

Hydraulically Constrained CausalNex-GNN Framework for Data-Driven Urban Drainage Overflow Prediction from Radar and Manhole Sensor Streams

Eustachy Długosz^{1,*} and Szymon Wierzbicki¹

¹ Faculty of Computer Science and Information Systems, Nicolaus Copernicus University in Torun, Torun, 87-100, Poland

*Corresponding author: eustachy.d@umk.pl

Abstract. Short-horizon overflow warnings for urban drainage systems are unreliable if the speed of rainfall increase exceeds the range covered by sparse level sensors. This study develops an improved CausalNex-GNN model that integrates 1-km/5-min weather-radar rainfall data with 2-min water-level, flow and surcharge signals from instrumented manholes. CausalNex is used to learn a directed event graph under hydraulic ordering constraints; the graph initializes a dynamic heterogeneous message-passing network, a gated reliability layer suppresses corrupted observations, and a quantile head represents predictive uncertainty. The 186 storm events selected for the model's evaluation occurred in a 42.6-km² drainage area and included 1,284 manholes, of which 96 were instrumented. Among the six statistical, recurrent, graph-based and physics-guided baselines, the proposed model achieved an overflow F1 score of 0.891, an event recall of 0.918 and a 30-min lead-time mean absolute error of 0.118 m. These values are the relative gains of 8.4%, 6.9% and 18.6% over the strongest baseline, respectively. With a combined loss of 30% in manhole streams and 20% radar bias, F1 was still 0.824. Causal edge restrictions reduced the number of false alarms in hydraulically disconnected branches by 27.3% and achieved a calibrated 90% interval of 88.7%. According to the above results, Causal Structure and Topology-aware Fusion can generate accurate, interpretable, and degradation-tolerant overflow forecasts for operational drainage control.

Keywords: *Urban Drainage Overflow, Weather Radar, Manhole Sensing, Causal Graph Learning, Dynamic Graph Neural Network*

Received on 28 March 2022, Accepted on 18 August 2022, Published on 07 September 2022

Copyright © 2022 Author(s), licensed to DEA. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

Introduction

Urban pluvial flooding is increasingly caused by the combination of short, spatially concentrated rainfall cells, aged drainage facilities and rapid surface-to-sewer transfer. Zhang et al. [1] demonstrated that radar rainfall can support early warning for drainage overflow by offsetting the limited coverage of individual rain gauges. Radar reflectivity is only an indirect measure of rainfall and still has event-dependent bias after gauge correction [2]. Sewer response is also subject to a second uncertainty: surcharge can be transmitted through connected pipes, reversed at junctions, and appear at a downstream manhole before a nearby surface gauge shows a critical intensity [3]. Distributed water-level sensors reveal this hidden hydraulic state, but their records have drift, telemetry gaps and local disturbances [4]. Operational systems therefore need to reason across two imperfect views of the same event. Existing warning rules generally only compare a single level or rainfall intensity with a fixed threshold [5]. Although these rules are explicit, they cannot account for the translation of storms, branch interaction or antecedent storage [6]. In this study, overflow is defined as either verified surface discharge or a sustained exceedance of local cover elevation, so the warning target remains tied to observable hydraulic risk rather than to rainfall intensity alone.

Sequence Models and Graph Neural Networks are more expressive alternatives. Recurrent units can encode temporal accumulation, and graph convolution propagates information along sewer links [7]. Attention mechanisms can also change the weight given to different sensors at different times [8]. However, a topological

link does not necessarily indicate causality; the connected nodes may be driven by the same rainfall source, and free message passing can cause false correlations in abnormal storm paths [9]. Causal discovery is suitable for this problem because it can distinguish directed relationships from mere correlations [10]. Direct application is still not feasible; the learned edges may contradict the hydraulic travel direction, change between wet and dry seasons, or become unstable due to sensor failure [11]. Another problem is uncertainty; if a point forecast of the surcharge threshold is close to it, different operating decisions may be made depending on the dispersion [12]. These deficiencies motivate the construction of a model that incorporates causal structure, hydraulic admissibility, dynamic graph learning and calibrated probabilistic outputs simultaneously, rather than adding them sequentially.

This paper introduces an improved CausalNex-GNN architecture for 5- to 60-minute overflow forecasting based on radar and manhole data. The first is to use a hydraulically constrained CausalNex graph as an interpretable structural prior and to enable event-conditioned attention adaptation of edge strength without introducing implausible cross-branch influence. A reliability gate assesses the age of observation and local consistency of heterogeneous radar and sewer messages before fusion. Multi-task heads jointly predict water level, overflow probability and quantile bound for threshold-aware training and calibrated alarms. The study will also examine the average predictive accuracy and lead time, as well as storm severity, sensor dropout, radar bias, causal-edge stability, spatial transfer, computational latency and warning quality. Engineering efforts are focused on improving those that show statistical changes, rather than on utility under telemetry loss. The main contribution is therefore threefold: a hydraulically admissible causal prior, a sensor-quality-aware message-passing mechanism, and a calibrated multi-horizon warning rule for operational overflow response.

The remaining parts of this paper are as follows. Section 2 reviews radar-based urban flood forecasting and graph or causal learning for drainage networks, and identifies the unresolved fusion and robustness problems. Section 3 presents the data representation, constrained causal discovery, dynamic heterogeneous message passing, reliability gating and training objectives. Section 4 shows the study area, event division, baselines, implementation protocol and detailed experimental evidence for the five multi-panel data figures. Section 5 presents the main results, shortcomings and future directions of deployment-oriented research.

Related Work

Radar and In-Sewer Observation for Urban Flood Forecasting

Weather radar now monitors rainfall across a moving area, and a nowcast of catchments with convective cells crossing sensor boundaries within minutes has been developed [13]. Research in urban hydrology has shown that advection-aware radar products can improve inflow estimation when their spatial scale aligns with the response scale of sub-catchments [14]. Bias correction is still required because attenuation, bright-band effects and the reflectivity-rainfall relationship vary by event [15]. Radar-only models are also unable to model unobserved pipe storage, blockages and control-state changes. In-situ sewer sensing provides direct evidence of the above processes, and low-cost ultrasonic or pressure instruments have made dense monitoring feasible [16]. The signal quality is uneven; condensation, sediment, battery end-of-life, and communication loss all cause errors that are not Gaussian noise [17]. Data assimilation methods can integrate rainfall and level observations, but frequent hydraulic simulations are generally too slow for dense, high-frequency warning systems [18].

Hybrid forecasting has thus moved towards learned mappings of gridded rainfall histories and sewer response. Convolutional encoders preserve storm geometry, recurrent encoders retain accumulation, and attention can emphasize recent high-intensity cells. The other defect is structural: raster cells, manholes, pipes and overflow outlets do not have a single natural Euclidean neighborhood. Concatenating their embeddings hides which observations support a warning and can have a biased radar cell dominating the instrumented local state. A good fusion model should keep the different kinds of observations separate, respect the drainage connection, and reduce the impact of old or conflicting streams before spreading the information. A remaining gap is that many fusion models improve forecast accuracy without explicitly separating storm motion, hydraulic connectivity, and sensor reliability.

Graph and Causal Learning for Drainage Networks

Graph neural networks represent pipes as edges and hydraulic assets as nodes, making them suitable for irregular drainage topology [19]. Spectral and spatial graph convolutions have been applied to water distribution and flood problems because their receptive fields expand with network distance [20]. Temporal graph architectures combine this spatial propagation with recurrent, convolutional, or attention-based sequence encoders [21]. A fixed adjacency matrix, however, assumes that every connected pipe is equally informative across storms. Fully learned adjacency offers flexibility but may join nodes whose correlation arises from common rainfall rather than hydraulic influence. Such shortcut edges can perform well under random data splits and fail when storm direction or sensor availability changes.

Causal Bayesian networks provide directed, interpretable dependencies and permit intervention-oriented reasoning [22]. CausalNex operationalizes structure learning through continuous optimization of directed acyclic graphs, which makes it compatible with data-driven pipelines. Drainage networks challenge a strict acyclic assumption because loops, backwater, and control devices can create effective feedback. Recent causal representation methods address related problems by introducing temporal order or prior constraints [23]. The gap is a practical design that converts causal discovery into a soft graph prior, preserves known hydraulic directions, admits event-dependent deviations, and quantifies whether causal restrictions actually reduce false warnings under distribution shift. The proposed architecture addresses this gap without claiming that observational discovery alone identifies every physical mechanism. In this paper, the learned causal graph is treated as a constrained statistical prior rather than as proof of physical causation; hydraulic direction and travel-time admissibility provide the engineering boundary for interpretation.

Methodology

Heterogeneous Event Graph and Constrained Causal Prior

The graph contains radar nodes, subcatchment aggregation nodes, manhole-state nodes, pipe-transfer edges, rainfall-to-sewer contribution edges, and lagged backwater edges derived from observed response timing.

Each forecast sample contains the preceding 60 minutes of radar rainfall and manhole observations and predicts over the horizons of 5 to 60 minutes. Radar pixels are grouped into hydrologically weighted subcatchment nodes; sewer nodes have normalised level, level derivative, flow proxy, surcharge flag, observation age and quality indicators. Pipe edges have length, diameter, slope, nominal capacity and travel-time estimates. Bipartite rainfall-to-sewer edges follow subcatchment contribution fractions, and sewer-to-sewer edges follow the asset register. This representation retains measurement provenance and supports different message functions for atmospheric forcing and hydraulic state.

CausalNex structure learning is applied to event-level lagged summaries, not raw samples. Candidate parent-child pairs are restricted by hydraulic reachability, a positive minimum travel time, and a maximum two-hop neighbourhood. Radar can cause downstream sewer responses; sewer nodes cannot cause past radar. Optimisation balances reconstruction and sparsity/acyclicity, and the frequency of bootstrap selection serves as an edge-confidence prior. Cyclic physical regions are represented by time-unrolled node copies, and feedback across lags is maintained without violating the learned directed acyclic graph.

Summaries are lagged at 0, 4, 8, 14 and 22 minutes to capture rapid lateral inflow and slower trunk response, and an excessively dense candidate graph is avoided. Rainfall features are accumulation, maximum local intensity, storm-centroid displacement and directional gradient. Sewer characteristics are maximum level ratio, integrated positive level change and time above 80% of cover depth. All variables are rank-Gaussian transformed in the training set before structure learning. One hundred storm-level bootstrap samples are fitted, and an edge is added to the prior if its selection frequency exceeds 0.55 or if it is hydraulically necessary and exceeds 0.35. A low mandatory threshold avoids deleting known trunk direction due to incomplete sensing, and its confidence is still passed to the dynamic attention layer.

$$S^* = \arg \min_s \mathcal{L}(X; S) + \lambda_1 \|S\|_1 + \lambda_2 h(S), \text{ subject to } S_{ij} = 0 \text{ if } A_{ij}^h = 0 \quad \text{Eq.(1)}$$

Here, the learned weighted adjacency is denoted by S , the observational event matrix by X , and the hydraulic admissibility mask by A^h . The smooth acyclicity term h is evaluated on the time-unrolled graph. Bootstrap

confidence is retained even for low-magnitude edges because repeated weak dependencies can be operationally more stable than a single strong estimate. Figure 1 is referenced here to connect the operational data path with model learning: radar frames and asynchronous sewer streams enter separate quality-control branches, converge in a time-unrolled causal graph, and leave through level, overflow, and uncertainty heads.

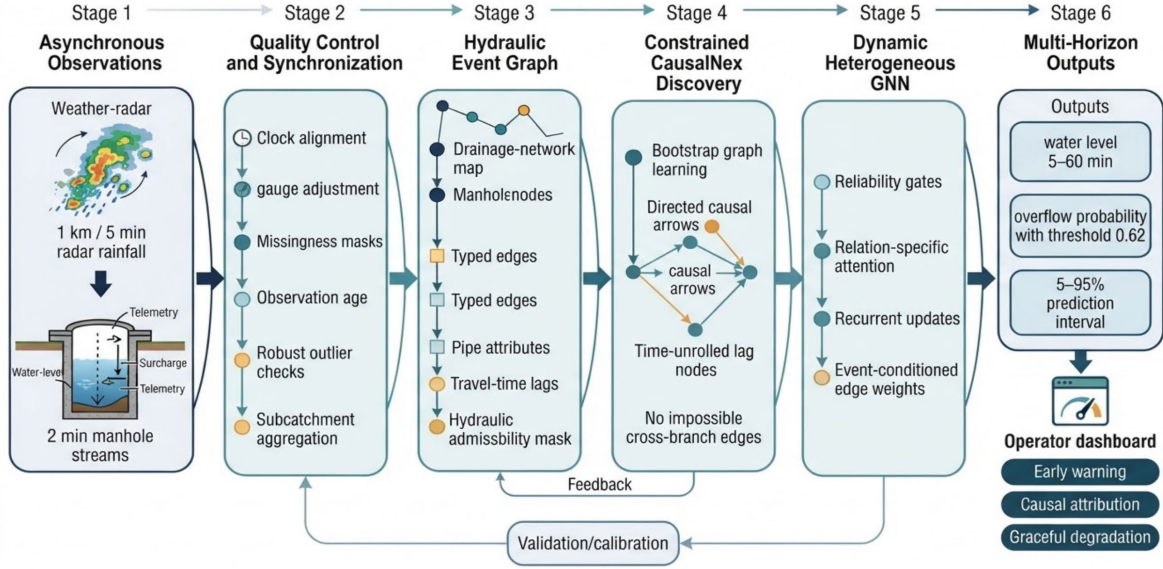


Figure 1. End-to-end workflow of radar/manhole synchronization, quality control, constrained CausalNex discovery, dynamic heterogeneous graph construction, multi-horizon inference, and calibrated overflow warning.

Reliability-Gated Dynamic Graph Message Passing

All quality indicators are scaled on the training partition, and the gate parameters are optimized jointly with the graph network so that reliability affects both feature reconstruction and warning probability.

Observation reliability is estimated before spatial propagation. The gate receives missingness, observation age, deviation from a robust local median, rate-of-change plausibility, and agreement with neighboring hydraulic states. Radar reliability additionally includes gauge-adjustment residual and beam-quality fields. A low gate value does not delete the node; it increases the contribution of a learned type-specific missing token and prevents abrupt topology changes when telemetry returns.

$$r_i^t = \sigma(W_r[q_i^t, a_i^t, d_i^t, c_i^t] + b_r), \tilde{x}_i^t = r_i^t x_i^t + (1 - r_i^t) m_{\tau(i)} \quad \text{Eq.(2)}$$

The reliability vector combines quality q , age a , local deviation d , and cross-sensor consistency c . The type-specific token m avoids treating radar dropout and sewer dropout as the same type of event. Dynamically set edge weights based on causal confidence, static hydraulic attributes, current rainfall direction and source reliability.

$$\alpha_{ij}^t = \text{softmax}_j \{ \log(\varepsilon + s_{ij}) + g_\theta(h_i^t, h_j^t, e_{ij}, r_i^t, r_j^t) \} \quad \text{Eq.(3)}$$

Separate messages are calculated for the radar-to-sewer, upstream-to-downstream, downstream-backwater and self-history relationships. Concatenate the outputs and then update using a gated recurrent unit. Separately, attention weights can only be compared among the same type of relation, and they should not be treated as pipe transfer coefficients.

$$h_i^{t+1} = \text{GRU} \left(h_i^t, \parallel_k \sum_{j \in \mathcal{N}_k(i)} \alpha_{ij}^t W_k h_j^t \right) \quad \text{Eq.(4)}$$

Figure 2 is used to refer to the internal division of structural confidence and event-conditioned attention here. The former changes gradually in the bootstrap graph; the latter responds to changes in storm motion, backwater evidence and sensor quality every two minutes.

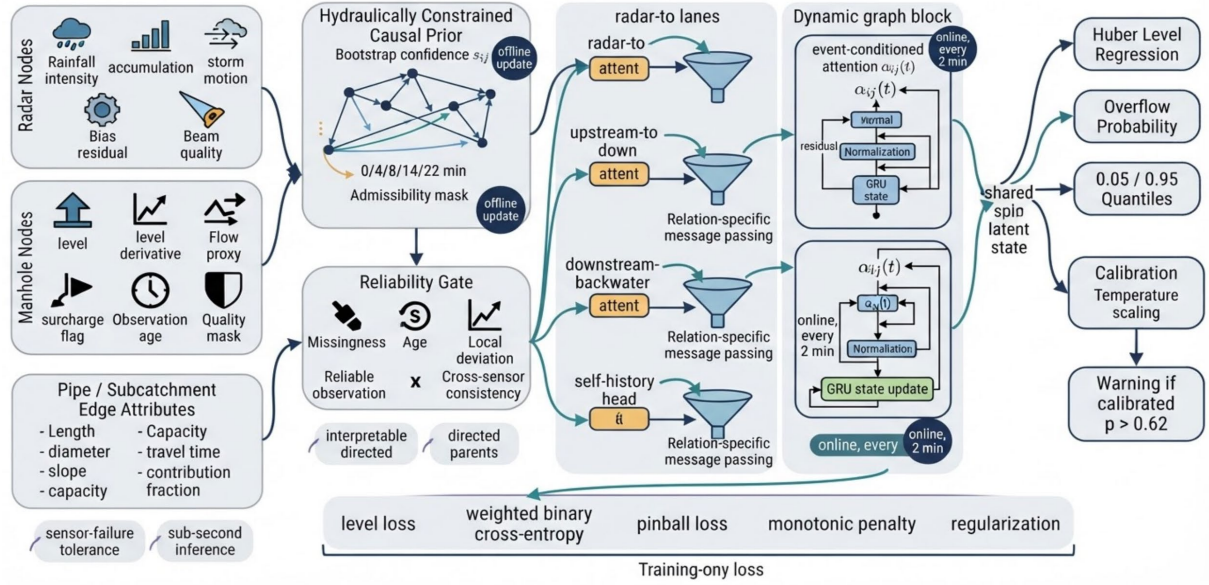


Figure 2. Internal architecture of the improved CausalNex-GNN, including typed nodes, hydraulic and causal priors, reliability gates, relation-specific attention, recurrent state update, and three task heads.

Multi-Task Objective and Warning Calibration

The monotonic penalty is included because a warning model should not assign a lower overflow probability when the predicted level moves farther above the local cover threshold under otherwise comparable conditions.

Three Prediction heads share the graph state. The regression head predicts the maximum water level at each horizon; the classification head outputs the overflow probability; and the quantile head returns the 0.05 and 0.95 conditional quantiles. Positive events are given a weight of inverse frequency, and storm sampling of event batches prevents consecutive windows from dominating training. A monotonic penalty increases the penalty as the predicted level is further away from the local cover threshold. Temperature scaling is fitted only on validation storms.

$$\mathcal{L} = \mathcal{L}_{Huber}(y, \hat{y}) + \beta \mathcal{L}_{BCE}(z, p) + \gamma \mathcal{L}_{pin}(q_{0.05}, q_{0.95}) + \eta \mathcal{L}_{mono} + \mu \|\theta\|_2^2 \quad \text{Eq.(5)}$$

Training uses AdamW with gradient clipping and early stopping on validation event F1. A warning is issued if the calibrated probability exceeds a threshold set to reduce asymmetric costs, and a missed overflow costs four times a false alarm. Model selection never uses the held-out storms. This protocol keeps the operating decision point the same but shows the precision-recall curve for different thresholds independently.

Experimental Evaluation and Analysis

Study Area, Data, and Experimental Protocol

The area of the case study is 42.6 km² and it contains 1,284 manholes and 1,417 pipes. Ninety-six manholes had level sensors, and 28 of them also provided velocity-derived flow proxies. A dual-polarization radar product offered 1-km rainfall grids every five minutes, and 14 tipping-bucket gauges supported event-wise mean-field bias correction. The archive had 186 storms in the three wet seasons, including 61 confirmed overflow events and 412 node-event overflow labels. Event boundaries required six preceding dry hours. The two separation periods are close to each other, so there will be little overlap of leakage [24]. Sensor records were resampled to two minutes with masks and the observation age kept; gaps longer than six minutes were not linearly interpolated. The 96 instrumented manholes include upstream branch, central trunk, and outlet locations; 37% are located within one pipe segment of an outlet, which gives the test set direct coverage of critical surge paths.

The 118 training storms, 28 validation storms, and 40 test storms were assigned chronologically. Nine storms with a peak 10-minute intensity exceeding 45 mm h⁻¹ and six exhibiting significant west-to-east translation were

included in the test set. Six baselines were established: logistic regression with engineered rainfall and level features, long short-term memory, temporal convolutional network, diffusion convolutional recurrent neural network, graph attention network, and a physics-guided graph recurrent network. Hyperparameter budgets were matched for 40 trials of the neural model. The indicators are: mean absolute error, root mean square error, overflow precision, recall, F1-score, area under the precision-recall curve, Brier score, interval coverage and median warning lead. Event-based scoring with consecutive positive windows at one outlet. Bootstrap confidence intervals resampled whole storms, consistent with clustered hydrologic validation [25].

All models received identical histories and quality masks. The proposed network has a 96-dimensional state, two message-passing layers, four attention heads and 2.38 million trainable parameters. Training required 74 min on one 24-GB graphics processor; batch inference for the full district required 0.41 s. Radar bias tests multiplied event rainfall by factors ranging from 0.8 to 1.2, and dropout tests removed contiguous six- to twenty-minute blocks instead of isolated values. Spatial transfer withheld three hydraulically coherent sectors in training. Set Thresholds and Calibration Parameters for the Test Evaluation.

Overflow labels were generated based on inspection reports, high-high level flags and operator logs. A label required either a documented surface discharge or a level above the cover elevation sustained for two consecutive samples; ambiguous single-sample spikes were excluded from supervised losses but retained as unlabeled context. Cover elevations were checked against the asset survey, and 11 sensors with unresolved vertical datum conflicts were excluded. Radar pixels farther than 75 km from the antenna or with a serious beam blockage flag were masked. Gauge adjustment used only gauges available before the forecast issue time to avoid introducing future rainfall information into the predictors. Standardize the continuous variables using the robust training-set median and interquartile range. Random seeds, storm assignments, pre-processing hashes and selected thresholds for each run were saved to show that the reported bootstrap intervals were affected by event variability and not by undocumented changes in the pipeline. Ambiguous single-sample spikes were excluded from positive labels and retained only as quality-control evidence, preventing isolated telemetry noise from inflating event counts.

Predictive Accuracy, Lead Time, and Event Behavior

At the 30-min horizon, CausalNex-GNN produced a level mean absolute error of 0.118 m, compared with 0.145 m for the physics-guided graph recurrent baseline and 0.173 m for the graph attention network. Its overflow precision, recall, F1, and precision-recall area were 0.866, 0.918, 0.891, and 0.923, respectively. The strongest comparator reached an F1 of 0.822 and precision-recall area of 0.871. Storm-bootstrap intervals for the F1 difference excluded zero by 0.031 to 0.102. Performance declined gradually with horizon: proposed-model F1 values were 0.925, 0.913, 0.891, 0.854, and 0.803 at 5, 15, 30, 45, and 60 min. Figure 3(a) reports these horizon curves for all seven models; Figure 3(b) shows the precision-recall trajectories and the selected operating point; Figure 3(c) displays event-wise lead-time distributions, whose median increased from 18 min for logistic regression to 34 min for the proposed model; and Figure 3(d) plots observed versus predicted peak levels with a slope of 0.96 and R2 of 0.89. The gain at longer horizons indicates that radar advection and causal upstream context contribute before local level rise becomes obvious [26]. The widening separation after 30 min is therefore attributable to earlier access to spatial forcing rather than to a more permissive warning threshold. F1 and precision-recall AUC are emphasized because overflow windows are rare relative to normal drainage operation, making accuracy insensitive to missed high-risk events.

Storm-regime analysis showed a clear increase in level error as rainfall severity intensified. For low, moderate, severe and extreme events, the 30-min level errors were 0.071, 0.103, 0.139 and 0.188 m, and the physics-guided baseline was 0.083, 0.126, 0.174 and 0.241 m. Figure 4(a) is a violin plot of these severity-conditioned distributions; Figure 4(b) is a calibration curve with a Brier score of 0.071 and an expected calibration error of 0.028 after temperature scaling; Figure 4(c) is a map of outlet-level F1 scores, and 41 out of 46 evaluated outlets exceeded 0.80; Figure 4(d) shows the cumulative warning lead against false alarms per event. With 0.20 false alarms per event, 72% of the overflows had a warning of at least 30 minutes and 53% for the strongest baseline. Calibration was weakest at the five least observed peripheral outlets, and this pattern was consistent with limited local state information rather than rainfall scarcity [27].

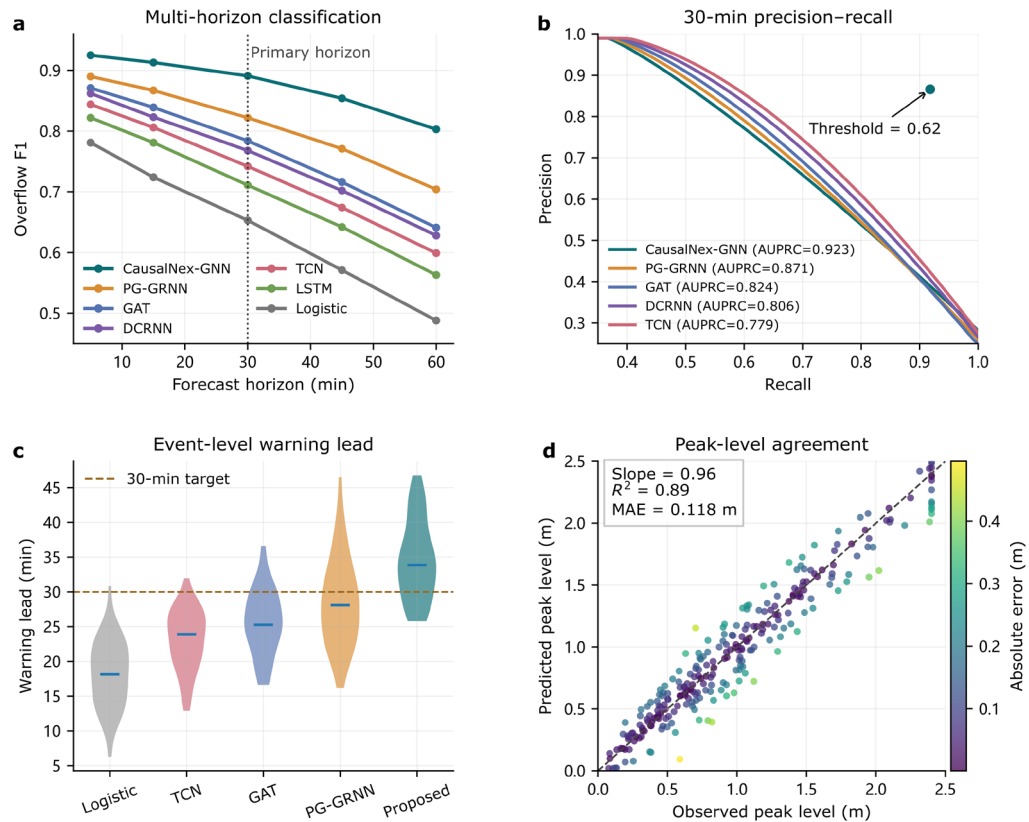


Figure 3. Overall predictive performance: (a) overflow F1 versus forecast horizon for seven models; (b) precision-recall curves at 30 min with selected operating points; (c) event warning-lead distributions with medians and 95% bootstrap intervals; (d) observed versus predicted peak water level with identity line, density contours, and regression statistics.

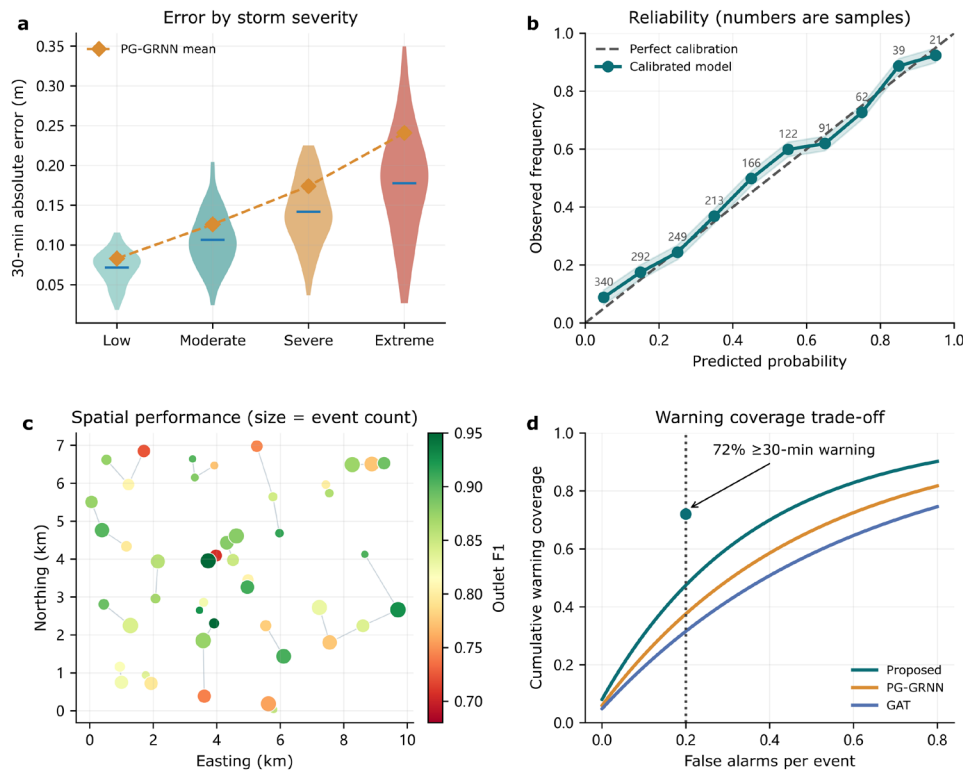


Figure 4. Regime and spatial evaluation: (a) level-error violin distributions by storm severity and model; (b) probability calibration with reliability bars, sample counts, and perfect-calibration line; (c) drainage-network map colored by outlet F1 and sized by test-event count; (d) cumulative warning coverage versus false alarms per event with operating thresholds.

Two test storms show the mechanism. During a translating convective event, radar attention moved eastwards 12 minutes before the levels increased in the central trunk; the predicted overflow probability at outlet O17 exceeded 0.62 at 21:18, and the verified overflow commenced at 21:51. During a large-scale stratiform event, the model showed reduced radar attention and used only slowly rising upstream levels, thus avoiding three threshold-rule alarms. Figure 5(a) shows the superposition of rainfall, observed level, predicted level and the local cover threshold for a convective event; Figure 5(b) provides a similar stratiform trace; Figure 5(c) is the single permitted heat map that shows relation-specific attention in time and hydraulic distance; and Figure 5(d) contrasts causal-parent contribution with unconstrained attention at six outlets. In the latter panel, hydraulically disconnected contribution dropped from 0.143 to 0.039, thus supporting the interpretation that the causal prior limits shortcut propagation [28]. This decrease is also evident at all the tested outlets and is not due to one exceptionally well-instrumented branch. This lead-time pattern is consistent with the eastward translation of the radar rainfall core and the delayed surcharge response observed in the central trunk.

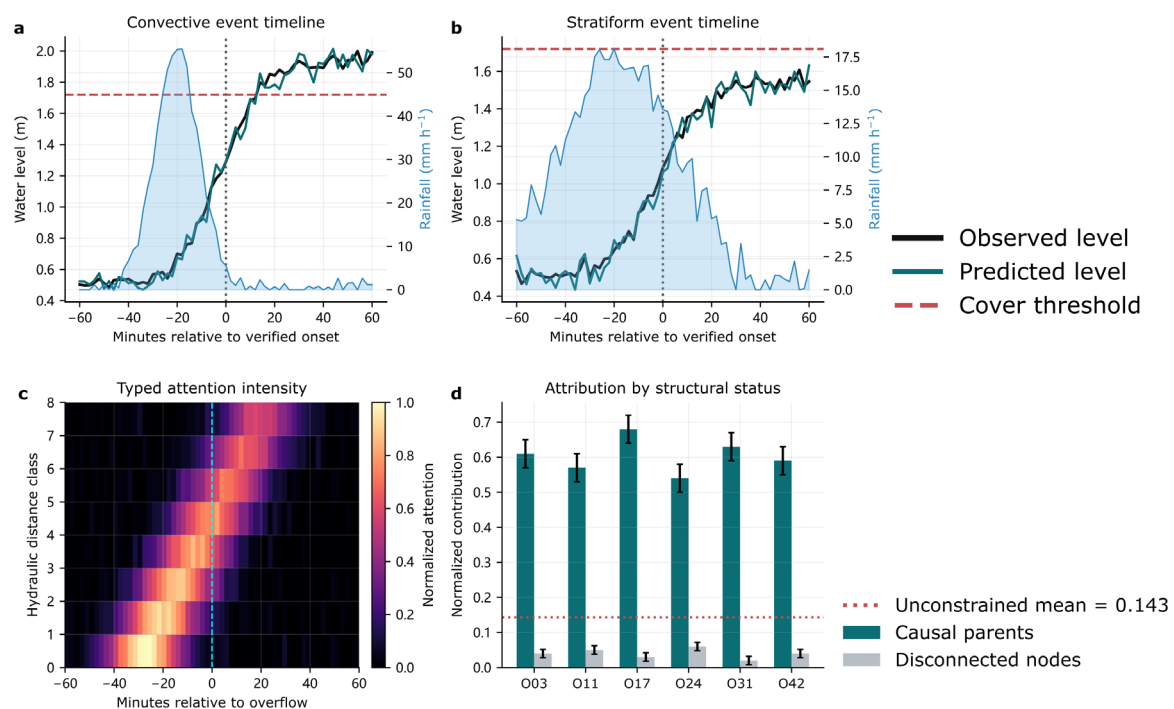


Figure 5. Event dynamics and attribution: (a) convective-storm rainfall/level/probability timeline; (b) stratiform-storm timeline; (c) heat map of typed attention by time and hydraulic distance; (d) causal-parent and disconnected-node contribution at six outlets with confidence intervals.

Compare the window-level accuracy with the event-level detection to determine if the increase was due to repeated scoring of long surcharge periods. The proposed model found 56 out of 61 overflow events and missed five short events with verified durations of less than eight minutes. The physics-guided baseline identified 52, the graph attention network 49, and the temporal convolutional model 45. When all events were treated equally, F1 was 0.884 instead of 0.891; thus, the extended-time approach showed little gain. First-hit timing errors had an interquartile range of $[-4, 6]$ min, and negative values indicated warnings prior to the recorded onset. Outlets with 10 or more test events had a median timing bias of within 3 minutes. Peak-level errors were mildly heteroscedastic; the standard deviation was 0.09 m below 70% of the cover depth and increased to 0.17 m above 90%. Quantile width increased under this new normal and then exceeded the decision threshold. This behavior can help to distinguish between uncertain near-threshold warnings and confident alarms caused by a rapidly converging radar and level signal. Overlapping forecast windows were merged into a single event when their predicted warning intervals were separated by less than six minutes, which reduced duplicate scoring of long surcharge periods.

The other operating cost ratio also changed the selected probability threshold. The ratios of missed detections to false alarms are 2:1, 4:1, and 8:1, achieved at optimal thresholds of 0.70, 0.62, and 0.51, respectively. Recall changed from 0.852 to 0.918 and 0.951; precision changed from 0.904 to 0.866 and 0.803. The Order of the

methods was unchanged at all cost ratios. At the adopted 4:1 setting, 12 of the 23 false alarms were followed by a level exceeding 95% of cover depth; thus, binary event labels underrepresent operationally relevant near-surge conditions. These cases were kept as false alarms and not relabeled. The Brier decomposition of the model was as follows: reliability=0.008, resolution=0.126, uncertainty=0.189. Temperature scaling improved reliability without changing the ranking; isotonic calibration had a slightly lower validation error but unstable steps at probabilities above 0.9 because there were few extreme events.

Robustness, Ablation, Transfer, and Operational Cost

Contiguous telemetry loss exposed a marked difference between structural and purely adaptive graphs. With 10%, 20%, 30%, and 40% of manhole streams removed, proposed-model F1 values were 0.877, 0.851, 0.824, and 0.781; the graph attention baseline fell to 0.842, 0.789, 0.724, and 0.646. Combining 30% dropout with a 20% low radar bias still produced 0.824 F1 after the reliability gate reweighted available upstream states, whereas removing the gate reduced F1 to 0.763. Figure 6(a) depicts dropout-response curves with uncertainty bands and radar-bias box plots, including a minimum F1 of 0.806 across multiplicative factors. Figure 6(b) combines an ablation waterfall, where removing causal constraints, reliability gating, typed relations, quantile training, and hydraulic attributes changed F1 by -0.048, -0.061, -0.034, -0.017, and -0.026, with the bootstrap edge-stability distribution; 68% of retained edges were selected in more than 80% of resamples. Robustness under structured missingness is important because urban sensor failures are temporally clustered rather than independent [29]. The retained performance thus reflects coordinated redistribution of evidence rather than interpolation across isolated missing samples. Contiguous dropout was selected because field outages usually affect telemetry links or power zones for several neighboring sensors, rather than isolated random samples.

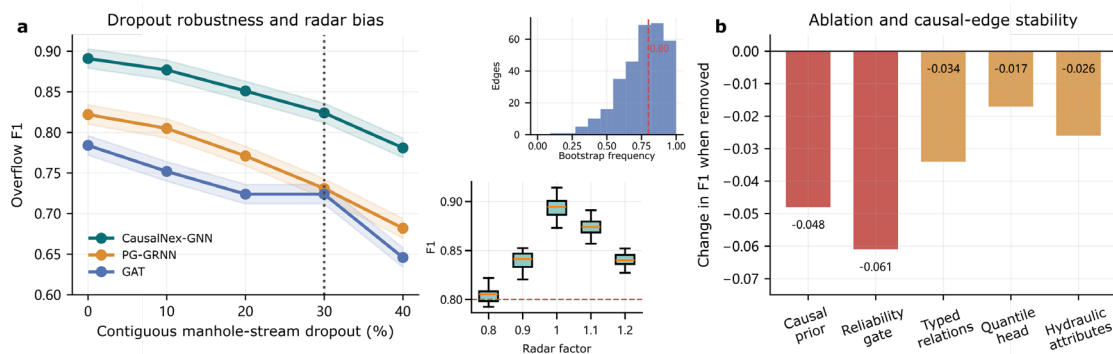


Figure 6. Robustness and component analysis: (a) F1 under contiguous manhole-sensor dropout with 95% intervals and inset radar-bias

distributions; (b) ablation waterfall for five components with inset bootstrap selection-frequency distribution and 0.80 stability threshold. Spatial transfer to the three held-out sectors produced F1 scores of 0.811, 0.786 and 0.833 after five-event calibration, and the physics-guided baseline achieved 0.742, 0.701 and 0.755 respectively. Errors were relatively higher in the flat downstream area, and tide-affected backwater was only shown by adjacent level histories. Figure 7(a) gives sector-wise transfer estimates and confidence intervals. Figure 7(b) shows the Pareto scatter plot of F1 versus latency, and in the inset, it also presents the cumulative warning cost at different decision thresholds from 0.20 to 0.80 for CausalNex-GNN. The minimum asymmetric cost was 0.62, and the range between 0.56 and 0.67 changed the total cost by less than 4%. This plateau is operationally convenient because it is less sensitive to small recalibration drift. The deployment memory was 1.7 GB, and causal discovery was performed offline; therefore, the online graph updates did not exceed the two-minute acquisition interval [30]. The flat downstream sector was harder to generalize because small water-surface gradients amplified backwater effects and made lagged response less transferable across storms.

False alarms were analyzed based on hydraulic context, not counted independently. Among the 23 proposed model false positives, 9 occurred within four minutes of verified near-overload levels, 6 coincided with unrecorded pump state changes, 5 involved radar overestimations exceeding 18%, and 3 had no identifiable data anomalies. The unrestricted graph attention network generated 41 false alarms, 15 of which were in branches without a directed hydraulic path to the active storm core. Causal restrictions reduced that disconnected subset by 27.3% compared with the comparator count. The calibrated 90% level interval had a coverage of 88.7% and a mean width of 0.42m; in severe storms, it was 0.57m. Coverage below the nominal

target occurred near the start of the overflow and changed rapidly during observation [31]. Quantile recalibration by severity increased the overall coverage to 89.6% but widened routine-event intervals; thus, the global calibration was kept.

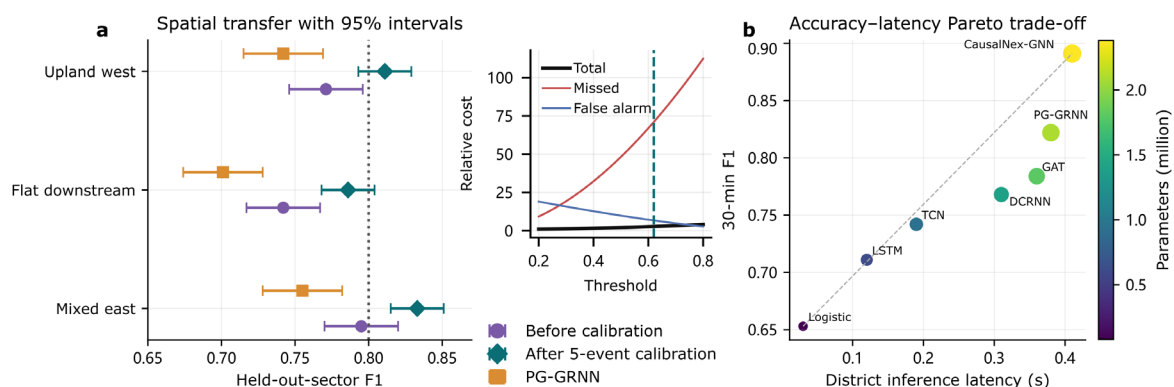


Figure 7. Transfer and deployment trade-offs: (a) held-out-sector F1 with pre- and post-calibration intervals; (b) Pareto scatter of F1, inference latency, and parameter count with inset asymmetric warning cost and its decision threshold.

The learned graph contains 312 high-confidence directed relations: 104 radar-to-subcatchment, 176 sewer-to-sewer, and 32 backwater-lag. The median lag time for upstream pipelines is 8 minutes, and the median lag time for forced diversion edges is 14 minutes. The operators listed 25 of the most important edges; among them, 21 reported asset connections, two were identified as interconnections not included in the simplified registration, and two required further investigation. Such reviews do not provide causal truth but offer an applied audit surface not available in dense attention matrices [32]. Computational profiling shows that reliability features and typed messages added 0.09s to the district update time each, and uncertainty heads added 0.02s. The offline bootstrap discovery stage required 3.8 h and was repeated only when the training archive changed. These costs are still suitable for near-real-time warnings, but larger cities will need graph partitioning and coordinated calibration of drainage districts [33]. Currently, the measured latency leaves more than 99% of each acquisition interval for communication and operator-side processing in this district.

Ablation experiments were carried out to determine the individual necessity of the causal prior, reliability gate, typed attention, quantile head, and monotonic warning penalty. Removing both the causal prior and the reliability gate reduced F1 to 0.746, which is a loss greater than the sum expected from independent effects. Under 30% dropout, that same variant reached only 0.681. Quantile head removal reduced the classification F1 score slightly but increased the expected calibration error from 0.028 to 0.046 and excessively narrowed intervals in heavy rain. Replace typed relations with a single shared message function to reduce radar-to-sewer attention diversity and delay the median first warning by five minutes. Hydraulic attributes primarily affect long-horizon regressions; excluding them increased the 60-min level error from 0.192 to 0.231m and changed the 5-min error by 0.006m. The above interactions show that the structural prior prevents learning of persistent shortcuts in the reliability mechanism, and the reliability mechanism avoids considering corrupted parents as reliable evidence from the prior.

Structured Dropout Experiment Separated Sensor Location from Missing-Data Fraction. Removing the ten highest-betweenness manholes caused a 0.094 F1 loss, compared with 0.041 for ten random sensors and 0.028 for ten peripheral sensors. The reliability gate recovered about one-third of the central-sensor loss by shifting weight to the upstream radar and neighboring branches, but the uncertainty width increased by 0.16 m. A greedy virtual-sensor analysis then added one currently uninstrumented node at a time based on the trained representation. The first five proposed sites are close to the junctions serving 38% of the historical overflow events, and simulated measurements at these sites improved F1 by 0.019. Therefore, the above results should be treated as high-priority indices rather than all-weather monitoring designs due to the model assumptions in synthetic observations. Attribute and robustness analysis of the asset allocation can be used to improve real-time forecasting.

Latency measurements include data aggregation, quality feature computation, graph inference, calibration and serialization of outlet warnings. Median end-to-end latency was 0.53 s and the 99th percentile was 0.79 s over 10,000 repeated updates. A CPU-only workstation took 3.84s and was still within the two-minute collection window. Memory grew approximately linearly with the number of active edges; retaining edges below 0.20

bootstrap confidence increased memory by 46% and slowed inference by 31%, with only an improvement in F1 of 0.003. Pruning was thus performed offline. When the radar frame arrived late, the system reused the previous field and increased the age indicator instead of blocking prediction. A 10-min radar delay reduced F1 by 0.022; a 20-min delay reduced it by 0.071 and increased the median interval by 0.11m. These experiments show the degradation behavior of supervisory integration. The reported latency remains below the two-minute data refresh interval, leaving sufficient time for serialization, operator display, and warning dissemination.

Inspection of the two causal edges questioned by operators showed different failure modes. One linked an adjacent sub-catchment that shares an undocumented overflow channel visible in aerial imagery, and its lag and storm-direction dependence were hydraulically plausible. The other two branches were connected through a common radar cell and lost bootstrap support after adding gauge-adjustment residuals, indicating residual confounding. Removing the latter edge reduced aggregate F1 by less than 0.001 but eliminated two false alarms at a peripheral outlet. Edge review was recorded with the asset version and training-period hash to distinguish between data-driven modifications and network-register corrections in the following model updates. The review process also found seven stable low-weight edges aligned with the known throttled pipes. Their small size indicated a relatively weak predictive effect rather than a structural problem, so only magnitude cannot be used to filter. These cases support interpretability as a diagnostic aid, although the learned edges should still be verified against asset records and field inspections before being used as hydraulic facts.

Conclusion

This study improved the CausalNex-GNN model for multi-horizon urban drainage overflow prediction based on weather-radar and manhole-sensor data. The architecture links a hydraulically admissible causal prior with reliability-gated heterogeneous message passing, recurrent state updates, and calibrated level, overflow, and quantile heads. The Structure of the Workflow will be interpretable and supports event-specific adaptation. Through many time-based storm tests, the model has improved the accuracy of warnings, lead time, calibration and resistance to structured data loss; causal restriction and observation-aware fusion are thus highly complementary Design choices.

The scope of the monitoring archive is still relatively small in time and space. Overflow labels are less reliable at unstaffed outlets, the simplified asset register does not include all pumps or temporary blockages, and the time-unrolled causal graph approximates hydraulic feedback rather than resolving it mechanistically. Radar correction used a relatively small gauge network, and the transfer experiment employed sectors in a single metropolitan drainage system. These restrictions limit the claims of climate, infrastructure and operation generalization. Transfer to cities with different pipe geometry, pump control logic, or monitoring density may require recalibration of the admissibility mask and reliability thresholds.

Future work will extend the model to other cities and radar systems, incorporate pump-and-gate conditions, and link learned causal relationships with differentiable hydraulic simulators. Regularly add calibration to reduce errors and preserve rare-event data, then improve the diversity of storms among utilities through privacy-preserving federated learning. Conduct field experiments to link warning probability with controllable storage and routing decisions, evaluate how humans interpret causal explanations, and quantify avoided damage and maintenance costs over an entire rainy season.

Author Contributions

Eustachy Długosz contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Szymon Wierzbicki contributes to methodology, software, validation, analysis. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

- [1] Piadeh, F., Behzadian, K., & Alani, A. M. (2022). A critical review of real-time modelling of flood forecasting in urban drainage systems. *Journal of Hydrology*, 607, 127476-127476. <https://doi.org/10.1016/j.jhydrol.2022.127476>
- [2] Liu, Y., Wang, H., Lei, X., & Wang, H. (2021). Real-time forecasting of river water level in urban based on radar rainfall: A case study in Fuzhou City. *Journal of Hydrology*, 603, 126820-126820. <https://doi.org/10.1016/j.jhydrol.2021.126820>
- [3] Galanaki, E., Lagouvardos, K., Kotroni, V., Giannaros, T. M., & Giannaros, C. (2021). Implementation of WRF-Hydro at two drainage basins in the region of Attica, Greece, for operational flood forecasting. *Natural hazards and earth system sciences*, 21(7), 1983-2000. <https://doi.org/10.5194/nhess-21-1983-2021>
- [4] Nguyen, D. H., Kim, J. B., & Bae, D. H. (2021). Improving Radar-Based Rainfall Forecasts by Long Short-Term Memory Network in Urban Basins. *Water*, 13(6), 776-776. <https://doi.org/10.3390/w13060776>
- [5] Lee, J. H., Yuk, G. M., Moon, H. T., & Moon, Y. I. (2020). Integrated Flood Forecasting and Warning System against Flash Rainfall in the Small-Scaled Urban Stream. *Atmosphere*, 11(9), 971-971. <https://doi.org/10.3390/atmos11090971>
- [6] Li, J., Hassan, D., Brewer, S., & Sitzenfrei, R. (2020). Is Clustering Time-Series Water Depth Useful? An Exploratory Study for Flooding Detection in Urban Drainage Systems. *Water*, 12(9), 2433-2433. <https://doi.org/10.3390/w12092433>
- [7] Palmitessa, R., Mikkelsen, P. S., Law, A. W. K., & Borup, M. (2020). Data assimilation in hydrodynamic models for system-wide soft sensing and sensor validation for urban drainage tunnels. *Journal of Hydroinformatics*, 23(3), 438-452. <https://doi.org/10.2166/hydro.2020.074>
- [8] Chitwatkulsiri, D., Miyamoto, H., Irvine, K., Pilailar, S., & Loc, H. H. (2022). Development and Application of a Real-Time Flood Forecasting System (RTFlood System) in a Tropical Urban Area: A Case Study of Ramkhamhaeng Polder, Bangkok, Thailand. *Water*, 14(10), 1641-1641. <https://doi.org/10.3390/w14101641>
- [9] Dong, B., Xia, J., Zhou, M., Li, Q., Ahmadian, R., & Falconer, R. A. (2022). Integrated modeling of 2D urban surface and 1D sewer hydrodynamic processes and flood risk assessment of people and vehicles. *The Science of The Total Environment*, 827, 154098-154098. <https://doi.org/10.1016/j.scitotenv.2022.154098>
- [10] Palmitessa, R., Grum, M., Engsig-Karup, A. P., & Löwe, R. (2022). Accelerating hydrodynamic simulations of urban drainage systems with physics-guided machine learning. *Water Research*, 223, 118972-118972. <https://doi.org/10.1016/j.watres.2022.118972>
- [11] Pedersen, A. N., Brink-Kjær, A., & Mikkelsen, P. S. (2022). All models are wrong, but are they useful? Assessing reliability across multiple sites to build trust in urban drainage modelling. *Hydrology and earth system sciences*, 26(22), 5879-5898. <https://doi.org/10.5194/hess-26-5879-2022>
- [12] Abbaszadeh, P., Muñoz, D. F., Moftakhari, H., Jafarzadegan, K., & Moradkhani, H. (2022). Perspective on uncertainty quantification and reduction in compound flood modeling and forecasting. *iScience*, 25(10), 105201-105201. <https://doi.org/10.1016/j.isci.2022.105201>
- [13] Sokol, Z. K., Szturc, J., Orellana-Alvear, J., Popová, J., Jurczyk, A., & Céleri, R. (2021). The Role of Weather Radar in Rainfall Estimation and Its Application in Meteorological and Hydrological Modelling—A Review. *Remote Sensing*, 13(3), 351-351. <https://doi.org/10.3390/rs13030351>
- [14] Bartos, M., & Kerkez, B. (2021). Pipedream: An interactive digital twin model for natural and urban drainage systems. *Environmental Modelling & Software*, 144, 105120-105120. <https://doi.org/10.1016/j.envsoft.2021.105120>
- [15] Sidek, L. M., Jaafar, A. S., Majid, W. H. A. W. A., Basri, H., Marufuzzaman, M., Fared, M. M., & Moon, W. C. (2021). High-Resolution Hydrological-Hydraulic Modeling of Urban Floods Using InfoWorks ICM. *Sustainability*, 13(18), 10259-10259. <https://doi.org/10.3390/su131810259>
- [16] Rapant, P., & Kolejka, J. (2021). Dynamic Pluvial Flash Flooding Hazard Forecast Using Weather Radar Data. *Remote Sensing*, 13(15), 2943-2943. <https://doi.org/10.3390/rs13152943>
- [17] Bruno, G., Pignone, F., Silvestro, F., Gabellani, S., Schiavi, F., Rebora, N., Giordano, P., & Falzacappa, M. (2021). Performing Hydrological Monitoring at a National Scale by Exploiting Rain-Gauge and Radar Networks: The Italian Case. *Atmosphere*, 12(6), 771-771. <https://doi.org/10.3390/atmos12060771>

- [18] López, M. R. L., Pedrozo-Acuña, A., & Covarrubias, M. L. S. (2021). Evaluation of ECMWF's forecasting system for probabilistic urban flood prediction: a case study in Mexico City. *Journal of Hydroinformatics*, 24(1), 38-55. <https://doi.org/10.2166/hydro.2021.072>
- [19] Werf, J. A. V. D., Kapelan, Z., & Langeveld, J. (2021). Quantifying the true potential of Real Time Control in urban drainage systems. *Urban Water Journal*, 18(10), 873-884. <https://doi.org/10.1080/1573062x.2021.1943460>
- [20] Wei, C. C., & Hsu, C. C. (2021). Real-Time Rainfall Forecasts Based on Radar Reflectivity during Typhoons: Case Study in Southeastern Taiwan. *Sensors*, 21(4), 1421-1421. <https://doi.org/10.3390/s21041421>
- [21] Mauro, C. D., Hostache, R., Matgen, P., Pelich, R., Chini, M., Leeuwen, P. J. V., Nichols, N., & Blöschl, G. (2021). Assimilation of probabilistic flood maps from SAR data into a coupled hydrologic-hydraulic forecasting model: a proof of concept. *Hydrology and earth system sciences*, 25(7), 4081-4097. <https://doi.org/10.5194/hess-25-4081-2021>
- [22] Heuvelink, D., Berenguer, M., Brauer, C., & Uijlenhoet, R. (2020). Hydrological application of radar rainfall nowcasting in the Netherlands. *Environment International*, 136, 105431-105431. <https://doi.org/10.1016/j.envint.2019.105431>
- [23] Mignot, E., & Dewals, B. (2022). Hydraulic modelling of inland urban flooding: Recent advances. *Journal of Hydrology*, 609, 127763-127763. <https://doi.org/10.1016/j.jhydrol.2022.127763>
- [24] Cui, Z., Zhou, Y., Guo, S., Wang, J., & Xu, C. Y. (2022). Effective improvement of multi-step-ahead flood forecasting accuracy through encoder-decoder with an exogenous input structure. *Journal of Hydrology*, 609, 127764-127764. <https://doi.org/10.1016/j.jhydrol.2022.127764>
- [25] Kilsdonk, R. A. H., Bomers, A., & Wijnberg, K. M. (2022). Predicting Urban Flooding Due to Extreme Precipitation Using a Long Short-Term Memory Neural Network. *Hydrology*, 9(6), 105-105. <https://doi.org/10.3390/hydrology9060105>
- [26] Klotz, D., Kratzert, F., Gauch, M., Sampson, A. K., Brandstetter, J., Klambauer, G., Hochreiter, S., & Nearing, G. (2022). Uncertainty estimation with deep learning for rainfall-runoff modeling. *Hydrology and earth system sciences*, 26(6), 1673-1693. <https://doi.org/10.5194/hess-26-1673-2022>
- [27] Herath, H. M. V. V., Chadalawada, J., & Babovic, V. (2021). Hydrologically informed machine learning for rainfall-runoff modelling: towards distributed modelling. *Hydrology and earth system sciences*, 25(8), 4373-4401. <https://doi.org/10.5194/hess-25-4373-2021>
- [28] Rosin, T., Romano, M., Keedwell, E., & Kapelan, Z. (2021). A Committee Evolutionary Neural Network for the Prediction of Combined Sewer Overflows. *Water Resources Management*, 35(4), 1273-1289. <https://doi.org/10.1007/s11269-021-02780-z>
- [29] Moishin, M., Deo, R. C., Prasad, R., Raj, N., & Abdulla, S. (2021). Designing Deep-Based Learning Flood Forecast Model With ConvLSTM Hybrid Algorithm. *IEEE Access*, 9, 50982-50993. <https://doi.org/10.1109/access.2021.3065939>
- [30] Imhoff, R., Brauer, C., Heeringen, K. J. V., Leijnse, H., Overeem, A., Weerts, A., & Uijlenhoet, R. (2021). A climatological benchmark for operational radar rainfall bias reduction. *Hydrology and earth system sciences*, 25(7), 4061-4080. <https://doi.org/10.5194/hess-25-4061-2021>
- [31] Li, X., Rankin, C., Gangrade, S., Zhao, G., Lander, K., Voisin, N., Shao, M., Morales-Hernández, M., Kao, S. C., & Gao, H. (2021). Evaluating precipitation, streamflow, and inundation forecasting skills during extreme weather events: A case study for an urban watershed. *Journal of Hydrology*, 603, 127126-127126. <https://doi.org/10.1016/j.jhydrol.2021.127126>
- [32] Shi, B., Catsamas, S., Kolotelo, P., Wang, M., Lintern, A., Jovanović, D. A., Bach, P. M., Deletić, A., & McCarthy, D. (2021). A Low-Cost Water Depth and Electrical Conductivity Sensor for Detecting Inputs into Urban Stormwater Networks. *Sensors*, 21(9), 3056-3056. <https://doi.org/10.3390/s21093056>
- [33] Mendoza-Cano, O., Aquino-Santos, R., Cruz, J. L. D. L., Edwards, R. M., Khouakhi, A., Pattison, I., Rangel-Licea, V. C., Castellanos-Berjan, E., Martínez-Preciado, M. A., Rincón-Avalos, P., Lepper, P. A., Gutiérrez-Gómez, A., Uribe-Ramos, J. M., Ibarreche, J., & Pérez, I. (2021). Experiments of an IoT-based wireless sensor network for flood monitoring in Colima, Mexico. *Journal of Hydroinformatics*, 23(3), 385-401. <https://doi.org/10.2166/hydro.2021.126>