraulic simulation data used for training are clean by construction, a model trained only with masking is never explicitly exposed to corrupted sensor readings. We therefore extend the sparse-sensing pipeline with dataset-specific drift models and use them both to quantify sensitivity to corrupted observations and to train a corruption-aware variant of the same GATRes architecture.

Figure 2 summarizes the training and evaluation procedure used for the sensor-drift robustness study. The original clean-training pipeline is compared with a corruption-aware training strategy in which 50% of the observed sensor inputs are corrupted during training. Both models retain the same GATRes architecture and hidden-node estimation objective. After training, the clean and corruption-aware models are evaluated under the same test-time drift protocol by progressively increasing the fraction of corrupted observed sensors and comparing the resulting MAE and relative MAE growth on the hidden nodes.

Heusden corruption model. For Heusden, the estimated variable is flood depth and sensor drift is represented as a time-varying additive offset. For a corrupted observed sensor i, the input is modified as

$$\tilde{x}_i(t) = x_i(t) + \beta_{0,i} + B_i(t), \tag{2}$$

where $x_i(t)$ is the clean simulated flood-depth value, $\beta_{0,i}$ is a sensor-specific initial bias, and $B_i(t)$ is a cumulative random-walk component. The temporal component is defined as

$$B_i(0) = 0, \quad B_i(t) = B_i(t-1) + \epsilon_i(t), \quad \epsilon_i(t) \sim \mathcal{N}(0, \sigma_B^2). \tag{3}$$

The initial offset is sampled within $\beta_{0,i} \in [-\beta_{\max}, \beta_{\max}]$, with $\beta_{\max} = 0.25$ m, while the random-walk increment standard deviation is fixed to $\sigma_B = 0.01$ m per five-minute timestep. The corruption is applied only to the observed sensor subset $\mathcal{V}_{\text{obs}}$. Values at hidden nodes $\mathcal{V}_{\text{unobs}}$ remain masked exactly as in the original sparse-sensing setup, and the regression targets are never corrupted. The offset is introduced in physical flood-depth units before applying the same normalization statistics derived from the training split. These synthetic perturbations represent measurement bias that already exists during a rainfall event. They do not simulate the long-term physical development of sensor drift.

The additive formulation represents the effect of a biased sensor reading, while the random-walk component introduces variation in that bias over consecutive timesteps. Wastewater sensor-fault studies have similarly represented drift as an increasing additive deviation from the normal measurement, while random-walk formulations provide a stochastic model for gradual sensor drift. Although the random-walk corruption changed across consecutive timesteps, each resulting corrupted snapshot is processed independently by the GatRes model.

Shunqing corruption model. Shunqing's target variable is a discrete, event-driven flood volume that is exactly zero during dry weather; an additive noise model would introduce non-physical volume during these periods. We therefore use a multiplicative corruption that vanishes wherever the true reading is zero:

$$\tilde{x}_i(t) = x_i(t) \cdot (1 + \gamma_i), \qquad \gamma_i \sim \mathcal{U}(-\gamma_{\max}, \gamma_{\max}) \tag{4}$$

with $\gamma_{\max} = 0.25$, drawn independently per corrupted node. Corruption is restricted to the visible sensor subset $\mathcal{V}_{\text{obs}}$ and applied in raw, pre-normalization feature space: because the StandardScaler transform $z = (x - \mu)/\sigma$ does not commute with multiplicative drift unless $\mu = 0$, drift must be injected before normalization to remain physically meaningful. A random fraction p_{drift} of $\mathcal{V}_{\text{obs}}$ is corrupted per sample; the remaining sensors, and all of $\mathcal{V}_{\text{unobs}}$, are left untouched, so the model has access to a mix of trustworthy and corrupted readings from which to learn.

Experimental design. The same high-level robustness experiment is applied to both the Heusden and Shunqing case studies. In each dataset, we compare two training conditions under 70% masking (30% sensor coverage):

- **Clean baseline.** The model is trained using the original sparse-sensing masking protocol without sensor-drift augmentation. At test time, sensor drift is introduced into a fraction of the observed sensor inputs to quantify the vulnerability of a model that has only been exposed to clean observations during training.
- **Corruption-aware training.** The same GATRes architecture is trained on the same state-estimation task, but sensor drift is injected into $p_{\text{drift,train}} = 0.5$ of the observed sensor inputs during training. A fresh corruption realization is sampled for every mini-batch so that the model is exposed to varying corrupted sensor subsets and drift values rather than a single fixed corruption pattern.

For each dataset, the Clean and corruption-aware models use the same GATRes architecture, feature normalization, sensor-coverage setting, and dataset partitioning used for that case study. The corruption-aware model does not explicitly reconstruct or correct the corrupted sensor readings. Instead, it learns the original hidden-node state-estimation task while part of the available sensor information is unreliable.

The training fraction $p_{\text{drift,train}} = 0.5$ is used as a fixed intermediate corruption condition rather than an optimized hyperparameter.

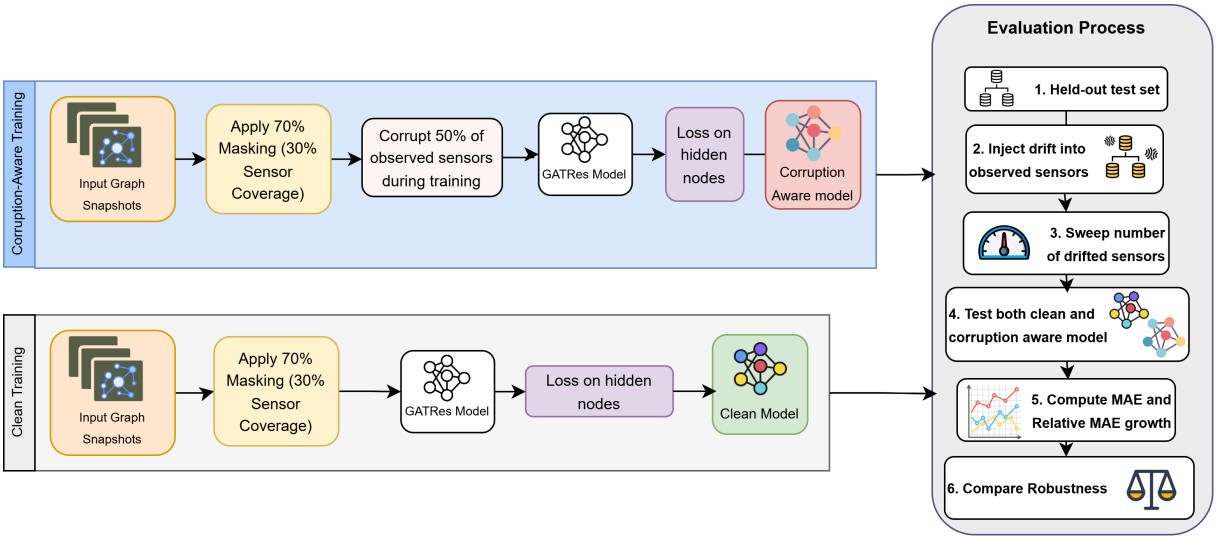


Figure 2: Training and evaluation pipeline for the sensor-drift robustness exploration.

The purpose is to investigate whether exposure to a moderate fraction of corrupted sensors during training improves robustness when the fraction of drifting sensors changes at test time.

Test-time corruption protocol. For both datasets, the fraction of observed sensors affected by drift is evaluated at

$$p_{\text{drift}} \in \{0,\ 0.3,\ 0.5,\ 0.7,\ 1.0\}. \tag{5}$$

Here, p_{drift} represents the fraction of the available sensors whose measurements are corrupted, rather than the magnitude of the corruption itself. The corruption magnitude is controlled separately by the dataset-specific drift parameters defined above.

Both case studies are evaluated using three independently generated random sensor placements to reduce dependence on a single sensor configuration. Performance is measured only on the unobserved nodes $\mathcal{V}_{\text{unobs}}$, i.e., the nodes whose states must be reconstructed from the available sensor observations and graph structure.

The dataset-specific evaluation setup follows the corresponding data organization of each case study. For Heusden, the 70% masking configuration contains 268 observed and 625 unobserved nodes, and robustness is evaluated on the held-out rainfall events Bui#02, Bui#08, and Bui#10, with ten corruption realizations for each test corruption fraction. The Heusden drift additionally evolves over consecutive timesteps through the random-walk term in Eq. 3. For Shunqing, the same Clean-versus-corruption-aware comparison and p_{drift} sweep are evaluated using the Shunqing data partition described previously and the multiplicative corruption model in Eq. 4.

5 RESULTS

This section presents the evaluation of the proposed GNN surrogate, with the Heusden wastewater network as the primary case study and the Shunqing stormwater dataset as an external validation benchmark. The analysis focuses on the Heusden network's spatial error patterns, temporal dynamics, and operational performance, followed by validation results from the Shunqing dataset to demonstrate cross-infrastructure generalization.

5.1 Performance on the Heusden Network

The Heusden case study represents a realistic Dutch urban drainage network developed in collaboration with the local municipality. Across the evaluated rainfall scenarios, the magnitude of the estimation error varies across nodes and hydraulic conditions: at non-flooding manholes, estimation errors for maximum water levels reach up to 50 cm, a range validated by domain experts as operationally sufficient for decision-making in smart city infrastructures, whereas much tighter tolerances (0–10 cm) are expected at weir locations, where larger deviations typically indicate backflow effects or local losses rather than acceptable estimation noise. Figure 3 reports the spatial distribution of node-level MAE under 30% sensor coverage.

The spatial analysis reveals distinct error patterns across the network. Figure 3 shows the spatial distribution of MAE under 30% sensor coverage (268 sensors out of 894 nodes). Most nodes display low estimation error, indicating effective spatial interpolation capabilities of the graph-based approach. Higher MAE values are concentrated near hydraulic control structures (pumps, weirs) and at outfalls, where nonlinear behavior and strong dependence on upstream flows create estimation challenges. These localized errors indicate that performance should be interpreted according to location and hydraulic condition, particularly near control structures, outfalls, and flooding nodes.

5.2 Node-level Dynamics in Heusden

The temporal analysis examines how well the surrogate captures flood wave propagation and peak dynamics across the network. Figure 4 compares GNN predictions with InfoWorks simulations at two representative nodes, revealing the model's ability to track both routine and extreme hydraulic conditions.

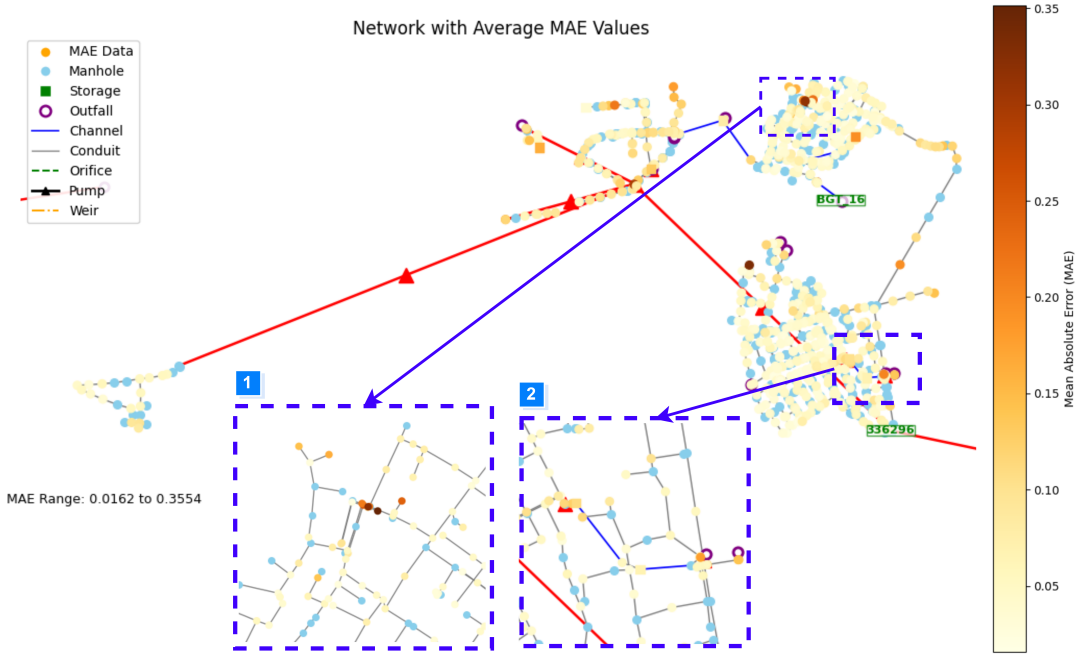


Figure 3: Spatial distribution of MAE across the Heusden network under 30% sensor coverage.

At Node 335242 (panel a), the model captures both rising and falling limbs of the hydrograph, reproducing multi-peak structures and confirming its ability to track flood wave propagation through the network topology. At Node 335169 (panel b), the surrogate follows the overall pattern but underestimates sharp peaks, indicating sensitivity to highly dynamic conditions that require additional training diversity. These results demonstrate the strengths of graph-based learning in capturing spatiotemporal propagation while highlighting areas for improvement in extreme event prediction.

5.3 External Validation on the Shunqing Network

To evaluate cross-infrastructure generalization, the surrogate was applied to the Shunqing stormwater dataset from Sichuan Province, China. This external validation serves to confirm that the approach is not limited to the Heusden network or InfoWorks simulation framework. Unlike Heusden, which emphasizes flood depth in a large gravity-driven system, Shunqing focuses on cumulative flood volumes in a smaller stormwater network developed using SWMM.

Table 3 reports predictive performance under the 70% masking setup (only 30% of sensors available). The surrogate achieves a mean absolute error of 0.453 m^3 with RMSE of 2.145 m^3, demonstrating robust estimation accuracy even under severe sensor scarcity.

Figure 5 shows the evolution of MAE as a function of sensor coverage (10–40%). The best performance occurs at 30% sensor availability (70% masking), consistent with the Heusden findings. Figure 6 complements this with the spatial distribution of node-level MAE under the same configuration, revealing that errors are well-contained across the network despite limited sensing. These results

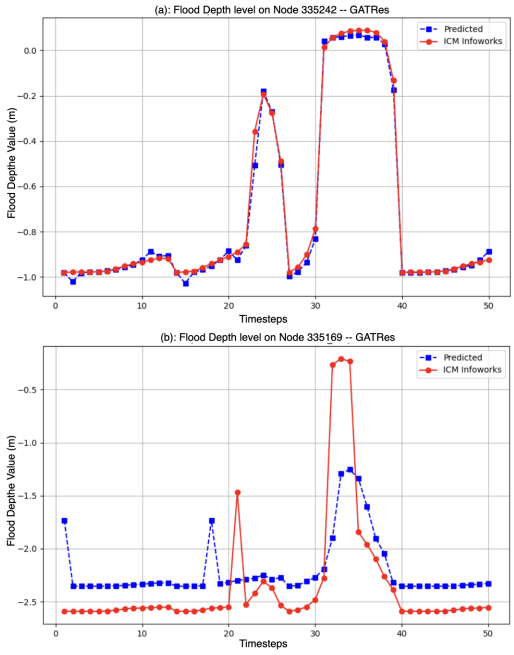


Figure 4: Comparison between GNN predictions and InfoWorks simulations in Heusden at (a) Node 335242 and (b) Node 335169.

demonstrate that the *same surrogate pipeline* can be instantiated on a distinct network and simulation framework with strong accuracy under sparse sensing. This replication across datasets indicates

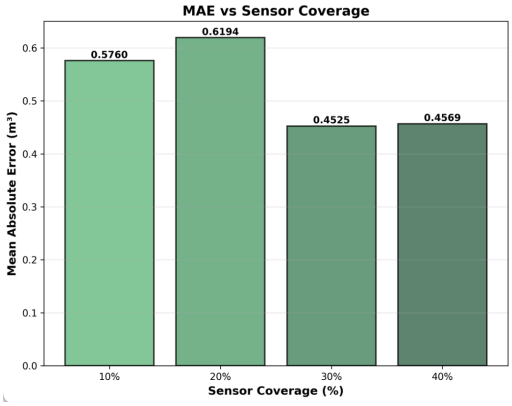


Figure 5: Impact of sensor coverage on MAE for the Shunqing network. Best performance at 30% sensor availability (70% masking).

portability of the approach and supports its use in varied urban drainage contexts.

Table 3: Performance of GATRes on the Shunqing dataset (70% masking).

Sensor Availability	MAE (m^3)	RMSE (m^3)
30% (70% masking)	**0.453**	2.145

To contextualize these values, the Shunqing dataset exhibits flood volumes ranging from 0 to 164.36 m^3, with a mean of 20.48 m^3 and median of 14.41 m^3. Out of 10.5 million simulation records, 21% correspond to non-zero flooding events (2.2 million points). Under 70% masking, the MAE of 0.453 m^3 corresponds to approximately 2.2% of the mean flood volume, while even at 90% masking the MAE remains below 3%. This indicates that errors are small relative to the typical scale of flood volumes, reinforcing the robustness of the surrogate under sparse sensing conditions.

5.4 Heusden: Robustness to Sensor Drift

Table 4 reports the MAE on $\mathcal{V}_{\text{unobs}}$ as the fraction of corrupted observed sensors increases, for both the clean-trained Baseline and the corruption-aware Denoising model. Results are reported as mean ± standard deviation over $N = 3$ random sensor placements, after aggregating performance across the held-out rainfall events Bui#02, Bui#08, and Bui#10.

Under the clean condition ($p_{\text{drift}} = 0$), the Baseline achieves a slightly lower MAE than the Denoising model (0.068 m versus 0.077 m): a small nominal-accuracy cost from corruption-aware training when all available sensor measurements are reliable. This behavior changes once sensor drift is introduced. At $p_{\text{drift}} = 0.3$, the Baseline MAE increases to 0.112 m, while the Denoising model remains at 0.084 m. The difference continues to increase as a larger fraction of sensors becomes corrupted. When all observed sensors

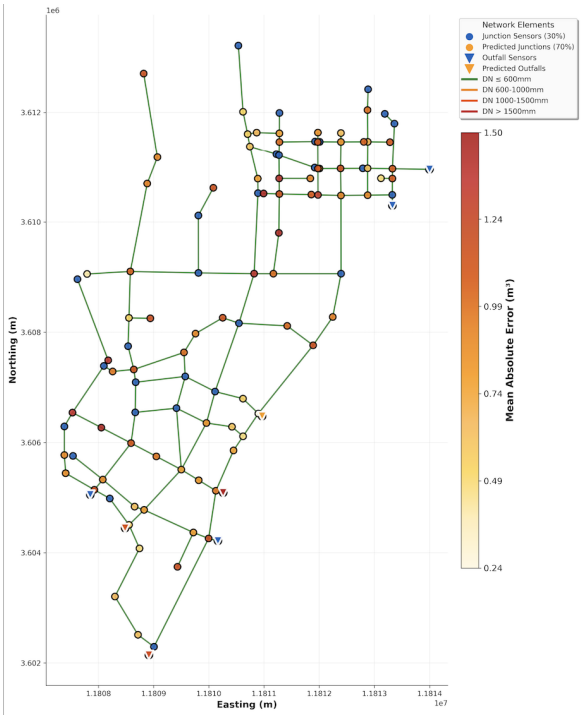


Figure 6: Node-level MAE distribution in Shunqing under 70% masking (30% sensors available).

Table 4: MAE (m) on hidden nodes under sensor drift, Heusden, 30% sensor coverage. $N = 3$ random sensor placements.

p_{drift} (corruption fraction)	Baseline	Denoising
0% (none)	0.068 ± 0.019	0.077 ± 0.025
30%	0.112 ± 0.014	0.084 ± 0.025
50%	0.137 ± 0.013	0.090 ± 0.027
70%	0.156 ± 0.012	0.096 ± 0.027
100% (all)	0.181 ± 0.011	0.104 ± 0.028

are affected by drift, the Baseline reaches an MAE of 0.181 m, compared with 0.104 m for the Denoising model, a 42% lower absolute MAE.

Figure 7 shows the same trend across the complete corruption-fraction sweep. The Baseline error increases sharply as more observed sensors become unreliable, while the Denoising model's increase is much flatter. The uncertainty bands show that absolute performance depends on the particular random sensor placement, especially for the Denoising model.

Because absolute MAE is influenced by the nominal accuracy of each sensor placement, we additionally examine the relative MAE growth with respect to each model and placement's own $p_{\text{drift}} = 0$ condition. This isolates the sensitivity to sensor corruption from differences in the initial clean-condition error. Table 5 reports the relative growth at full corruption for the three random sensor placements.

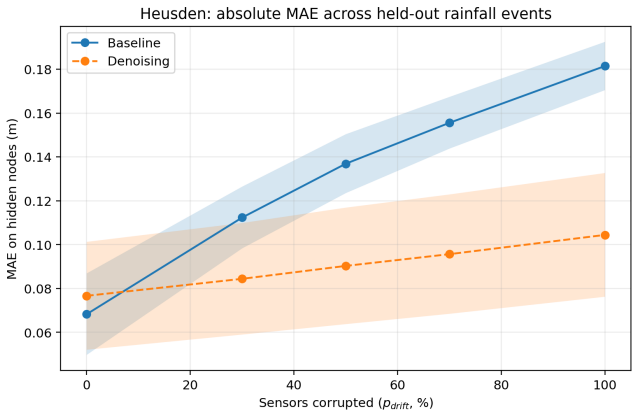


Figure 7: Absolute MAE on hidden nodes in Heusden as the fraction of corrupted observed sensors increases. Values are aggregated across the held-out rainfall events Bui#02, Bui#08, and Bui#10 and shown as mean ± standard deviation over three random sensor placements.

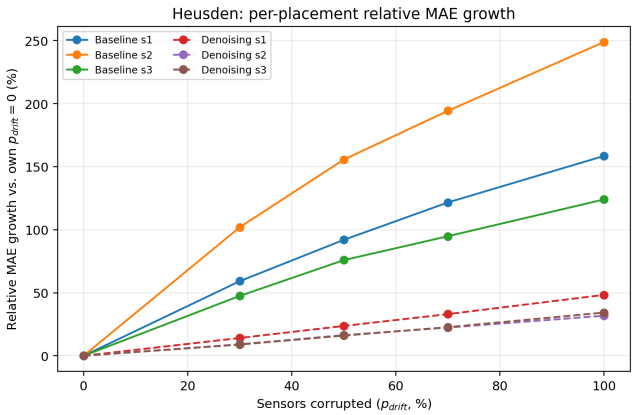


Figure 8: Relative MAE growth in Heusden for each random sensor placement, normalized to the corresponding model and placement's own $p_{\text{drift}} = 0$ performance.

Table 5: Relative MAE growth from $p_{\text{drift}} = 0$ to $p_{\text{drift}} = 1.0$ for each random sensor placement in Heusden.

Model	Placement s1	Placement s2	Placement s3
Baseline	+158.6%	+249.0%	+124.1%
Denoising	+48.3%	+31.8%	+34.3%

Figure 8 shows the relative MAE growth over the complete test sweep for each placement. The Baseline is consistently sensitive to sensor drift, although the magnitude of degradation varies widely across placements. At full corruption, its relative MAE growth ranges from approximately 124% to 249%. In contrast, the Denoising model shows markedly flatter and more closely grouped degradation curves, with full-corruption growth of approximately 32–48%. Thus, even the most drift-sensitive Denoising placement degrades considerably less than the least sensitive Baseline placement.

When relative growth is aggregated across the three placements, the same distinction remains clear. At full corruption, the Baseline increases by 177.2 ± 64.5% relative to its own clean-condition performance, while the Denoising model increases by only 38.1 ± 8.9%. The smaller relative growth and lower variation across placements show that corruption-aware training reduces the average sensitivity to sensor drift and makes the model's response less dependent on the particular random sensor configuration.

Overall, the Heusden results reveal a trade-off between nominal accuracy and robustness. Corruption-aware training introduces a small increase in MAE when no sensor drift is present, but provides a large advantage as soon as part of the observed sensor set becomes corrupted. This benefit is already visible at $p_{\text{drift}} = 0.3$ and persists up to complete corruption, showing that exposure to unreliable sensor inputs during training improves GATRes robustness across random sensor placements and held-out rainfall conditions.

Table 6: MAE (m^3) on hidden nodes under sensor drift, Shunqing, 30% sensor coverage. $N = 3$ seeds.

p_{drift} (corruption fraction)	Baseline	Denoising
0% (none)	1.305 ± 0.955	1.294 ± 0.767
30%	1.461 ± 0.942	1.321 ± 0.781
50%	1.565 ± 0.930	1.340 ± 0.790
70%	1.667 ± 0.917	1.358 ± 0.799
100% (all)	1.806 ± 0.895	1.382 ± 0.810

Table 7: Relative MAE growth from $p_{\text{drift}} = 0$ to $p_{\text{drift}} = 1.0$, per seed.

Model	Seed 42	Seed 43	Seed 44
Baseline	+79.0%	+93.7%	+15.7%
Denoising	+8.2%	+7.2%	+6.3%

5.5 Shunqing: Robustness to Sensor Drift

Table 6 reports MAE on $\mathcal{V}_{\text{unobs}}$ as the corruption fraction p_{drift} increases, for both the clean-trained baseline and the denoising-trained model of Section 4.3.1 (mean ± std over $N = 3$ random seeds).

The absolute std values above are large relative to the mean, driven almost entirely by one of the three seeds whose *clean-condition* accuracy was already much worse than the other two (test R^2 of 0.67 versus 0.95–0.97), most likely due to which random sensor placement that seed happened to draw, consistent with the sensitivity to sensor placement noted for this network in our SAC '26 paper's masking evaluation. Because seed-to-seed variance in absolute accuracy is large, the *relative* error growth from $p_{\text{drift}} = 0$ to $p_{\text{drift}} = 1.0$ is more informative and is reported per seed in Table 7.

Figures 9 and 10 visualize both views. Figure 9 shows the absolute MAE trend (mean ± std band) for each model; the wide,

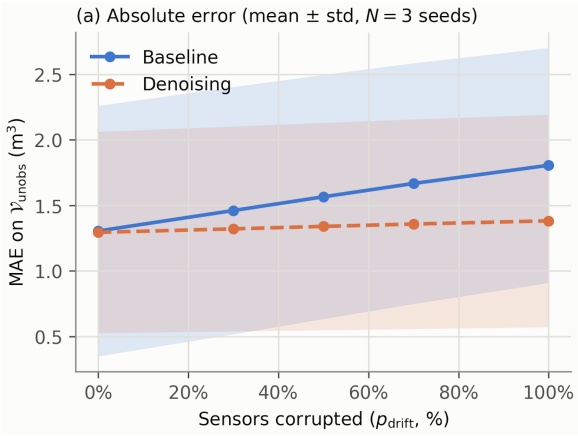


Figure 9: Shunqing sensor-drift robustness, $N = 3$ seeds: absolute MAE on $\mathcal{V}_{\text{unobs}}$, mean ± std.

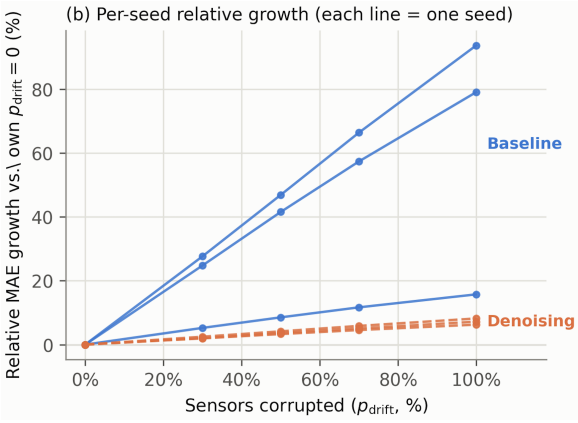


Figure 10: Shunqing sensor-drift robustness, $N = 3$ seeds: relative MAE growth per seed, normalized to that seed's own $p_{\text{drift}} = 0$ value. Baseline lines fan out erratically; denoising lines stay tightly clustered.

overlapping bands reflect the same large seed-to-seed variance in absolute accuracy discussed above, not noise in the measurement. Figure 10 isolates the relative-growth comparison per seed. The three baseline lines fan out sharply and inconsistently, while the three denoising lines stay tightly clustered near zero across the entire drift range.

The baseline model, which never observed corrupted sensors during training, degrades sharply as the corruption fraction increases in two of the three seeds (+79–94% relative MAE from $p_{\text{drift}} = 0$ to $p_{\text{drift}} = 1.0$). This confirms that message passing propagates sensor corruption outward to unobserved nodes when left unaddressed. The third seed shows smaller relative growth (+15.7%), but only because that run's clean-condition accuracy was already poor: its absolute MAE stays worse than the denoising model's at every drift level, including $p_{\text{drift}} = 0$, leaving less room for it to degrade further in relative terms. The denoising-trained model behaves differently and more consistently. Its relative MAE growth stays small

Table 8: Relative MAE growth at $p_{\text{drift}} = 1.0$ across drift severities, per seed.

Model	$\gamma_{\max}$	Seed 42	Seed 43	Seed 44
Baseline	0.10	+20.5%	+20.3%	+3.8%
	0.25	+79.0%	+93.7%	+15.7%
	0.50	+213.0%	+256.3%	+42.5%
Denoising	0.10	+1.3%	+1.0%	+1.1%
	0.25	+8.2%	+7.2%	+6.3%
	0.50	+43.7%	+46.0%	+23.9%

across all three seeds (+6.3% to +8.2%), a far tighter range than the baseline's. At $p_{\text{drift}} = 0$, the denoising model is comparable to or better than the baseline depending on seed, so there is no consistent accuracy cost from training under augmented noise for this network, and it is more accurate at high corruption levels in every seed tested. Exposing the network to sensor drift during training, rather than only to missing sensors, does more than reduce average degradation under drift. It makes the model's response to drift reliable across independent training runs, where the undefended baseline's response is erratic and depends heavily on incidental factors such as sensor placement.

Drift severity sweep. The results above fix the corruption bound at $\gamma_{\max} = 0.25$ (Section 4.3.1). Table 8 additionally sweeps $\gamma_{\max} \in \{0.10, 0.25, 0.50\}$ (mild, default, severe multiplicative drift) at $p_{\text{drift}} = 1.0$ (all sensors corrupted, the worst-case coverage), reusing the same checkpoints. The pattern holds at every severity level: at mild drift ($\gamma_{\max} = 0.10$), the denoising model is essentially unaffected (+1.0–1.3% across all three seeds) while the baseline already shows meaningful, uneven degradation (+3.8–20.5%); at severe drift ($\gamma_{\max} = 0.50$), both models degrade far more, but the baseline's worst seeds (+213%, +256%) are an order of magnitude beyond the denoising model's worst case (+46%). The two models' relative growth converges somewhat at this extreme. Denoising training is not a substitute for bounding how corrupted a sensor can plausibly become, but the consistency advantage persists throughout the tested range.

6 DISCUSSION

The experiments highlight the value of graph neural networks as surrogates for wastewater digital twins. Five key themes emerge: calibration and learning challenges, behavior under extreme versus routine conditions, generalization across networks, robustness to sensor drift, and broader implications for smart infrastructures.

6.1 Implications for Calibration and Learning

The Heusden results reveal important insights for calibration and learning in wastewater digital twins. Estimation accuracy is not uniform across the network, with errors clustering near pumps, weirs, and outfalls. This pattern may reflect localized nonlinear hydraulic dynamics near pumps, weirs, and outfalls, combined with limited observational information from the surrounding sensors.

The spatial error distribution suggests that calibration strategies for wastewater DTs should explicitly incorporate hydraulic structures and flow paths when retraining or adapting models, potentially requiring weighted loss functions or targeted training data augmentation near critical infrastructure elements.

The learning process benefits from the graph structure's ability to propagate information across the network topology. However, the concentration of errors near control structures indicates that the current approach may need additional training diversity or specialized attention mechanisms to better capture the nonlinear behavior at these locations. Future work could explore adaptive masking strategies that provide more training exposure to challenging network regions. In addition, extending GATRes to a *heterogeneous variant* that explicitly models different node and edge types (e.g., pumps, weirs, outfalls versus conduits and manholes) could further improve estimation accuracy by incorporating the distinct hydraulic roles of these components.

Equally important, the ability to produce reliable estimates with only 30% sensor coverage illustrates the surrogate's value as a virtual sensing mechanism. This capability supports cost-efficient monitoring strategies, where installing dense sensor networks is often infeasible for municipalities managing large wastewater infrastructures.

6.2 Performance under Extreme and Routine Conditions

The analysis reveals distinct performance characteristics under different operational conditions. From an operational perspective, peak conditions are critical for emergency response and flood management. The surrogate successfully captures flood wave timing and general propagation patterns, but amplitude errors remain during sharp transitions, particularly at nodes with complex hydraulic behavior. These limitations point to the need for training on a wider variety of rainfall scenarios, particularly extreme events with rapid intensity changes that challenge the model's ability to capture nonlinear dynamics.

Under routine conditions, the surrogate delivers consistent and stable estimates, supporting applications in daily monitoring and operational decision-making. The temporal analysis shows that the model effectively tracks gradual changes in water levels and maintains accuracy during normal flow conditions. This stability is crucial for continuous monitoring applications where false alarms or missed detections could lead to unnecessary operational interventions or missed maintenance opportunities.

The contrast between routine and extreme performance highlights a key trade-off in surrogate modeling: while the approach excels at capturing typical system behavior, it requires additional training diversity to handle the full spectrum of operational conditions. This suggests that future implementations should prioritize comprehensive scenario coverage during training, including both frequent low-intensity events and rare high-intensity storms.

6.3 Generalization across Networks and Frameworks

The Shunqing case study provides complementary validation of the proposed surrogate framework across different network scales and simulation environments. After demonstrating strong performance on the Heusden wastewater system, the same pipeline was retrained and applied to the Shunqing stormwater network, which differs substantially in topology, scale, and target variables. Despite these differences, the model maintained reliable accuracy under sparse sensing, indicating that the graph-based learning approach captures transferable hydraulic patterns relevant across infrastructures.

The application to both InfoWorks ICM (Heusden) and SWMM (Shunqing) highlights the adaptability of the framework to different simulation tools, an important property for practical deployment given the diversity of modeling practices used by municipalities worldwide. This flexibility suggests that the approach can be adopted in varied contexts, supporting real-time monitoring in both large, complex drainage systems and smaller stormwater networks.

Importantly, this cross-network result should be interpreted as framework portability rather than zero-shot model transfer. The GATRes model is retrained for each network using its corresponding graph structure, simulated data, and target variable. The transferable element is therefore the overall state-estimation pipeline and sparse-sensing formulation rather than a single trained model shared across infrastructures. This distinction is relevant for practical deployment, where network-specific adaptation is still expected even when the same modeling methodology is reused.

At the same time, the results emphasize that dataset-specific preprocessing remains necessary, particularly when switching between different target variables (flood depth in Heusden versus cumulative volume in Shunqing). Future work should explore transfer learning strategies that reduce the reliance on manual adjustments while preserving robustness across infrastructures.

6.4 Robustness to Sensor Drift

Robustness to *missing* sensors does not carry over to robustness against *corrupted* sensors. A masking-only surrogate has no reason to distinguish a corrupted reading from a clean one, because every sensor value it sees during training is correct by construction. Injecting drift during training gives the network both trustworthy and corrupted inputs to weigh differently, and Sections 5.4 and 5.5 show this in two networks with different drift models and different target variables.

The two networks differ in where the benefit appears. On Shunqing, denoising training costs nothing at zero drift: the denoising model is comparable to or better than the baseline depending on seed. On Heusden, denoising training costs a small amount of accuracy at zero drift, 0.077 m versus 0.068 m (Table 4), a 13% increase in MAE when every sensor is trustworthy. That cost disappears quickly: by $p_{\text{drift}} = 0.3$, the denoising model already beats the baseline on Heusden, 0.084 m versus 0.112 m, and the gap widens at every higher drift level.

Both networks show the same pattern in the spread across sensor placements. On Heusden, baseline relative error growth ranges from 124% to 249% across three placements; denoising training brings this down to 32–48%. On Shunqing, baseline growth ranges from 16% to 94%; denoising training brings this down to 6–8%. An operator installing sensors on a new network cannot know in advance which placement they will end up with. A model whose drift response depends heavily on that placement has field behavior that cannot

be predicted before deployment. Denoising training narrows that dependence on both networks, under an additive random-walk drift model for Heusden and a multiplicative drift model for Shunqing, and for two different target variables (flood depth versus cumulative volume).

6.5 Broader Impact for Smart Cities and Critical Infrastructures

By bridging detailed physics-based simulations with graph-based learning, the surrogate provides a scalable and adaptive monitoring tool for smart city infrastructures. In the Heusden case, it enables reliable state estimation without relying exclusively on computationally demanding hydraulic models, thereby supporting cost-effective and responsive infrastructure management.

The approach directly addresses three key challenges in urban wastewater management: limited sensor coverage, the computational cost of detailed hydraulic modeling, and the need for scalable methods that can be adapted to different infrastructures. By producing accurate network-wide estimates from sparse sensor inputs, the surrogate offers a cost-efficient alternative to dense instrumentation or continuous physics-based simulations. This makes it particularly valuable for municipalities operating under budget constraints or with legacy monitoring systems. The sensor-drift results further show that this reliability can be preserved even as real sensors degrade over their operational lifetime, an essential property for any surrogate intended for continuous, unattended deployment rather than one-off evaluation on clean simulation data.

The validation across two distinct case studies, Heusden and Shunqing, reinforces the broader potential of this approach. By demonstrating consistent performance in networks of different scales and simulation frameworks, the surrogate contributes to the development of resilient and adaptive wastewater management strategies, advancing the agenda of smart cities and critical infrastructures.

7 CONCLUSION

This study demonstrates the potential of graph neural networks as surrogates for wastewater monitoring within digital twin ecosystems. Using the Heusden wastewater network as the main case study, the surrogate achieved expert-validated accuracy (errors of up to 50 cm for maximum water levels at non-flooding manholes, with much tighter tolerances of 0–10 cm expected at weir locations) under sparse sensing conditions. Validation on the Shunqing stormwater dataset confirmed its transferability to different infrastructures and modeling frameworks.

By enabling accurate, scalable, and cost-effective state estimation, this work contributes to the development of smart, resilient urban drainage systems. Graph-based surrogates can play a central role in the broader agenda of smart cities and critical infrastructures, bridging the gap between detailed hydraulic simulations and practical real-time monitoring.

Beyond missing sensors, a GNN surrogate trained only for sparse sensor coverage does not generalize reliably to corrupted sensing, on either network. On Shunqing, across three random seeds, relative error growth at unobserved nodes ranges from 16% to 94% as sensor drift increases from none to total. On Heusden, across three random sensor placements, relative error growth ranges from 124% to 249%. In both cases this range depends on incidental factors such as sensor placement rather than on the severity of the drift itself.

Training with the same physically-motivated drift model injected on every batch closes most of this gap on both networks. On Shunqing, relative error growth falls to 6–8% across all three seeds. On Heusden, it falls to 32–48% across three placements, and the denoising model is more accurate than the baseline at every drift level above zero. The one difference between the two networks is at zero drift: denoising training costs nothing on Shunqing but costs a small amount of accuracy on Heusden, 0.077 m versus 0.068 m. On both networks, denoising training is more accurate than the undefended baseline once any drift is present, and its response to drift is consistent across independent training runs where the baseline's is not.

8 LIMITATIONS AND FUTURE WORK

This study demonstrates the potential of graph neural networks as surrogate models for wastewater monitoring, but several aspects offer opportunities for further development.

- **Synthetic data.** The evaluation relied on data generated from hydraulic simulations checked against known critical locations, rather than a fully calibrated model, which provide a controlled test environment. Future work will extend the approach to real sensor data to assess robustness under operational uncertainties and noise.
- **Synthetic sensor corruption.** The experiments model the effects of existing sensor bias, rather than the long-term process through which drift develops. Future work should evaluate naturally occurring drift using continuous operational sensor records.
- **Fault detection and diagnosis.** This study evaluates whether the surrogate remains accurate under biased sensor inputs, but it does not detect corrupted sensors or determine the cause of unusual measurements. Future work should investigate how to distinguish sensor faults from genuine hydraulic changes, such as sediment accumulation or developing blockages.
- **Temporal modeling.** The current setup treats system states as independent snapshots. Incorporating temporal information through spatio-temporal GNNs or sequence-based extensions could improve the prediction of peak events and enhance flood wave tracking.
- **Model design.** While the standard GATRes architecture already shows strong performance, extending it toward a heterogeneous and temporally aware variant could better capture nonlinear dynamics around pumps, weirs, and outfalls.
- **Seed-to-seed variance in absolute accuracy (Shunqing).** Across $N = 3$ seeds, one run's clean-condition accuracy was much worse than the other two, most likely due to which sensor placement that seed happened to draw.
- **Comparative evaluation.** Although the focus here was on GNN surrogates, benchmarking against simpler methods (e.g., interpolation or regression) would help contextualize their added value in different operational scenarios.

The present evaluation also does not separately assess error during flooding and non-flooding conditions. Future work should evaluate flooded nodes separately and determine whether the surrogate correctly identifies operationally important flooding states. This does not change the qualitative conclusion: the denoising model's *relative* response to drift is consistently small (6–8% across all three seeds) while the baseline's is erratic (16–94%). But the absolute MAE values in Table 6 carry large variance and should not be read as tightly estimated point values. A masking-rate or sensor-placement sensitivity study, beyond the fixed 30% coverage used here, would clarify how much of this variance is placement-driven versus optimization noise.

- **Mechanism not directly verified.** We show that denoising training improves robustness, but have not confirmed *why*, for example via attention-weight inspection showing the network down-weighting corrupted sensors, which would strengthen the causal story beyond the aggregate error metrics reported here.
- **Single masking rate (Shunqing).** All drift results are reported at one sensor-coverage level (30%); a sweep across coverage levels, mirroring Figure 5's masking-only sweep, would clarify how the robustness benefit scales with sensor scarcity.
- **Chronological split is a positional proxy**, not a literal per-rainfall-event split, since the Shunqing public release does not retain individual event boundaries.

In future work, we aim to explore heterogeneous GATRes architectures with temporal extensions, integrate richer feature sets, and benchmark against non-graph baselines. We also aim to evaluate drift robustness using real continuous water-level monitoring data from Heusden and to investigate the attention-based mechanism underlying the denoising model's improved robustness. Ultimately, testing on live operational data will be an important step toward transitioning these surrogates from proof-of-concept to fully integrated components of wastewater digital twins.

9 ARTIFACT AVAILABILITY

Code implementing the pipeline described in this paper is available at: https://github.com/revinnaufal/gatres-denoising

ACKNOWLEDGMENTS

This research was funded by the Dutch Research Council NWO Merian Fund - Cooperation China (CAS) program, through the project DDTClean: Intelligent Wastewater Treatment - Distributed Digital Twin for Clean Water (482.22.007). The authors would like to thank Marco van Bijnen (sewage, urban water and data management consultant) for providing and explaining the InfoWorks model from the Heusden municipality, and Wijnand Turkensteen (urban water advisor, Waterschap AA en Maas) for insightful discussions.

REFERENCES

[1] M. Eren Akbiyik. 2023. Data Augmentation in Training CNNs: Injecting Noise to Images. arXiv:2307.06855 [cs.CV]. https://doi.org/10.48550/arXiv.2307.06855

[2] Pooja Aslami, Tara Aryal, Astha Rai, Niranjan Bhujel, Hossein Moradi Rekabdarkolaee, Kaiqun Fu, Reinaldo Tonkoski, Zongjie Wang, and Timothy M. Hansen. 2024. Power System Frequency Dynamics Modeling, State Estimation, and Control using Neural Ordinary Differential Equations (NODEs) and Soft Actor-Critic (SAC) Machine Learning Approaches. *ACM SIGAPP Applied Computing Review* 24, 1 (2024), 24–39. https://doi.org/10.1145/3663652.3663655

[3] Autodesk. 2024. InfoWorks ICM. https://www.autodesk.com/products/infoworks-icm/overview. Software. Originally developed by Innovyze, acquired by Autodesk in 2021.

[4] G. Baskar, A. N. Parameswaran, and R. Sathyanathan. 2025. Optimizing Papermaking Wastewater Treatment by Predicting Effluent Quality with Node-Level Capsule Graph Neural Networks. *Environmental Monitoring and Assessment* 197, 2 (2025). https://doi.org/10.1007/s10661-024-13581-3

[5] Yan Chen, Daoping Huang, Jing Wu, Gang Fang, Yan Li, Abid Aman, and Yiqi Liu. 2025. Towards Sludge Bulking Diagnosis via Dynamic Attention Graph Neural Network. *Journal of Water Process Engineering* 69 (2025), 106774. https://doi.org/10.1016/j.jwpe.2024.106774

[6] F. H. L. R. Clemens. 2001. *Hydrodynamic Models in Urban Drainage: Application and Calibration.* Ph.D. Dissertation. Delft University of Technology.

[7] Victoria Degeler, Mostafa Hadadian, Erkan Karabulut, Alexander Lazovik, Hester van het Loo, Andrés Tello, and Huy Truong. 2025. DiTEC: Digital Twin for Evolutionary Changes in Water Distribution Networks. In *Leveraging Applications of Formal Methods, Verification and Validation. Application Areas.* Springer Nature Switzerland, 62–82.

[8] Imane El Ghabi, Revin Naufal Alief, Dilek Düstegör, Guy Henckens, Alexander Lazovik, and Victoria Degeler. 2026. GNN-Based State Estimation for Urban Wastewater Digital Twins. In *Proceedings of the 41st ACM/SIGAPP Symposium on Applied Computing (SAC '26).* Thessaloniki, Greece. https://doi.org/10.1145/3748522.3779823

[9] A. Garzón, Z. Kapelan, J. Langeveld, and R. Taormina. 2022. Machine Learning-Based Surrogate Modeling for Urban Water Networks: Review and Future Research Directions. *Water Resources Research* 58, 5 (2022). https://doi.org/10.1029/2021wr031808

[10] Md. Nazmul Hasan, Sana Ullah Jan, and Insoo Koo. 2023. Wasserstein GAN-Based Digital Twin-Inspired Model for Early Drift Fault Detection in Wireless Sensor Networks. *IEEE Sensors Journal* 23, 12 (2023), 13327–13339. https://doi.org/10.1109/JSEN.2023.3272908

[11] MohammadHossein Homaei, Agustín Javier Di Bartolo, Mar Ávila, Óscar Mogollón-Gutiérrez, and Andrés Caro. 2024. Digital Transformation in the Water Distribution System Based on the Digital Twins Concept. arXiv:cs.CY/2412.06694 https://arxiv.org/abs/2412.06694

[12] Offir Inbar, Moni Shahar, and Dror Avisar. 2024. Predictive Modeling of BOD Throughout Wastewater Treatment: A Generalizable Machine Learning Approach for Improved Effluent Quality. *Environmental Science: Water Research & Technology* 10, 10 (2024), 2577–2588. https://doi.org/10.1039/d4ew00111g

[13] Zilin Li, Haixing Liu, Chi Zhang, and Guangtao Fu. 2024. Real-Time Water Quality Prediction in Water Distribution Networks Using Graph Neural Networks with Sparse Monitoring Data. *Water Research* 250 (2024), 121018. https://doi.org/10.1016/j.watres.2023.121018

[14] Shuaiyin Ma, Wei Ding, Yang Liu, Shan Ren, and Haidong Yang. 2022. Digital Twin and Big Data-Driven Sustainable Smart Manufacturing Based on Information Management Systems for Energy-Intensive Industries. *Applied Energy* 326 (2022), 119986. https://doi.org/10.1016/j.apenergy.2022.119986

[15] J. Meirlaen, B. Huyghebaert, F. Sforzi, L. Benedetti, and P. Vanrolleghem. 2001. Fast, Simultaneous Simulation of the Integrated Urban Wastewater System Using Mechanistic Surrogate Models. *Water Science and Technology* 43, 7 (2001), 301–309. https://doi.org/10.2166/wst.2001.0439

[16] Nikolaj T. Mücke, Prerna Pandey, Shashi Jain, Sander M. Bohté, and Cornelis W. Oosterlee. 2023. A Probabilistic Digital Twin for Leak Localization in Water Distribution Networks Using Generative Deep Learning. *Sensors* 23, 13 (2023), 6179. https://doi.org/10.3390/s23136179

[17] Duc-Thinh Ngo, Ons Aouedi, Kandaraj Piamrat, Thomas Hassan, and Philippe Raipin-Parvédy. 2023. Empowering Digital Twin for Future Networks with Graph Neural Networks: Overview, Enabling Technologies, Challenges, and Opportunities. *Future Internet* 15, 12 (2023), 377. https://doi.org/10.3390/fi15120377

[18] Prerna Pandey, Nikolaj T. Mücke, Shashi Jain, Pravin Ramachandran, Sander M. Bohté, and Cornelis W. Oosterlee. 2024. Machine Learning-Based Digital Twin for Water Distribution Network Anomaly Detection and Localization. In *Proceedings of the 3rd International Joint Conference on Water Distribution Systems Analysis & Computing and Control for the Water Industry (WDSA/CCWI 2024) (Engineering Proceedings)*, Vol. 69. Ferrara, Italy, 201. https://doi.org/10.3390/engproc2024069201

[19] RIONED. 2024. Ontwerpbuien met Statistische Herhalingstijd (Bui01–Bui10). https://www.riool.net/kennisbank/onderzoek/modelleren-hydraulisch-functioneren/stap-iv-hydraulische-belasting-bepalen/hemelwaterafvoer-hwa/ontwerpbuien-met-statistische-herhalingstijd/bui01-bui10

[20] Priit Roosipuu, Ivar Annus, Alar Kuusik, Nils Kändler, and Muhammad Mahtab Alam. 2023. Monitoring and Control of Smart Urban Drainage Systems Using NB-IoT Cellular Sensor Networks. *Water Science & Technology* 88, 2 (2023), 339–354. https://doi.org/10.2166/wst.2023.222

[21] L. A. Rossman. 2015. *Storm Water Management Model User's Manual, Version 5.1.* Technical Report EPA/600/R-14/413. U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC.

[22] Naghmeh Shafiee Roudbari, Shubham Rajeev Punekar, Zachary Patterson, Ursula Eicker, and Charalambos Poullis. 2024. From Data to Action in Flood Forecasting Leveraging Graph Neural Networks and Digital Twin Visualization. *Scientific Reports* 14, 1 (2024). https://doi.org/10.1038/s41598-024-68857-y

[23] Dhe Yeong Tchalla. 2025. ST-Hybrid: Dynamic Graph Learning with Multi-Scale Spatio-Temporal Attention for Traffic Forecasting. *ACM SIGAPP Applied Computing Review* 25, 4 (2025), 35–52. https://doi.org/10.1145/3787594.3787597

[24] Huy Truong, Andrés Tello, Alexander Lazovik, and Victoria Degeler. 2024. Graph Neural Networks for Pressure Estimation in Water Distribution Systems. *Water Resources Research* 60, 7 (2024). https://doi.org/10.1029/2023WR036741

[25] Guido Vaes, Patrick Willems, and Jean Berlamont. 2002. The Use of Reservoir Models for the Assessment of the Input from Combined Sewer Overflows into River Models. *Global Solutions for Urban Drainage* (2002), 1–16. https://doi.org/10.1061/40644(2002)139

[26] Charlotte Weil, Simon Elias Bibri, Régis Longchamp, François Golay, and Alexandre Alahi. 2023. Urban Digital Twin Challenges: A Systematic Review and Perspectives for Sustainable Smart Cities. *Sustainable Cities and Society* 99 (2023), 104862. https://doi.org/10.1016/j.scs.2023.104862

[27] Guang yao Zhao, Shunya Suzuki, Jia hao Deng, and Masafumi Fujita. 2022. Machine Learning Estimation of Biodegradable Organic Matter Concentrations in Municipal Wastewater. *Journal of Environmental Management* 323 (2022), 116191. https://doi.org/10.1016/j.jenvman.2022.116191

[28] Zhiyu Zhang, Wenchong Tian, Chenkaixiang Lu, Zhenliang Liao, and Zhiguo Yuan. 2024. Graph Neural Network-Based Surrogate Modelling for Real-Time Hydraulic Prediction of Urban Drainage Networks. *Water Research* 263 (2024), 122142. https://doi.org/10.1016/j.watres.2024.122142

[29] Jinjun Zhou, Yali Pang, Hao Wang, Wentao Li, Jiahong Liu, Zhuoran Luo, Weiwei Shao, and Haijia Zhang. 2023. Sewage Network Operational Risks Based on InfoWorks ICM with Nodal Flow Diurnal Patterns Under NPIs for COVID-19. *Water Research* 246 (2023), 120708. https://doi.org/10.1016/j.watres.2023.120708

[30] Qiande Zhu, Kaibin Fang, Dexun Zhu, Xinran Li, Xiaoyu Chen, Song Han, Feng Chen, Chuang Gao, Jun Sun, RongJie Tang, et al. 2024. Simulation and Analysis of Water Quality Improvement Measures for Plain River Networks Based on InfoWorks ICM Model: Case Study of Baoying County, China. *Water* 16, 18 (2024), 2698. https://doi.org/10.3390/w16182698

ABOUT THE AUTHORS:



Imane El Ghabi received the B.Sc. degree in Data Science and the M.Sc. degree in Data Engineering, both from Mohammed VI Polytechnic University, Morocco, in 2022 and 2024, respectively. She is currently a PhD Researcher at the Informatics Institute, University of Amsterdam, The Netherlands. Her research focuses on knowledge-sharing mechanisms for distributed digital twins, combining graph-based learning and foundation models to support scalable monitoring of urban drainage networks. She has prior experience at the World Bank, and the CERN ATLAS experiment, and has published work presented at the ACM/SIGAPP Symposium on Applied Computing (SAC 2026).



Revin Naufal Alief received the B.Eng. degree in Telecommunications Engineering from Telkom University, Indonesia, in 2020, and the M.Eng. degree in IT Convergence Engineering from Kumoh National Institute of Technology, South Korea, in 2023. He is currently a PhD student in the Distributed Systems group, Bernoulli Institute, University of Groningen, The Netherlands. His research focuses on blockchain, federated learning, and graph neural networks, particularly sensor placement and state estimation for digital twins. He has published in venues including the IEEE Internet of Things Journal, IEEE Access, and IEEE ICC.



Guy Henckens received the M.Sc. degree in Civil Engineering, specializing in Water Management, from Delft University of Technology, The Netherlands, in 2001. He is currently Leading Expert in Urban Water and Climate Adaptation at Aveco de Bondt, The Netherlands, with over 20 years of experience in urban water and sewerage engineering, including senior roles at Witteveen+Bos, Waterschap Brabantse Delta, and Royal HaskoningDHV. He chairs the sewerage expert network of Koninklijke NLingenieurs and serves on the program council of Stichting RIONED, and has taught professional courses in urban water and sewerage systems for over a decade.



Dr. Alexander Lazovik studied at Belarusian State University from 1997 to 2002 and received the Ph.D. degree in Computer Science from the University of Trento, Italy, in 2006. He is currently a Full Professor in the Distributed Systems group, Bernoulli Institute, University of Groningen, The Netherlands, having previously held ERCIM Fellowships at CWI and INRIA. His research focuses on distributed systems and service-oriented computing, particularly smart buildings, IoT, and, more recently, digital twins and graph neural networks for water infrastructure. He has published in venues including ACM Transactions on the Web, IEEE Transactions on Smart Grid, and Water Resources Research, with over 2,600 citations and an h-index of 26.



Dr. Dilek Düstegör received the B.S. and M.S. degrees in Computer Engineering from Boğaziçi University, Türkiye, in 1998 and 2000, respectively, and the Ph.D. degree in Industrial Computing from the University of Lille 1, France, in 2005. She is currently an Associate Professor in the Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands, having previously held faculty positions at Prince Mohammad Bin Fahd University and Imam Abdulrahman Bin Faisal University, Saudi Arabia. Her research focuses on distributed decision-making for smart environments, particularly digital twins for sustainable water and energy applications.



Dr. Victoria Degeler received the B.Sc. degree in Systems Analysis and Control from the National Technical University "Kharkiv Polytechnic Institute", Ukraine, in 2005, the M.Sc. degree in Intelligent Systems Design from Chalmers University of Technology, Sweden, in 2007, and the Ph.D. degree from the University of Groningen, The Netherlands, in 2014. She is currently an Associate Professor at the Informatics Institute, University of Amsterdam, where she coordinates the DiTEC project and the Big Data Engineering track of the joint UvA–VU M.Sc. in Computer Science. Her research focuses on reasoning and decision-making systems for smart environments, particularly digital twins for sustainable water and energy management.