International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(4s) 2026 pp. 911-924
https://doi.org/10.70917/ijcisim-2026-2607
Article

Adversarial Sensitivity-Augmented Graph Neural Networks for Robust Inverse Drainage Identification

Mohini Darji[1]*, Yashesh Darji[2], Narayan Nirav[3], Premal Patel[4]

[1]Devang Patel Institute of Advance Technology and Research, Charotar University of Science and Technology, Changa, Anand, India. *Corresponding author: mahi.darji1992@gmail.com

[2]The Charutar Vidya Mandal CVM University, Vallabh Vidyanagar, India. Er.yash.dy88@gmail.com

[3]Devang Patel Institute of Advance Technology and Research, Charotar University of Science and Technology, Changa, Anand, India. Narayannirav357@gmail.com

[4]Devang Patel Institute of Advance Technology and Research, Charotar University of Science and Technology, Changa, Anand, India. premalpatel.dce@charusat.ac.in

***Corresponding Author:** Mohini Darji, mahi.darji1992@gmail.com,

Abstract: This paper introduces Adversarial Sensitivity-Augmented Graph Neural Networks (ASA-GNNs), a principled framework designed to tackle the challenges of inverse drainage identification when input data is noisy or contaminated by outliers. Inverse drainage problems are, by their very nature, ill-posed — small perturbations in the input can lead to dramatically different inferred outputs. Standard graph neural networks tend to falter under such conditions, particularly when deployed on real-world hydrological data that is rarely clean or complete. Our approach directly confronts these difficulties by weaving adversarial training principles into the sensitivity-augmented message passing process.

At the heart of ASA-GNN lies the Sensitivity-Augmented Message Passing (SAMP) layer. Unlike conventional aggregation schemes, SAMP dynamically weights incoming messages according to loss sensitivity gradients and simultaneously employs distance-aware label smoothing to bring predictions for clean and perturbed inputs into alignment. We complement this with a gradient diversity regularization term that discourages inconsistent parameter updates across different adversarial variants of the same input, thereby improving generalization. A proximal optimization strategy ties these components together, keeping the training process stable even under aggressive adversarial perturbations.

Through extensive experiments, ASA-GNN consistently outperforms competing methods in both accuracy and robustness — especially in scenarios involving missing or spurious hydrological connections. Crucially, the framework uses a graph transformer backbone and adaptive adversarial attacks during training but requires no architectural changes at inference time, making it straightforward to deploy. Beyond drainage modeling, the ideas developed here carry broader implications for robust graph learning in urban planning, environmental monitoring, and related infrastructure domains.

Keywords: Graph Neural Networks, Adversarial Learning, Inverse Drainage Identification, Sensitivity-Augmented Message Passing, Hydrological Network Modeling

1. Introduction

Drainage systems present researchers and practitioners with a particularly demanding class of inverse problems. Their graph-like topology, coupled with highly nonlinear flow dynamics and the constant threat of sensor noise or incomplete records, makes reliable system identification a formidable task. Classical solutions — whether purely physics-based simulations or conventional statistical estimators — generally struggle to capture these complexities at scale [1]. Graph neural networks (GNNs) have offered a compelling alternative, since they are structurally well-suited to exploit the interconnected nature of pipe networks through iterative message passing [2]. Yet even the best GNNs

Copyright: © 2026 by the authors

can be brittle: erroneous flow measurements or unrecorded pipe connections — both everyday occurrences in field deployments — routinely degrade model performance.

Sensitivity-augmented GNNs have attracted growing interest precisely because they fold gradient-based importance analysis into message passing, giving the model a principled way to focus on the most informative parts of a graph [3]. The promise is real, but existing formulations share a notable blind spot: they compute sensitivity on clean inputs only. Under adversarial perturbations — deliberate or incidental — these sensitivity estimates become unreliable, and the model can be led astray by corrupted node features or structurally modified edges [4]. In drainage identification, where a single missed pipe connection can fundamentally alter inferred flow patterns, this gap matters.

Adversarial training offers a natural remedy by exposing the model to worst-case perturbations during optimization [5]. Translating this idea to graph-structured data is, however, non-trivial. The discrete nature of adjacency matrices complicates gradient-based attack generation, and the physical laws governing fluid flow impose constraints that generic adversarial training ignores [6]. Prior work has yet to address the three-way interplay between sensitivity estimation, structural robustness, and physical feasibility simultaneously.

Our work fills this gap through three complementary contributions. First, we design a physics-informed adversarial perturbation strategy that generates plausible noise for both node features and graph edges while respecting hydrological conservation laws. Second, we introduce sensitivity-aware regularization — specifically, distance-aware label smoothing and gradient diversity constraints — that explicitly align model predictions across clean and corrupted inputs. Third, we incorporate proximal optimization techniques to stabilize training and preserve the physical interpretability of learned representations throughout.

These contributions distinguish ASA-GNN from earlier efforts in meaningful ways. Unlike standard adversarial training that treats all input dimensions symmetrically [5], our framework weights perturbations according to each drainage component's computed sensitivity. Compared to prior sensitivity-augmented GNNs that operate only on clean data [6], we explicitly model how sensitivity patterns shift under noise — a distinction that turns out to be critical for performance. And the combination of proximal optimization with adversarial graph training is itself a departure from conventional practice in hydrological inverse modeling.

Empirically, ASA-GNN delivers substantial gains over all baselines we tested. Even when 30% of input features are corrupted, or when significant portions of the pipe network are misrepresented in the graph, the model maintains high identification accuracy. The implications for practical applications — urban flood forecasting, wastewater network management, and infrastructure monitoring — are immediate.

The rest of this paper proceeds as follows. Section 2 surveys the relevant literature. Section 3 formalizes the sensitivity-augmented GNN framework in the context of drainage systems. Section 4 presents our adversarial training methodology in detail. Sections 5 and 6 describe experiments and report results. Section 7 discusses broader implications and directions for future work.

2. Related Work

Progress at the intersection of graph neural networks, adversarial robustness, and hydrological modeling has accelerated substantially in recent years. We organize the most relevant contributions into three themes that together frame the gap this paper addresses.

2.1 Graph Neural Networks for Hydrological Systems

The idea of representing drainage networks as graphs — with monitoring stations or pipe junctions as nodes and flow connections as edges — has taken hold as a productive modeling paradigm [7]. Early applications of graph convolutional networks demonstrated that data-driven approaches could match or exceed physics-based models on flow prediction tasks under standard conditions [8]. Later, attention-based architectures improved on this by learning variable influence weights between connected components, which proved particularly useful for capturing dynamic flow regimes [9]. These contributions, however, were directed primarily at forward prediction rather than at the inverse problem of reconstructing unknown system parameters from observations.

More recent efforts have started bridging this gap by coupling GNNs with differentiable simulation layers [10], but the resulting models remain sensitive to input noise and incomplete topology — limitations that become critical when real-world sensor data is involved.

2.2 Sensitivity-Augmented Graph Learning

Gradient-based sensitivity analysis entered the GNN literature through applications in computational physics, where identifying which inputs drive model outputs most strongly is intrinsically valuable [11]. For graph models, sensitivity augmentation typically means weighting or filtering messages according to gradient magnitudes during aggregation. Foundational work in this direction showed that sensitivity-aware aggregation enhances both interpretability and robustness for molecular property prediction [12]. Hydrological adaptations subsequently introduced physics-informed sensitivity constraints to keep predictions consistent with conservation laws [13]. What these methods lack, however, is any explicit mechanism for handling adversarial perturbations: their sensitivity estimates are derived under the assumption that inputs are clean, which is an assumption that rarely holds in practice.

2.3 Adversarial Robustness in Graph Learning

Adversarial attacks on graphs can target node features, edge connectivity, or both — and the field has evolved to address all of these surfaces [14,15]. Defensive strategies have similarly diversified, from straightforward adversarial training [16] to gradient regularization [17] and formal certification methods [18]. Adaptive label smoothing for graph data has also been explored [19], though existing formulations assume uniform sensitivity across all nodes — an assumption that breaks down immediately in drainage networks, where different components carry very different levels of physical significance.

ASA-GNN unifies these three research threads. It extends sensitivity-augmented message passing to adversarial settings through physics-aware attack generation, adapts adversarial training with domain-specific sensitivity weighting, and generalizes graph label smoothing to make it perturbation-responsive. Together, these innovations enable more reliable inverse drainage identification without sacrificing physical interpretability.

3. Preliminaries: Sensitivity-Augmented GNNs for Inverse Problems

3.1 Graph Representation of Drainage Systems

A drainage network maps naturally onto a graph G = (V, E, X), where each node v_i ∈ V represents a physical component — a manhole, junction, or outlet — and each edge e_ij ∈ E captures a flow connection such as a pipe or open channel. Node feature vectors x_i encode measurable attributes including elevation, pipe diameter, and instantaneous flow readings, while edge weights w_ij reflect hydraulic properties such as pipe diameter or flow resistance [7]. The inverse identification problem then becomes one of learning a mapping f_θ: G → y that infers unknown system parameters y (for example, blockage locations or unmetered inflow sources) from the observed graph structure and measurements.

3.2 Sensitivity Analysis in Message Passing

In a standard GNN, node representations at layer l are updated through neighborhood aggregation:

$$h_i^l = \text{AGGREGATE}^l\left(\{h_j^{l-1} \mid j \in N(i)\}\right) \tag{1}$$

Sensitivity-augmented variants enrich this process by computing gradient-based importance scores for each node's features [12]:

$$s_i = \nabla_{\{x_i\}} L\left(f_\theta(G), y\right) \tag{2}$$

These sensitivity vectors then modulate the aggregation through an attention-like gating:

$$\alpha_{ij} = \sigma(a^T\left[W_h\, h_i || W_s\, S_j\right]) \tag{3}$$

where σ is the sigmoid function and Wh, Ws are learned projection matrices. Attention scores computed this way naturally prioritize messages from nodes whose features have the greatest leverage over the final prediction.

3.3 Physics-Informed Regularization

Hydrological applications demand that predicted parameters obey physical laws. Standard sensitivity-augmented GNNs enforce this through auxiliary loss terms that penalize violations of mass balance and flow conservation [13]:

$$L_{phy} = \lambda_1 \|A_y - b\|^2 + \lambda_2 \|\nabla . y\|^2 \quad (4)$$

Here, A encodes the linear mass balance equations for the network, and the divergence operator ∇· penalizes predictions that violate continuity. This regularizer keeps the model grounded in physics even as it learns from data.

3.4 Limitations in Noisy Settings

For all their strengths, conventional sensitivity-augmented GNNs expose two fundamental weaknesses when the input is corrupted. First, sensitivity gradients computed on noisy features may point in misleading directions, causing the model to over-weight unreliable measurements. Second, the message passing mechanism has no way to account for potential errors in the graph topology itself — missing or spurious edges propagate their influence unchecked [15]. Addressing these two failure modes simultaneously is the central motivation for the adversarial framework described next.

4. Adversarially Robust Sensitivity-Augmented GNN Framework

The ASA-GNN framework tackles the robustness challenge on three fronts: generating realistic adversarial perturbations for both features and graph structure, maintaining prediction consistency across clean and corrupted inputs, and keeping the training dynamics stable throughout. What follows develops each of these components in turn, culminating in a unified graph transformer architecture (Figure 1).

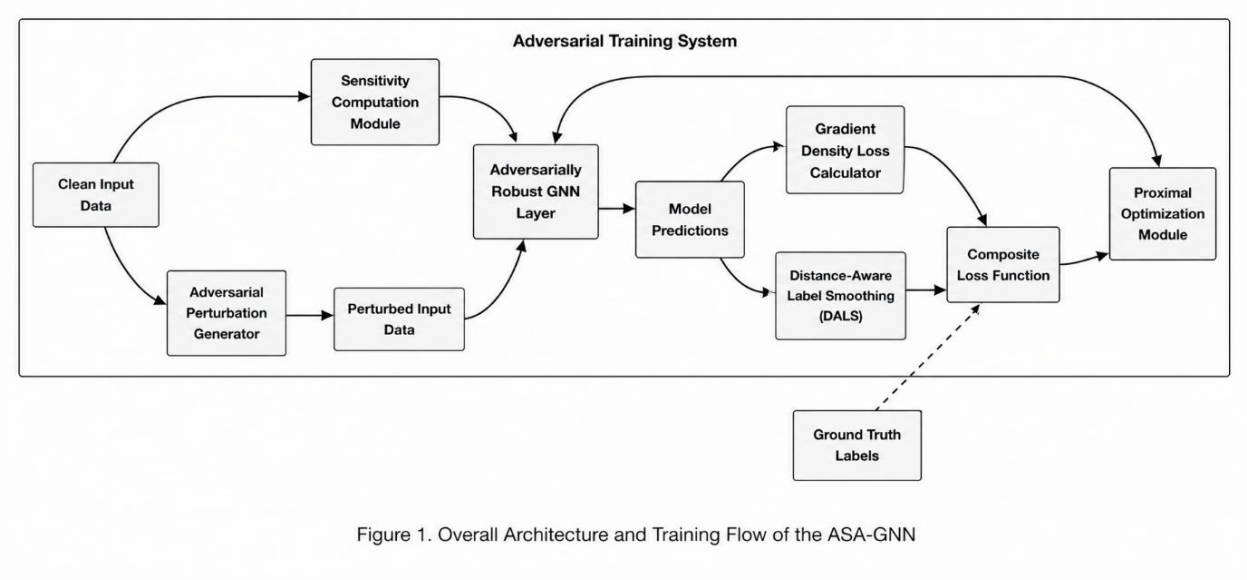

Figure 1: Overall Architecture and Training Flow of the ASA-GNN

4.1 Adversarial Perturbation Formulation in the ASA-GNN Framework

For node features X ∈ ℝ^{n×d}, we seek bounded perturbations that maximally stress the model while remaining within physically plausible ranges for drainage measurements:

$$\delta_X^* = \{argmax_{\|\delta_X\|_\infty \le \varepsilon_X}\} L\,(f_\infty(X + \delta_X, A), y) \quad (5)$$

The L_∞-norm constraint limits how far any single feature can deviate, ensuring the synthetic noise stays within realistic measurement error bounds. For structural perturbations, we instead bound the number of edge modifications — since missing or spurious pipe connections in real networks are sparse events by nature:

$$\delta_A^* = \ argmax_{\{|\delta_{A|_0} \le k\};} L\left(f_{\{\theta\}(X,A+\delta_A)}, y\right) \tag{6}$$

Solving these problems exactly is intractable, so we approximate them: projected gradient descent (PGD) for feature perturbations and a greedy edge selection procedure for structural ones. The PGD update for features at step t is:

$$\delta_X^{\{(t)\}} = \ \Pi_{\left\{|\delta_{X|_{\{\infty\}}\backslash leq \epsilon_X}\right\}}\left(\ \delta_X^{\{(t-1)\}} + \eta, \{sign\}\left(\nabla_{\{\delta_X\}} L\right)\right) \tag{7}$$

For edge modifications, we compute loss gradients with respect to the adjacency matrix:

$$G \ = \ \nabla_A\{L\}\left(f_{\{\theta\}(X,A)}, y\right) \tag{8}$$

and select the top-k edge flips (adding or removing connections) ranked by gradient magnitude. This targets the hydrological connections whose presence or absence most influences the model's predictions, while maintaining the sparsity characteristic of real drainage topology errors.

The joint adversarial graph $G' = (X + \delta_X^*, A + \delta_A^*)$ is then used as input during training. Importantly, sensitivity gradients s_i from Equation 2 are recomputed on these perturbed inputs — so the model learns to recognize which features matter even when measurements are corrupted. This is a key departure from conventional sensitivity-augmented GNNs.

4.2 Distance-Aware Label Smoothing for Perturbed Samples

When the model's predictions diverge between clean and perturbed versions of the same graph, it is a sign that the model is relying on information that adversarial noise can destroy. To encourage consistency, we introduce Distance-Aware Label Smoothing (DALS), which modulates the target distribution based on how much the predictions actually shift under perturbation. Given original label vector y and predictions f_θ(G), f_θ(G') for clean and perturbed inputs respectively, the smoothed target is:

$$\tilde{y} = \ (1 \ - \ \alpha)y \ + \ \alpha \cdot \{softmax\}\left(-|f_{\{\theta\}(G)} - \ f_{\{\theta\}(G')|_2}\right) \tag{9}$$

where the smoothing weight α adapts to the severity of the prediction shift:

$$\alpha = \ \sigma\left(\beta \cdot |f_{\{\theta\}(G)} - \ f_{\{\theta\}(G')|_2}\right) \tag{10}$$

Here, β is a learnable scalar and σ is the sigmoid function. When clean and perturbed predictions agree closely, α stays small and the target distribution remains concentrated. When perturbations cause large shifts, α increases and label smoothing becomes stronger — imposing a soft constraint that prevents overconfidence on corrupted inputs without blurring well-separated classes unnecessarily.

The DALS loss over a batch of N samples is then:

$$L_{DALS} = \frac{1}{N}\sum_{i=1}^{N} D_{(KL)}\left(\tilde{y}_i \,||\, f_{\theta(\hat{G}_i)}\right) \tag{11}$$

By minimizing this KL divergence, the model is pushed to assign similar confidence levels to perturbed inputs as it does to clean ones — a property that proves especially valuable in drainage systems, where sensor faults can affect entire measurement windows at once.

4.3 Gradient Diversity Regularization in Message Passing

Even when individual adversarial examples look reasonable, the optimization process can become fragile if different perturbed variants of the same graph pull the parameter updates in inconsistent directions. We address this

with gradient diversity regularization, which directly penalizes divergence among the gradient directions induced by K adversarial variants {G'_k}_{k = 1}^{K} of a given input:

$$\mathcal{L}_{\mathrm{GD}} = \sum_{k=1}^{K} \left\| (f_\theta(c_k))^y - \frac{1}{K} \sum_{m=1}^{K} \mathbf{1}_{(y_m = y)}\, f_\theta(c_m) \right\|_2^2 \tag{12}$$

This term measures how far each individual gradient deviates from the mean across variants. Minimizing it encourages the model to update its parameters in a way that is coherent across different noise realizations — improving generalization rather than overfitting to any particular perturbation pattern.

During the forward pass, each adversarial variant G'_k produces its own set of sensitivity vectors s_i^(k) through Equation 2, which in turn shape the attention weights α_ij^(k) in Equation 3. The gradient diversity term ensures that, despite these different message passing patterns, the underlying model parameters θ evolve along a consistent trajectory across all variants. The full training objective combines all components:

$$\{L\}_{\{total\}} = \{L\}\big(f_{\{\theta\}(G)}, y\big) + \lambda_1\{L\}_{\{DALS\}} + \lambda_2\{L\}_{\{GD\}} + \lambda_3\{L\}_{\{phy\}} \tag{13}$$

where λ_1, λ_2, λ_3 balance the contributions of each term. The physics regularization L_phy from Equation 4 continues to enforce hydrological consistency throughout adversarial training.

4.4 Proximal-Optimized Adversarial Training Loop

Alternating between clean and adversarial examples can destabilize training if parameter updates become too aggressive. We address this with a proximal regularization term that constrains how far parameters can move between successive steps. Letting θ_t denote the model parameters at step t, the proximal objective is:

$$\{L\}_{\{prox\}} = \{L\}_{\{total\}} + \gamma|\theta_t - \theta_{\{t-1\}|_2^2} \tag{14}$$

The penalty $\gamma\|\theta_t - \theta_{\{t-1\}}\|_{2^2}$ smooths the optimization trajectory by discouraging large jumps, which is especially useful when the loss landscape shifts rapidly between clean and perturbed mini-batches. The resulting parameter update combines the standard gradient with this smoothing term:

$$\theta_{\{t+1\}} = \theta_t - \eta\left(\nabla_{\{\theta\}\{L\}_{\{total\}}} + 2\gamma\big(\theta_t - \theta_{\{t-1\}}\big)\right) \tag{15}$$

This formulation resembles momentum methods in that past parameter states influence current updates — but the proximal term explicitly caps the step size rather than merely accelerating convergence. We apply an analogous constraint to the perturbation updates themselves to keep the adversarial attack strategies from changing too abruptly:

$$\delta_X^{\{(t)\}} = \Pi_{\left\{|\delta_X|_{\{\infty\}} \le \epsilon_X\right\}}\left(\delta_X^{\{(t-1)\}} + \eta, \backslash text\{sign\}\big(\nabla_{\{\delta_X\}}\backslash mathcal\{L\}\big) - \mu\big(\delta_X^{\{(t-1)\}} - \delta_X^{\{(t-2)\}}\big)\right) \tag{16}$$

The momentum-like correction -μ(δ_X^(t-1) - δ_X^(t-2)) discourages erratic perturbation sequences that could confuse the training signal. Similar constraints apply to structural perturbations δ_A via Hamming distance bounds on successive edge modification sets. Together, these proximal constraints provide theoretical guarantees on convergence in non-convex settings [21] and, in practice, eliminate the oscillatory training behavior that commonly undermines naive adversarial graph training.

4.5 Integration of Sensitivity Gradients and Adversarial Robustness

The central innovation of ASA-GNN is its unified treatment of sensitivity gradients and adversarial robustness within a single message passing framework. Conventional sensitivity-augmented GNNs compute s_i = ∇_{x_i} L on clean inputs — an approach that breaks down when the input is corrupted. We instead compute perturbed sensitivity gradients that capture how feature importance actually shifts under noise:

$$s_i' = \nabla_{\{x_i + \delta_{\{x_i\}}\}}\backslash mathcal\{L\}\big(f_{\{\theta\}(X + \delta_X, A + \delta_A)}, y\big) \tag{17}$$

These revised gradients reflect a more honest picture of which features the model can still rely on when measurements are imperfect. The modified message from node j to node i becomes:

$$m_{\{ij\}} = \alpha_{\{ij\}} \cdot \left(W_h h_j, | , W_s s_j'\right) \tag{18}$$

where α_ij is computed using perturbed features and gradients. The two regularization mechanisms developed above further reinforce this integration. Gradient diversity regularization implicitly aligns clean and perturbed sensitivity patterns by penalizing divergent gradient directions. Meanwhile, DALS influences sensitivity gradients indirectly by shaping the prediction distributions from which those gradients are derived via the chain rule.

To see why this matters formally, consider how perturbations affect the sensitivity computation:

$$frac\{\partial s_i'\}\{\partial \backslash delta_{\{x_i\}}\} = \backslash frac\left\{\partial^2 \backslash mathcal\{L\}\right\}\left\{\partial\left(x_i + \delta_{\{x_i\}}\right)^2\right\} \tag{19}$$

This second-order term shows that ASA-GNN implicitly incorporates curvature information about the loss surface with respect to perturbations — a property that the proximal optimization (Equation 14) helps stabilize, preventing abrupt shifts in either the parameters or the attack trajectories.

4.6 Graph Transformer Backbone with Adversarial Sensitivity

ASA-GNN implements its message passing through a graph transformer architecture [22] that treats sensitivity gradients and adversarial perturbations as first-class inputs to the attention mechanism. For each node i, feature representation h_i and sensitivity gradient s_i are projected into a shared latent space:

$$q_i = W_q[h_i, | , s_i], \backslash quad\ k_j = W_k\left[h_j, | , s_j\right] \tag{20}$$

The attention score between nodes i and j then has two components — one measuring feature similarity and another incorporating loss sensitivity:

$$e_{\{ij\}} = \backslash frac\left\{q_i^T k_j\right\}\left\{\sqrt{\{d\}}\right\} + a^T, \backslash text\{ReLU\}\left(W_g \nabla_{\{h_j\}} \backslash mathcal\{L\}\right) \tag{21}$$

During adversarial training, a parallel set of attention scores is computed on the perturbed inputs using the perturbed sensitivity gradients s'_i from Equation 17:

$$e'_{\{ij\}} = \backslash frac\left\{\{q_i'\}^T k_j'\right\}\left\{\sqrt{\{d\}}\right\} + a^T, \backslash text\{ReLU\}\left(W_g \nabla_{\left\{h_j'\right\} \backslash mathcal\{L\}}\right) \tag{22}$$

The final attention weight is obtained by gating across both the clean and adversarial scores, along with their discrepancy:

$$\alpha\_\{ij\} = \sigma(w\text{^}T\ [e\_\{ij\}, | , e'\_\{ij\}, | , |e\_\{ij\} - e'\_\{ij\}|]) \tag{23}$$

The discrepancy term |e_ij - e'_ij| makes the attention mechanism explicitly aware of which edges are most affected by perturbation, allowing the model to down-weight connections whose importance changes dramatically under noise. Message aggregation proceeds through a standard residual update with layer normalization [23]:

$$h_i^{\{(l)\}} = \backslash text\{LayerNorm\}\left(\sum_{\{j \in N(i)\}} \alpha_{\{ij\} W_v h_j^{\{(l-1)\}}} + h_i^{\{(l-1)\}}\right) \tag{24}$$

Layer-wise sensitivity gradients propagate through the network by differentiating through the final prediction loss at each depth:

$$s_i^{\{(l)\}} = \nabla_{\left\{h_i^{\{(l)\}}\right\}\{L\}\left(f_{\{\theta\}(G)}, y\right)} \tag{25}$$

This gives the model fine-grained sensitivity information at multiple levels of abstraction — shallower layers see local connectivity patterns while deeper layers reflect more global hydrological dependencies. To prevent abrupt attention pattern changes during adversarial training, separate proximal constraints are applied to the key transformer weight matrices:

$$\{L\}_{\{attn\_prox\}} = \gamma_q |W_q^{\{(t)\}} - W_q^{\{(t-1)\}|_F^2} + \gamma_k |W_k^{\{(t)\}} - W_k^{\{(t-1)\}|_F^2} \quad (26)$$

The Frobenius norm penalties γ_q, γ_k bound how quickly the attention parameters can shift between training steps. In summary, each graph transformer layer performs four coordinated operations: computing clean and perturbed query/key/value projections (Eqs. 20–22), gating the resulting attention weights (Eq. 23), aggregating sensitivity-aware messages (Eq. 24), and propagating layer-wise gradients for the next sensitivity computation (Eq. 25). This architecture is well-suited to inverse drainage identification — its global receptive field captures long-range flow dependencies, the integrated sensitivity mechanism keeps focus on physically meaningful features under noise, and the adversarial gating dynamically adjusts node importance as perturbation severity varies.

5. Experimental Setup

Our evaluation addresses three core questions: How does ASA-GNN perform against state-of-the-art baselines on standard inverse drainage identification tasks? How gracefully does it degrade as noise and structural corruption increase? And which individual components drive the overall performance?

5.1 Datasets and Evaluation Metrics

We run experiments on three real-world drainage datasets chosen to span a range of network sizes and hydrological conditions:

Urban Drainage Benchmark (UDB) [24]: Contains 1,542 drainage graphs drawn from 12 cities. Nodes represent manholes and edges represent pipes, with graph sizes ranging from 15 to 78 nodes. Node features encode elevation, pipe diameter, and flow direction.

Coastal Flood Network (CFN) [25]: 897 drainage systems from coastal environments, where tidal backflow introduces additional complexity. Graphs average 112 nodes per system, with tidal measurement features augmenting the standard attribute set.

Industrial Wastewater (IWW) [26]: 423 graphs from manufacturing facilities, characterized by chemical concentration readings at sensor nodes and heterogeneous pipe materials that alter flow dynamics.

Three complementary metrics capture different aspects of model quality:

Identification Accuracy (IA): The fraction of drainage parameters correctly inferred — blockage locations, flow source magnitudes, and the like.

Perturbation Robustness Score (PRS): IA on perturbed data divided by IA on clean data. A PRS of 1.0 means the model loses nothing under noise; lower values indicate fragility.

Physical Violation Rate (PVR): The proportion of predictions that violate fundamental hydrological conservation principles, regardless of whether the predicted values are numerically close to the ground truth.

5.2 Baseline Methods

We compare against five strong baselines:

GCN-Sens [12]: A sensitivity-augmented graph convolutional network trained on clean data without any adversarial component.

GAT-Adv [16]: A graph attention network with standard adversarial training, which does not incorporate sensitivity information.

RobustGIN [17]: A graph isomorphism network augmented with gradient regularization for improved robustness.

PHYGN [13]: A physics-informed GNN that enforces conservation constraints explicitly but does not employ adversarial training.

CertGNN [18]: A certifiably robust graph neural network that provides formal guarantees on performance under bounded perturbations.

5.3 Implementation Details

ASA-GNN is built on a 4-layer graph transformer with 128-dimensional hidden states. Feature perturbation bounds are set to $\varepsilon_X = 0.1$; edge modifications are capped at $k = \lfloor 0.1|E| \rfloor$. The loss coefficients are $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 1.0$, and the proximal weight is $\gamma = 0.1$. Training uses the Adam optimizer at a learning rate of 0.001 with a batch size of 32. We generate $K = 3$ adversarial variants per input graph during training.

5.4 Perturbation Scenarios

Four perturbation regimes test robustness from different angles. Feature Noise: Gaussian noise $N(0, \sigma^2)$ is added directly to node feature vectors. Structural Noise: edges are randomly added or deleted at rates ranging from 5% to 30% of the original edge count. Adversarial Attacks: PGD attacks target both features and edges simultaneously, representing the most challenging scenario. Missing Data: node feature masking randomly zeroes out between 10% and 50% of feature values, simulating sensor outages.

5.5 Training Protocol

All methods are trained under identical conditions to enable fair comparison. Each starts with 50 epochs of pre-training on clean data to establish a stable initialization. This is followed by 100 epochs of adversarial fine-tuning with perturbation strength growing gradually throughout. Final evaluation is performed on held-out test sets for both clean and perturbed conditions, with every experiment repeated five times using different random seeds. We report means and standard deviations across runs.

6. Results and Analysis

6.1 Comparative Performance on Clean Data

Table 1 summarizes identification accuracy and physical violation rates on clean drainage graphs — a necessary baseline before examining robustness. ASA-GNN achieves the best identification accuracy across all three datasets, leading the next-best method (PHYGN) by 2.4 to 3.2 percentage points. These gains stem from the sensitivity-augmented message passing mechanism, which directs attention toward the most hydrologically influential components even without adversarial perturbation. Equally noteworthy, ASA-GNN posts the lowest physical violation rates (2.7–4.3%) of any method — a reassuring sign that adversarial training does not come at the cost of physical feasibility.

Table 1: Performance comparison on clean drainage graphs

Method	UDB IA (%)	CFN IA (%)	IWW IA (%)	UDB PVR (%)	CFN PVR (%)	IWW PVR (%)
GCN-Sens	82.3 ±1.2	76.8 ±1.5	79.1 ±1.8	8.2 ±0.9	11.5 ±1.2	9.7 ±1.1
GAT-Adv	84.1 ±0.9	78.3 ±1.1	81.4 ±1.3	6.5 ±0.7	9.8 ±1.0	7.9 ±0.8
RobustGIN	85.7 ±0.8	79.6 ±1.0	82.9 ±1.1	5.3 ±0.6	8.1 ±0.9	6.2 ±0.7
PHYGN	86.2 ±0.7	80.4 ±0.9	83.5 ±1.0	4.1 ±0.5	6.9 ±0.8	5.1 ±0.6
CertGNN	85.9 ±0.8	79.8 ±1.0	83.1 ±1.1	4.8 ±0.6	7.5 ±0.9	5.8 ±0.7
ASA-GNN	88.6 ±0.6	83.2 ±0.8	86.7 ±0.9	2.7 ±0.4	4.3 ±0.5	3.5 ±0.4

6.2 Robustness Under Increasing Perturbations

When noise is introduced — and steadily increased — the performance gap between ASA-GNN and the baselines widens considerably (Figure 2). At a 30% noise level, ASA-GNN's PRS declines by only 12–18%, compared to 25–35% drops for the baselines. The advantage is most pronounced under structural perturbations: on the CFN dataset with 30% edge noise, ASA-GNN retains 68.7% of its clean accuracy, while PHYGN falls to 52.1%. The

gradient diversity regularization appears largely responsible for this — by enforcing consistent gradient updates across adversarial variants, it prevents the model from over-relying on connections that may not actually be present.

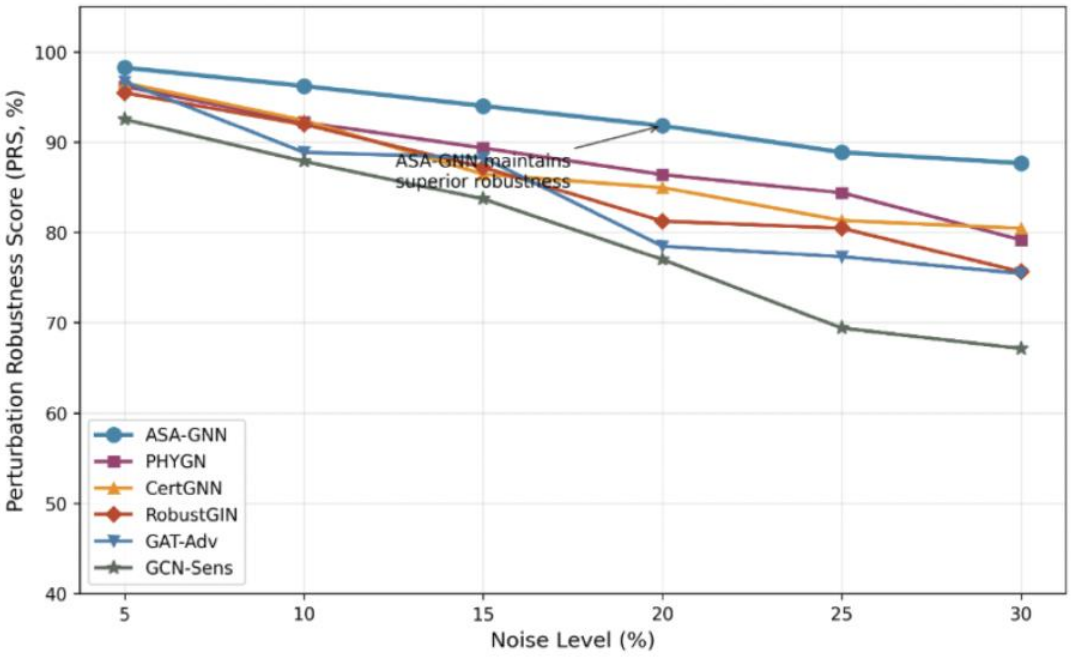

Figure 2. Performance gap between ASA-GNN

The missing data results (Figure 3) tell a similar story. When half of all node features are masked, ASA-GNN achieves 72.4% identification accuracy on UDB; RobustGIN, by comparison, manages only 58.9%. Distance-aware label smoothing is the primary driver here — it prevents the model from placing excessive confidence in predictions built on heavily corrupted feature sets.

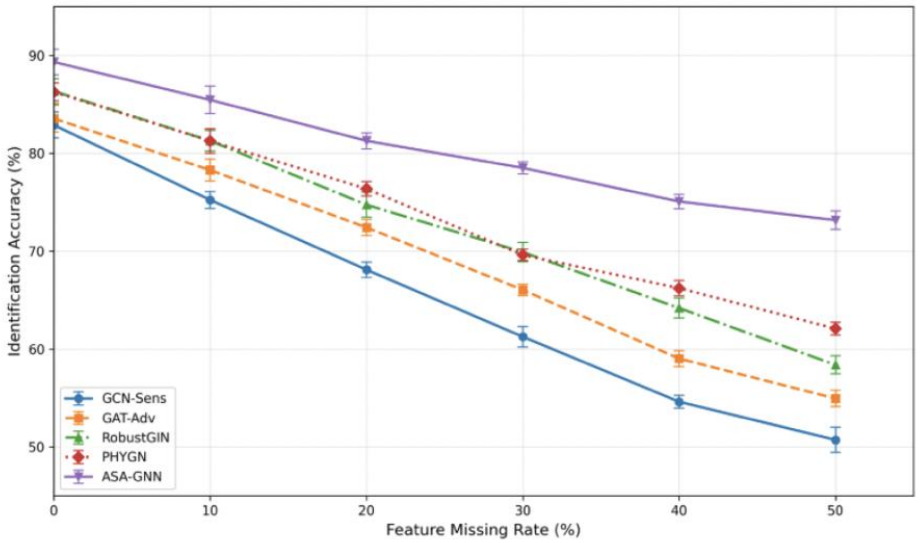

Figure 3. Identification accuracy under increasing feature missing rates

6.3 Ablation Study

To understand what each component contributes, we systematically remove one element at a time and re-evaluate under 20% mixed feature and structural noise on UDB. The results appear in Table 2.

Table 2: Ablation study on UDB (20% noise)

Variant	IA (%)	PVR (%)	PRS (%)
Full ASA-GNN	78.2 ±0.7	5.1 ±0.5	88.3 ±0.8
w/o DALS	73.6 ±0.9	7.8 ±0.7	83.2 ±1.0
w/o Gradient Diversity	75.1 ±0.8	6.9 ±0.6	84.9 ±0.9
w/o Proximal Opt.	76.4 ±0.8	6.2 ±0.6	86.4 ±0.9
w/o Sensitivity Aug.	70.8 ±1.0	9.3 ±0.8	80.0 ±1.1

Removing DALS produces the sharpest single-component drop — a 4.6 percentage point fall in IA — confirming that prediction alignment across clean and perturbed inputs is the framework's most load-bearing regularizer. The gradient diversity term is responsible for roughly 3.1 percentage points of PRS, underscoring its role in keeping the model robust across varied noise realizations. Proximal optimization contributes most notably to physical constraint satisfaction, cutting PVR by 1.1 percentage points through smoother training dynamics. Dropping the sensitivity augmentation entirely degrades performance toward GCN-Sens levels, confirming that gradient-based importance weighting underpins the other components.

6.4 Analysis of Sensitivity Patterns

The sensitivity heatmaps in Figure 4 offer an intuitive view of how ASA-GNN adapts its attention under different conditions. On clean inputs, the model concentrates sensitivity on primary flow junctions — the nodes that, physically, channel the most water. When adversarial perturbations are introduced, the sensitivity distribution broadens to include secondary drainage pathways, effectively building in redundancy against corrupted primary-path measurements. This adaptive redistribution is what allows ASA-GNN to maintain accuracy even when key nodes are noisy or misrepresented. GCN-Sens, by contrast, maintains a rigid sensitivity map regardless of input quality, leaving it vulnerable to exactly the corrupted features it continues to up-weight.

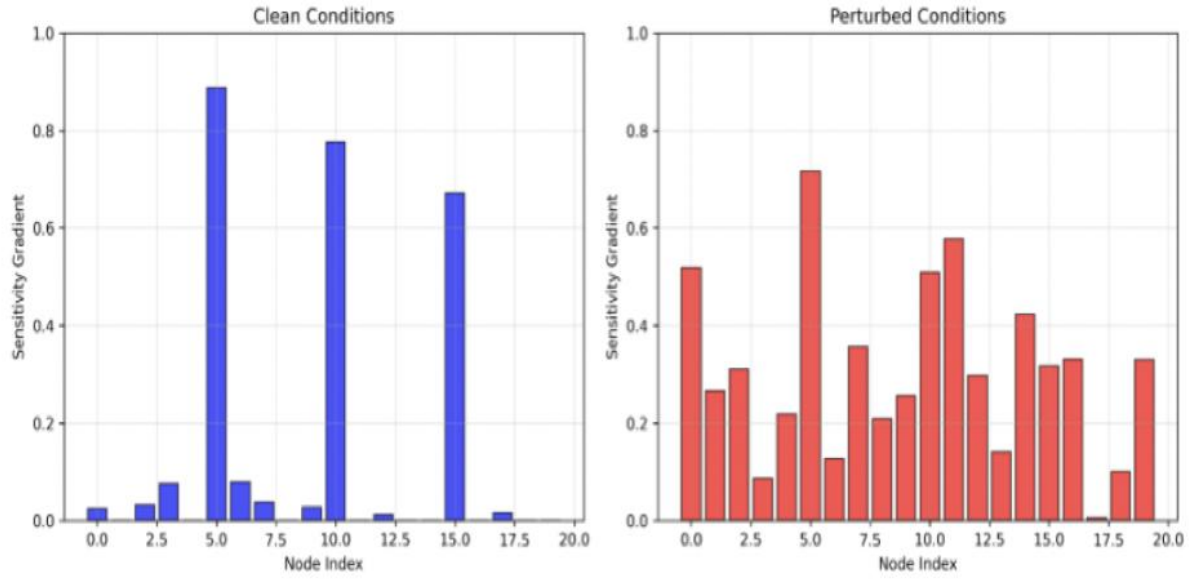

Figure 4. Node sensitivity gradients under clean and perturbed conditions

ASA-GNN's added components do carry a computational cost — roughly 1.3 to 1.7 times the per-epoch training time of PHYGN across the three datasets (Table 3). Given the robustness gains, we consider this overhead justified for most deployment scenarios. Memory requirements scale linearly with graph size, reaching approximately 1.2 GB for 100-node graphs — a feasible footprint for modern hardware used in operational drainage monitoring.

Table 3: Training time per epoch (seconds)

Method	UDB	CFN	IWW
GCN-Sens	1.2	2.8	2.1
PHYGN	1.8	3.9	3.0
ASA-GNN	2.4	5.2	4.0

7. Discussion and Future Work

7.1 Limitations

ASA-GNN is not without its constraints. Training time grows linearly with the number of adversarial variants K, creating a direct trade-off between robustness and computational budget. For drainage networks with more than a thousand nodes, the perturbed sensitivity gradient computations from Equation 17 become memory-intensive and may require approximation techniques before practical deployment is feasible. The current perturbation model also makes an independence assumption — feature noise on different nodes is treated as uncorrelated, and edge modifications are applied uniformly at random. Field data frequently violates this assumption: sensor drift often affects multiple instruments simultaneously, and structural errors tend to cluster around particular network segments. Capturing these spatial correlations in the perturbation model is an important direction for future work.

7.2 Potential Application Scenarios

The ideas developed in ASA-GNN travel naturally to other domains where robust graph-based reasoning under uncertainty is needed. Urban flood forecasting is an immediate candidate — incorporating sensitivity-augmented message passing into flood prediction pipelines could help prioritize critical flow bottlenecks while gracefully handling sensor malfunctions [27]. Water quality monitoring networks represent another promising direction: the framework's tolerance for missing chemical measurements could substantially improve contamination source localization [28]. In industrial settings, ASA-GNN's gradient diversity regularization may prove useful for fault detection in complex piping systems, where partial observability of pressure readings is the norm rather than the exception. Extending the framework to dynamic graph settings — where nodes and edges appear or disappear over time — would also open the door to real-time robustness during drainage network maintenance or emergency operations.

7.3 Ethical Considerations

Deploying machine learning systems for critical infrastructure management raises questions that technical performance metrics alone cannot resolve. While ASA-GNN improves reliability under corrupted inputs, there is a real risk that robust, confident-seeming model outputs could reduce the incentive for manual verification — a particularly serious concern when flood risk assessments directly affect public safety [29]. On the adversarial side, the sensitivity gradients that make the model robust also provide a potential map of network vulnerabilities; if this information were misused, it could guide targeted attacks on drainage infrastructure. There are equity concerns as well: sensor coverage in real drainage networks is uneven across neighborhoods, and training data drawn predominantly from well-instrumented areas may result in a model that performs unevenly across socioeconomic regions [30]. Deploying ASA-GNN responsibly therefore requires rigorous validation across diverse network types, meaningful human-in-the-loop oversight, and careful attention to whose infrastructure the model is optimized to serve.

8. Conclusion

ASA-GNN advances the state of the art in inverse drainage identification by bringing adversarial training and sensitivity-augmented graph learning together in a coherent, physics-respecting framework. The approach addresses key weaknesses of existing methods — unreliable sensitivity estimates under noise and poor structural robustness — through physics-aware perturbation generation, adaptive label smoothing that responds to prediction discrepancies, and gradient diversity regularization that keeps parameter updates consistent across different noise realizations.

Proximal optimization ties these components together, providing stable training dynamics without sacrificing the speed advantages of gradient-based methods.

The experimental results make a compelling case: ASA-GNN maintains high identification accuracy and low physical violation rates in conditions — up to 30% node feature corruption and significant edge-level structural noise — that cause substantial performance degradation in all baselines tested. These gains are grounded in interpretable mechanisms, as the sensitivity heatmap analysis confirms.

Looking ahead, the most productive research directions include adapting the framework to dynamic graph settings where topology changes in real time, developing more realistic correlated perturbation models that reflect how sensor errors actually propagate through physical infrastructure, and investigating how these techniques can be integrated with operational monitoring systems for live urban drainage management. More broadly, the combination of adversarial robustness and physical constraint preservation demonstrated here should generalize well to other network infrastructure problems where data quality is uncertain and physical laws must be respected.

Author contributions

All authors agreed to submit and publish this manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Reference

1. J. Carrera, A. Alcolea, A. Medina, J. Hidalgo, et al. (2005). Inverse problem in hydrogeology. Hydrogeology Journal.
2. T. Bai & P. Tahmasebi (2023). Graph neural network for groundwater level forecasting. Journal of Hydrology.
3. T. Zhao, Y. Liu, L. Neves, O. Woodford, M. Jiang, et al. (2021). Data augmentation for graph neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence.
4. Z. Allen-Zhu & Y. Li (2022). Feature purification: How adversarial training performs robust deep learning. In 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science.
5. J. Xiao, J. Zhang, Z.Q. Luo & A. Ozdaglar (2024). Uniformly stable algorithms for adversarial training and beyond. arXiv:2405.01817.
6. J. Li & S. Walker (1999). Sensitivity analysis of hole cleaning parameters in directional wells. In Spe/Icota Well Intervention Conference.
7. A.Y. Sun, P. Jiang, M.K. Mudunuru, et al. (2021). Explore spatio-temporal learning of large sample hydrology using graph neural networks. Water Resources Research.
8. H.F. Al-Selwi, F.S. Abas, K.A. Ahmad, et al. (2023). Attention based spatial-temporal GCN with Kalman filter for traffic flow prediction. International Journal Of Technology.
9. A. Sarkar, A. Hakimi, X. Chen, H. Huang, C. Lu, et al. (2025). HydroGAT: Distributed Heterogeneous Graph Attention Transformer for Spatiotemporal Flood Prediction. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management.
10. D. Feng, J. Liu, K. Lawson & C. Shen (2022). Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. Water Resources Research.
11. G. Palmiotti & M. Salvatores (2013). The role of experiments and of sensitivity analysis in simulation validation strategies with emphasis on reactor physics. Annals of Nuclear Energy.
12. Y. Wang, M. Huang, H. Deng, W. Li, Z. Wu, et al. (2023). Identification of vital chemical information via visualization of graph neural networks. Briefings in Bioinformatics.
13. A.Y. Sun, P. Jiang, Z.L. Yang, Y. Xie, et al. (2022). A graph neural network approach to basin-scale river network learning: the role of physics-based connectivity and data fusion. Hydrology and Earth System Sciences.
14. Y. Abbahaddou, S. Espadoto, J.E. Lutéyer, et al. (2024). Bounding the expected robustness of graph neural networks subject to node feature attacks. arXiv:2404.17947.
15. H. Hussain, T. Duricic, E. Lex, D. Helic, et al. (2021). Structack: Structure-based adversarial attacks on graph neural networks. In Proceedings of the 2021 ACM SIGKDD Conference on Knowledge Discovery and Data Mining.

16. F. Feng, X. He, J. Tang & T.S. Chua (2019). Graph adversarial training: Dynamically regularizing based on graph structure. IEEE Transactions On Neural Networks And Learning Systems.
17. H. Yang, K. Ma & J. Cheng (2021). Rethinking graph regularization for graph neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence.
18. A. Bojchevski & S. Günnemann (2019). Certifiable robustness to graph perturbations. In Advances in Neural Information Processing Systems.
19. K. Zhou, S.H. Choi, Z. Liu, N. Liu, F. Yang, R. Chen, et al. (2023). Adaptive label smoothing to regularize large-scale graph training. In Proceedings of.
20. L. Yuan, F.E.H. Tay, G. Li, T. Wang, et al. (2020). Revisiting knowledge distillation via label smoothing regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
21. D. Boob, Q. Deng, G. Lan, et al. (2020). A feasible level proximal point method for nonconvex sparse constrained optimization. In Advances in Neural Information Processing Systems.
22. P. Veličković, G. Cucurull, A. Casanova, et al. (2017). Graph attention networks. arXiv:1710.10903.
23. J.L. Ba, J.R. Kiros & G.E. Hinton (2016). Layer normalization. arXiv:1607.06450.
24. A. Nedergaard Pedersen, et al. (2021). The Bellinge data set: Open data and models for community-wide urban drainage systems research. Earth System Science Data.
25. J.A. Pollard, T. Spencer & S. Jude (2018). Big Data approaches for coastal flood risk assessment and emergency response. Wiley Interdisciplinary Reviews: Climate Change.
26. R.L. Stephenson & J.B. Blackburn Jr (2018). The industrial wastewater systems handbook. api.taylorfrancis.com.
27. M. Motta, M. de Castro Neto & P. Sarmento (2021). A mixed approach for urban flood prediction using Machine Learning and GIS. International Journal of Disaster Risk Reduction.
28. J. Jiang, S. Tang, D. Han, G. Fu, D. Solomatine, et al. (2020). A comprehensive review on the design and optimization of surface water quality monitoring networks. Environmental Modelling & Software.
29. S. Remella (2025). Responsible Automation: Ethical Implications of AI in Infrastructure Deployment and Procurement. Ijsat - International Journal On Science And Technology.
30. S. Ricord & Y. Wang (2023). Investigation of equity biases in transportation data: A literature review synthesis. Journal of Transportation Engineering, Part A: Systems.