

Artificial Intelligence for Urban Flood Early Warning: A Review of Recent Advances

Fanghao Liu *

College of Water Resources and Architectural Engineering, Northwest A&F University, Yangling, Shaanxi, 712100, China

* Corresponding author Email: lf7796068092021@163.com

Abstract: Urban flooding is among the most frequent and costly climate-related hazards, and short-term early warning is one of the few interventions that can reduce its impact without new physical infrastructure. This review examines how artificial intelligence has been applied to short-term urban flood early warning, based on 67 studies published between 2019 and 2026 and retrieved from Web of Science and CNKI. Studies were screened through a four-stage PRISMA process and coded across ten fields, including method family, data source, forecast lead time, sample size, validation protocol, and the use of social-attribute variables. Reported accuracy is high across all method families, with R^2 values typically above 0.95, but these figures are not comparable across studies because the metrics, sample sizes, validation protocols, and evaluation targets all differ. Four systematic limitations emerge. First, 55 of the 67 studies train on simulator-generated labels rather than observed inundation, so the reported accuracy measures fidelity to a simulator, not to a real flood. Second, 53 of the 67 studies include no social-attribute variable in the model input. Third, validation protocols and sample sizes vary by orders of magnitude and are reported inconsistently. Fourth, none of the studies reports what happened after a warning was issued, so the claim that AI-based early warning improves flood outcomes remains an assumption rather than a finding.

Keywords: Urban Flooding; Artificial Intelligence; Deep Learning; Early Warning System; Flood Forecasting; Machine Learning; Brier Score; PRISMA.

1. Introduction

Urban flooding occurs when rainfall exceeds the capacity of a city's drainage system, causing water to accumulate in streets, underpasses, and other low-lying areas. Unlike river floods, urban floods are characterised by shallow depths, short durations, and high frequency, and they occur precisely where population and infrastructure are most concentrated. The Intergovernmental Panel on Climate Change has documented a shift in the spatial and temporal patterns of the water cycle, with extreme precipitation frequency projected to increase by approximately 7% by 2030. In parallel, global impervious surface area expanded by 28% between 2000 and 2020. These two drivers have combined to raise urban flood risk: the frequency of flood disasters in 2021 was 41% higher than the average of the preceding decade, and the World Resources Institute projects that the population annually affected by coastal flooding will double from 7 million to 15 million by 2030. In China alone, compound flooding from rainfall and storm surges in coastal megacities causes more than USD 50 billion in annual economic losses.

Early warning is one of the few interventions that can reduce these losses without requiring new physical infrastructure. A warning issued with sufficient lead time allows drainage operators to pre-emptively lower storage levels, allows traffic to be rerouted, and allows residents in vulnerable locations to move to safety. The value of a warning system therefore depends on two things that are often conflated: the accuracy of the forecast, and the speed at which it can be produced and communicated. Traditional hydrodynamic models, such as SWMM, MIKE, and InfoWorks ICM, achieve high accuracy but require detailed terrain and pipe network data, and their computational cost scales steeply with resolution. For sub-hourly forecasting over an entire city, these models struggle to produce results

within the time window in which those results are still useful.

Artificial intelligence methods, machine learning, deep learning, and their hybrids with physical models, have been proposed as a way to close this gap. By learning a mapping from rainfall inputs to inundation outputs, a trained model can generate a prediction in seconds rather than hours, and can do so on hardware that is available to most municipal agencies. The number of studies applying these methods to urban flooding has grown rapidly since 2019, and the methodological diversity has grown with it. It is no longer possible to summarise this literature by listing individual studies, and it is not yet clear which of the reported advances are robust and which are artefacts of the way they were evaluated.

Four recent reviews have addressed adjacent questions. Li et al. [1] surveyed 181 studies published between 2004 and 2024 on AI-incorporated urban flood prediction, covering the field as a whole. Zeng et al. [2] reviewed the transition from traditional numerical models to deep learning approaches, with a focus on model structure and applicability. Piadeh et al. [3] reviewed real-time flood forecasting models specifically for urban drainage systems, with attention to data sources and performance assessment. Yan et al. [4] used CiteSpace co-citation analysis to map the application of advanced technologies, AI, remote sensing, crowdsourced data, and the Internet of Things, in urban flood model development. Each of these reviews covers a substantial portion of the field, and together they establish that AI methods are now central to urban flood forecasting research.

None of them, however, focuses specifically on the intersection that matters most for early warning: short-term and near-real-time forecasting, as distinct from risk mapping, offline simulation, and post-event reconstruction. Li et al. [1] report that among their 181 studies, 123 addressed flood risk mapping, while only 38 addressed forecasting with a lead

time of one hour or less. Aggregating risk mapping and real-time forecasting into a single review blurs the methodological distinction between them, because the two tasks differ in their input data, their validation protocols, and their evaluation metrics. A second gap concerns language. Li et al. [1] restricted their search to English-language publications, excluding 24 non-English studies on that basis. A substantial share of the global research output on urban flooding originates in Chinese cities and is published in Chinese, and excluding it introduces a systematic regional bias. A third gap concerns what is recorded about each study. Existing reviews report the distribution of accuracy across algorithms, but they do not separately tabulate sample size, validation protocol, or the use of social-attribute variables. Without these columns, the reported accuracy distributions cannot be interpreted, because a high R^2 on a small hold-out split and a high R^2 on a large cross-validated dataset are not equivalent evidence.

This review addresses these gaps by narrowing its scope to short-term and near-real-time forecasting, by including Chinese-language literature alongside English-language sources, and by coding each included study across ten fields, including sample size, validation protocol, and the use of social-attribute variables. Three research questions guide the review:

RQ1. What are the publication trends, methodological distributions, and geographic concentrations of AI-based short-term urban flood early warning research?

RQ2. What input variables, forecast lead times, data sources, and reported performance levels characterise the models used in this field?

RQ3. What systematic limitations exist in variable selection, validation practices, data provenance, and the evaluation of early warning performance?

The remainder of this paper is organised as follows. Section 2 describes the search strategy and inclusion criteria. Section 3 presents the bibliometric results, addressing RQ1. Section 4 reviews the six method families identified in the corpus, addressing RQ2. Section 5 discusses the four systematic limitations that emerge from the review, addressing RQ3. Section 6 concludes.

2. Methods

2.1. Research Questions and Scope

This review aims to examine how artificial intelligence (AI) has been applied to short-term urban flood early warning, and to identify the systematic limitations that persist across this body of work. Three research questions guide the review:

RQ1. What are the publication trends, methodological distributions, and geographic concentrations of AI-based short-term urban flood early warning research?

RQ2. What input variables, forecast lead times, data sources, and reported performance levels characterise the models used in this field?

RQ3. What systematic limitations exist in variable selection, validation practices, data provenance, and the evaluation of early warning performance?

This review is positioned relative to the recent critical review by Li et al. [1], which examined AI applications in urban flood prediction over the period 2004–2024. That review covered the field at the urban scale as a whole, aggregating studies across risk mapping, offline simulation, and real-time forecasting. It reported that among 181 included studies, 123 addressed flood risk mapping, while only 38

addressed forecasting within a one-hour lead time. Its exclusion criteria also removed studies without an AI component, studies at non-urban scales, and non-English publications.

The present review differs in three respects. First, it narrows the scope to short-term and near-real-time forecasting, excluding static risk mapping and post-event reconstruction. Second, it extends the search window to 2026, capturing literature published after the cut-off of [1]. Third, it includes Chinese-language journal articles alongside English-language publications, thereby reducing the regional bias that a single-language search may introduce. These differences are reflected in the inclusion and exclusion criteria described in Section 3.3.

2.2. Search Strategy

Literature was retrieved from two databases: the Web of Science Core Collection and the China National Knowledge Infrastructure (CNKI). The search was conducted on 30 August 2026. The search window was set to 2022–2026, reflecting the rapid growth of this field after 2019 and the objective of capturing the most recent methodological developments.

3. Results

3.1. Publication Trends

The 67 included studies span the period from 2019 to 2026, with a pronounced concentration in the most recent three years. Studies published in 2024, 2025, and 2026 together account for the majority of the corpus, while work published before 2022 constitutes a small minority. This distribution is consistent with the broader trajectory reported in [1], where publication volume in AI-based urban flood research remained low until 2018 and then increased sharply. The concentration in the present corpus is sharper still, because the search window was deliberately restricted to 2022–2026.

Two observations follow from this distribution. First, the field is young and still expanding; the majority of the methods reviewed here were introduced within the last three years, which means that comparative evidence across methods remains limited. Second, the recency of the corpus means that long-term validation records are rare. Studies that report performance over a single rainfall season, or over a small number of storm events, cannot yet demonstrate that their models remain reliable across years and across changing urban surfaces.

3.2. Method Distribution

Across the 67 studies, a wide range of model families is represented, but a small number of approaches account for a disproportionate share of the corpus. Deep learning architectures, particularly convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and their hybrid combinations, appear most frequently. Convolutional architectures are typically used where the prediction target is a spatial field, such as an inundation depth map, while recurrent architectures are used where the target is a time series at a fixed location, such as the water depth at a monitoring point. A substantial number of studies combine the two, using a convolutional encoder to extract spatial features and a recurrent module to model their temporal evolution.

Ensemble tree methods, including random forest, XGBoost,

LightGBM, and CatBoost, form the second major group. These methods are frequently applied to point-scale prediction, where the input is a rainfall time series and the output is a water depth or a risk level, and where sample sizes are moderate. Their persistence in a field otherwise dominated by deep learning suggests that, for small tabular datasets, tree-based ensembles remain competitive.

A third, smaller group applies optimisation algorithms to tune the hyperparameters of a base model. Particle swarm optimisation, genetic algorithms, simulated annealing, and Bayesian optimisation all appear in the corpus, generally coupled to a neural network or a support vector machine. These studies consistently report improvements over the untuned baseline, but the improvements vary widely in magnitude, and the baselines differ across studies.

Finally, a small number of studies adopt physics-informed or hybrid physics-data approaches, in which a hydrodynamic model is either embedded in the loss function or used to generate training data. This group is small but growing, and it addresses a limitation discussed in Section 5.

3.3. Forecast Horizon and Spatial Scale

Lead times across the corpus range from minutes to 24 hours. Studies at the shortest end of this range report lead times of 5 to 30 minutes, typically for point-scale prediction of water depth at a specific monitoring station. Studies at the longer end report lead times of 1 to 6 hours, and in a few cases up to 12 or 24 hours, typically for spatial prediction of inundation extent. A small number of studies do not state a lead time explicitly and instead describe their system as "real-time"; where such studies reported a predictive output with quantitative validation, they were retained, and their lead time is recorded as not reported in the coding table.

Spatial resolution varies equally widely. Studies that predict a spatial field report grid resolutions between 2 m and 80 m, with 5 m and 10 m being the most common. Studies that predict at a point scale do not report a grid resolution at all. The size of the study area ranges from a single street intersection to an entire metropolitan region, and the majority of studies focus on a sub-district or district-scale catchment rather than a whole city.

This distribution indicates that the field has not converged on a standard spatial or temporal scale. A model validated at 2 m resolution over a 3 km² catchment is not directly comparable to a model validated at 80 m resolution over a 100 km² catchment, even when both report the same metric. This point is developed further in Section 5.1.

3.4. Geographic Distribution

The included studies are distributed unevenly across regions. China accounts for the largest share of the corpus, followed by South Korea, Germany, and the United States. Studies from the Middle East, South America, and Africa are comparatively rare, despite the fact that urban flooding is a recurring hazard in those regions.

This distribution has two implications. First, the methodological evidence base is dominated by a small number of countries, most of which have relatively dense monitoring networks and well-documented drainage infrastructure. Whether the reported performance levels transfer to cities with sparser observational data is not established by the current literature. Second, the regions least represented in the corpus are not necessarily the regions least affected by urban flooding. The gap between where the

hazard occurs and where the research is conducted remains wide.

3.5. Summary of the Bibliometric Findings

Taken together, the bibliometric results describe a field that is expanding rapidly, methodologically diverse, and geographically concentrated. Deep learning methods dominate, lead times and spatial resolutions vary by more than an order of magnitude, and a small number of countries produce most of the evidence. These patterns form the basis for the critical discussion in Section 5.

4. Application

4.1. Hybrid Spatiotemporal Architectures

The largest group in the corpus comprises studies that combine convolutional and recurrent components in a single architecture. Eighteen of the 67 included studies fall into this category. The common motivation is structural: urban inundation is simultaneously a spatial and a temporal phenomenon, and a model that processes only one dimension at a time cannot capture both. A purely convolutional model can map a rainfall field to a water-depth field, but it has no mechanism for carrying information forward from one time step to the next. A purely recurrent model can carry temporal state forward, but it must treat space as a flat vector of independent locations. Hybrid architectures are an attempt to resolve this tension.

The dominant design pattern is an encoder-decoder structure in which the encoder alternates convolutional and recurrent blocks, and the decoder reconstructs a spatial field at each forecast step. Cao et al. [5] provide a representative example. Their U-RNN architecture combines a skip-connected convolutional GRU with a latent autoregressive U-shaped recurrent network, and is trained on a self-supervised warm-up paradigm. Evaluated over a 1 km² catchment in Shenzhen at 2 m spatial resolution and 1 min temporal resolution, the model produces six-hour-ahead nowcasts of inundation extent, peak depth, and hydrographs. Reported MAE is 0.001-0.003 m and RMSE is 0.008-0.014 m, with CSI values of 0.892-0.922 for observed rainfall events and above 0.970 for design rainfall events.

A second design pattern separates the spatial and temporal tasks into two models. One study couples an InfoWorks ICM hydrodynamic model with a PSO-calibrated LSTM for point-scale water-depth time series and a CNN for city-scale peak-depth maps. The LSTM component reports a mean MAE of 3.05 cm, and the CNN reports a mean depth error of 3.6-4.3 cm with an IoU of 0.87 on the test set. The reported NSE exceeds 0.90, and the LSTM inference time for a 24-hour event is approximately 0.05 seconds.

A third pattern addresses generalisability rather than accuracy. In [6], a context-aware architecture that combines a multi-scale CNN encoder, an attention mechanism, an RNN, and a U-Net decoder, and evaluate it not only on the training city but also on two unseen cities. On the Zurich test events, the model achieves RMSE values of 0.0153-0.0196 m and CSI values of 0.97-0.98. When transferred to Lucerne and Singapore without retraining, the RMSE at the 0.1 m threshold rises to 0.17 m and 0.10 m respectively, and CSI falls to 0.72 and 0.66. The degradation is substantial, and the authors present it as the central result: a model that performs well in one city does not automatically perform well in another.

Taken together, these studies show that hybrid architectures are the most widely adopted approach in the current literature, and that they consistently achieve low point-wise error and high spatial overlap with hydrodynamic reference simulations. What they do not yet show is that this performance transfers across cities, rainfall regimes, or infrastructure configurations. The reported metrics are almost always computed against a hydrodynamic model rather than against observed inundation, and the one study that explicitly tests cross-city transfer reports a marked loss of accuracy. This limitation is taken up again in Section 5.

4.2. Ensemble Tree Models

Thirteen of the 67 included studies apply ensemble tree methods, either as the primary predictor or as a benchmark against which deep learning models are compared. The appeal of these methods in an urban flooding context is threefold. They perform well on small tabular datasets, which is often all that is available at a single monitoring station. They are robust to noisy and missing inputs, which matters when sensor records are irregular. And they provide a direct measure of variable importance, which makes the resulting model easier to interpret than a deep network. The trade-off is that they cannot natively process spatial fields: a random forest or an XGBoost model sees a flat feature vector, not an inundation map.

The most common application is point-scale prediction of water depth, in which the input is a rainfall time series and the output is the water depth at one or more monitoring locations. One study provides a clear example of how far this approach can be pushed when combined with domain knowledge. Working in Fengxi New Town, Shaanxi, over a 3.68 km² catchment, they combine five machine learning algorithms, KNN, random forest, extremely randomised trees, ridge regression, and lasso, with a fuzzy rain-pattern classifier, so that a different algorithm is used depending on whether the incoming storm is single-peaked, double-peaked, or uniform. The reported mean absolute percentage error is 5.32% for single-peak events, 7.73% for double-peak events, and 2.49% for uniform events. The full prediction for the 3.68 km² area is completed in 14.07 seconds. The design choice worth noting is that the authors do not select a single best algorithm; they select the algorithm that performs best under each rainfall pattern.

A second application is to predict a quantity that is not directly observable but that drives flooding. Another study forecasts the overflow volume at manholes in Nanchong, Sichuan, using a Bayesian neural network with a spatiotemporal effect term, and compare it against ANN, LightGBM, and LSTM baselines. The proposed model achieves an R^2 of 0.982 and a root mean square error of 0.147 on the validation set, against 0.920 for ANN, 0.944 for LightGBM, and 0.955 for LSTM. The inclusion of uncertainty estimates, CRPS 0.074, prediction interval coverage 0.948, is unusual in this literature and directly addresses a limitation discussed in Section 5.

A third application is surrogate modelling, in which a tree-based model is trained on the output of a hydrodynamic simulator and then used in place of it. Guo et al. [7] train nine machine learning models on 806 rainfall-tide scenarios generated by a MIKE+ 1D-2D coupled model of Ho Chi Minh City, covering 270,000 inundation points. Genetic programming achieves the highest R^2 of 0.997, with random forest close behind at 0.996, followed by neural networks at

0.995. The trained surrogate produces predictions in seconds, against 25 minutes to 1.5 hours for a single MIKE+ simulation, and the authors report that the system has been deployed to issue inundation warnings three hours in advance. This study is one of the few in the corpus that connects a machine learning model to an operational warning service.

A fourth application uses tree methods to predict not depth but water level in the drainage network. Another study combines PCA-based point selection with XGBoost to nowcast water levels at pipe nodes and surface nodes over a 24-hour horizon. The surface-level model reports an R^2 of 0.909 and an RMSE of 0.144, while the network-level model reports an R^2 of 0.899 and an RMSE of 0.162. At two independent monitoring stations, the 24-hour forecast achieves R^2 values of 0.973 and 0.962. Total computation time is 0.29 seconds on CPU and 0.25 seconds on GPU.

Across these studies, ensemble tree methods consistently report high R^2 values, often above 0.95, and consistently demonstrate inference times short enough for operational use. The recurring limitation is that their inputs are almost always rainfall and, where available, upstream water level; the spatial configuration of the city enters only indirectly, through the hydrodynamic model that generated the training labels. Whether a tree-based surrogate trained on one city's drainage network can be transferred to another is not addressed by any of the thirteen studies.

4.3. Convolutional Neural Networks

Eleven of the 67 included studies use a convolutional neural network (CNN) as their primary architecture. The defining characteristic of this group is that the prediction target is a spatial field: a grid of water depths covering a catchment, a district, or a city. A CNN is a natural fit for this task because its convolutional layers are designed to detect local spatial patterns, and because the same set of learned filters can be applied across the entire grid. The trade-off is that a standard CNN has no memory: it processes each input independently, and must therefore be given the rainfall history as an additional input channel if the prediction is to reflect how the storm has evolved up to the present moment.

The most direct application is to map a rainfall scenario to a peak-depth map. One study trains a one-dimensional CNN on 49 design storm scenarios generated by a coupled SWMM/LISFLOOD-FP model of the Liede River basin in Guangzhou, covering 14.38 km² at 5 m resolution. The model reports an average correlation coefficient of 0.982, a mean absolute error of 0.021 m, and a root mean square error of 0.035 m. The most notable result is computational: a 100-year storm scenario that takes 31 minutes in the coupled hydrodynamic model is predicted in 20 seconds on a laptop CPU, and the reported speed-up is 210 times on GPU and 51 times on CPU. The authors also apply SHAP analysis to identify which input features drive the predictions, and report that a derived elevation feature (R_DEM) is the most influential, followed by manhole density and raw elevation.

A similar approach but at a finer temporal grain. Working in Zhoukou, China, they train a 1D-CNN on samples generated by a MIKE URBAN-MIKE 21 coupled model, using the current and five preceding rainfall intervals as input. The model predicts water depth at 20 flood-prone points with a mean error of 5.74%, compared with 6.97% for the hydrodynamic model itself. The overall NSE is 0.967 and the KGE is 0.969, and a three-hour rainfall scenario is predicted in approximately eight seconds, against roughly ten hours for

the numerical model. Feature importance analysis identifies elevation, land use, and pipe density as the three strongest predictors.

A third variant addresses interpretability more directly. (2026b) train a two-dimensional CNN on 42 storm scenarios generated by a WCA2D-SWMM coupled model of a 70-hectare community in Guangzhou, and apply both SHAP and a gradient-based attribution method. The test-set metrics are a mean absolute error of 0.0181 m, a root mean square error of 0.0407 m, and a critical success index of 0.707 at the 0.05 m threshold. The attribution analysis reports that building footprints contribute 30.3% of the prediction, elevation 28.9%, and aspect 17.9%, with terrain-related features accounting for 90.1% of the total and infrastructure-related features for 9.9%.

The most methodologically distinct study in this group is presented in [8], which moves beyond single-catchment prediction to a multi-catchment setting. They train a U-Net with a temporal attention encoder on 1,500 rainfall events across 100 Swiss catchments, and evaluate both on catchments seen during training and on catchments held out entirely. On the seen catchments, the mean absolute error ranges from 0.75 cm to 23.59 cm depending on water depth, and remains as low as 27 cm even for depths exceeding 1 m. On the unseen catchments, the error range widens to 1.18–57.06 cm. The gap between the two is the paper's central finding: a CNN trained on a set of catchments does not generalise automatically to a new catchment, even within the same country and the same rainfall regime.

Taken together, these studies show that CNNs can reproduce hydrodynamic simulation output with high spatial fidelity and at a fraction of the computational cost, and that attribution methods can identify which spatial features the model relies on. They also show that the reported performance is almost always measured against a hydrodynamic model rather than against observed inundation, and that the one study that tests cross-catchment transfer reports a substantial loss of accuracy. Both points are revisited in Section 5.

4.4. Recurrent Neural Networks

Eleven of the 67 included studies use a recurrent neural network as their primary architecture, most commonly a long short-term memory (LSTM) network or one of its variants. The defining characteristic of this group is that the prediction target is a time series at a fixed location, a water depth recorded at a monitoring station, a water level in a pipe, or an overflow volume at a manhole. A recurrent network is a natural fit for this task because its internal state carries information forward from one time step to the next, allowing the model to represent the accumulation and recession of water without being told explicitly how the catchment responds. The trade-off is that the model is spatially blind: each monitoring point is modelled independently, or all points are flattened into a single feature vector, and the spatial arrangement of the drainage network enters only indirectly through the training data.

One study provides an example of the rainfall-forecasting side of this group. Working in Guwahati, India, over a 227 km² urban catchment, they train a modified LSTM to forecast rainfall at lead times of one to six hours, and then drive a HEC-RAS Rain-on-Grid hydrodynamic model with the forecast rainfall to simulate inundation depth. The rainfall model achieves an R^2 of 0.942 on the test set, with R^2 above

0.7 at lead times of one to two hours. Relative to a baseline LSTM, RMSE is reduced by 72.4% and SMAPE by 74.0%. The simulated inundation is validated against Sentinel-1 SAR imagery, with a critical success index of 0.62, a hit rate of 0.78, and a false alarm ratio of 0.21 at the 0.05 m threshold. This study is one of the few in the corpus that chains a rainfall forecast to a hydrodynamic model and validates the final product against satellite observations rather than against a simulator.

Another study takes a different approach, using LSTM not to predict rainfall or inundation directly, but to supply a boundary condition to a hydrodynamic model. Working in the Doumen area of Fuzhou over a 33.7 km² catchment with 2,093 manholes, they first apply a time-lag correlation analysis to identify the lag between upstream rainfall and downstream water level, then train an LSTM to predict the downstream river level. That predicted level is used as the real-time downstream boundary condition for a one-dimensional Saint-Venant solver, which in turn predicts the water level in the manholes. The LSTM component achieves an NSE of 0.969–0.971 for river level, and the coupled system achieves an NSE of 0.826–0.945 for manhole water level, with RMSE between 0.047 m and 0.127 m. The reported computation time is under one minute, against four minutes and seven seconds for the equivalent hydrodynamic simulation.

Another study addresses a problem that the other studies in this group do not: what happens when the model must remain accurate after it has been deployed. Their RP-SN architecture splits the prediction task into two segments, one predicting lateral inflow at pipe nodes and one predicting surface inundation, and adds a fine-tuning update mechanism that ingests new sensor data and adjusts the model online. On the test set, the un-updated model achieves an accuracy above 85% and an NSE above 0.85 for water level and flow. After updating with new data, spatial accuracy rises from 83% to 86%, RMSE for inundation depth falls from 0.011 m to 0.008 m, and the correlation for flow prediction rises from 0.40 to 0.80. The flow correlation before updating is the more revealing number: a model that reports a high NSE for water level can still perform poorly for flow, and the gap between the two is only closed once new data is incorporated.

Taken together, these studies show that recurrent architectures can predict point-scale water levels and depths with high accuracy and at inference times short enough for operational use, and that they can be coupled to hydrodynamic solvers in either direction, driving them with forecast rainfall, or supplying them with boundary conditions. The recurring limitation is that each model is trained and evaluated on a single catchment or a single set of monitoring points, and none of the eleven studies reports whether the trained model transfers to a different drainage network, a different rainfall regime, or a different city.

4.5. Physics-Informed and Surrogate Models

Seven of the 67 included studies use a physics-informed or surrogate modelling approach, in which a hydrodynamic simulator is either embedded inside the learning process or used to generate the training data that the model then reproduces. The defining characteristic of this group is that physical conservation laws are treated as a constraint rather than as a target to be learned from examples alone. The appeal of this approach is that it can produce predictions that respect mass and momentum balance, even in scenarios for which no

training data exists. The trade-off is that the resulting model is more complex to implement and typically requires greater computational resources during training.

The clearest example of the surrogate approach is presented in [9], which trains four neural network architectures, a standard deep neural network, a Bayesian neural network, a physics-guided neural network, and an LSTM, on water depth and velocity fields generated by a HEC-RAS 2D model. The training data cover a one-hectare university campus in Houston at 1 m resolution with over one million pixels, and a nine-hectare residential area used for transfer testing. On the training catchment, the physics-guided variant achieves a median RMSE of approximately 2 mm over a one-hour rainfall event, and is between 15.9 and 52.2 times faster than the HEC-RAS model. When transferred to the residential catchment without fine-tuning, the median RMSE rises to 44.4 cm, and 73.4% of pixels remain below a 1 m error. The contrast between 2 mm and 44.4 cm is the paper's central result, and it echoes the transfer findings in Sections 4.1 and 4.3.

The second example addresses a question that the surrogate studies leave open: how should a training set be partitioned when the prediction target is a street-level time series? One study compares four strategies for training an LSTM surrogate on 20 storm events in Norfolk, Virginia, covering 40 flood-prone streets. The four strategies range from training a single global model to training a separate model for each street. The street-specific model achieves the lowest RMSE at 0.0281 m, followed by the cluster-based models at 0.0300 m, and the global model at 0.0325 m. Peak depth errors are generally below 0.1 m and peak timing errors below one hour. The difference between the best and worst strategies is less than 5 mm, but the direction of the result is consistent: a model trained on a subset of streets that behave similarly outperforms a single global model.

Taken together, these studies show that physics-informed and surrogate models can reproduce hydrodynamic output with high accuracy at low computational cost, and that the choice of how to partition the training data across locations has a measurable effect on performance. They also show that the transfer problem identified in Section 4.1 persists: a surrogate trained on one site does not automatically perform well on another, and the loss of accuracy can be two orders of magnitude larger than the within-site error.

5. Conclusion

This review has examined 67 studies published between 2019 and 2026 on the application of artificial intelligence to short-term urban flood early warning. The studies were retrieved from Web of Science and CNKI, screened through a four-stage PRISMA process, and coded across ten fields including method family, data source, forecast horizon, sample size, validation protocol, and the inclusion of social-attribute variables. Three research questions guided the review: how the field is distributed across methods and regions, what input variables and forecast horizons characterise the models, and what systematic limitations persist across the literature.

With respect to the first question, the field is expanding rapidly and is dominated by deep learning. Hybrid architectures that combine convolutional and recurrent components account for the largest share of the corpus, followed by ensemble tree models, standalone CNNs, standalone recurrent networks, physics-informed and

surrogate models, and conventional machine learning. Publication volume is concentrated in the most recent three years, and the geographic distribution is uneven, with China, South Korea, Germany, and the United States accounting for the majority of studies. The regions where urban flooding causes the greatest damage are not the regions where most of the research is conducted.

With respect to the second question, the models in this corpus use rainfall and terrain as their primary inputs, operate at forecast horizons ranging from minutes to twenty-four hours, and report accuracy metrics that are consistently high. RMSE values below 0.05 m and R^2 values above 0.95 appear across all six method families. Inference times are short in every family, and speed relative to a hydrodynamic simulator is reported as the principal advantage of the data-driven approach.

With respect to the third question, four limitations recur across the literature. First, the reported accuracy metrics are not comparable across studies, because the metrics themselves, the sample sizes, the validation protocols, and the evaluation targets all differ. Second, input variables are rarely selected on the basis of a stated hydrological rationale, and the resulting variable sets are shaped more by data availability than by an explicit theory of what drives urban flooding. Third, 55 of the 67 studies train their models on simulator-generated labels rather than on observed inundation, which means that the reported accuracy measures the model's fidelity to the simulator rather than its ability to predict real floods. Fourth, none of the 67 studies reports what happened after a warning was issued, so the claim that AI-based early warning improves urban flood outcomes remains an assumption rather than an empirical finding.

References

- [1] Li, Z., Zhou, Z., Wang, H., et al. (2025). Artificial intelligence-incorporated prediction for urban flooding processes in the past 20 years: A critical review. *Environmental Modelling & Software*, 192, 106525. <https://doi.org/10.1016/j.envsoft.2025.106525>.
- [2] Zeng, B., Huang, G., & Chen, W. (2025). Research progress and prospects of urban flooding simulation: From traditional numerical models to deep learning approaches. *Environmental Modelling & Software*, 183, 106213. <https://doi.org/10.1016/j.envsoft.2024.106213>.
- [3] 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. <https://doi.org/10.1016/j.jhydrol.2021.127476>.
- [4] Yan, Y., Zhang, N., & Zhang, H. (2023). Applications of advanced technologies in the development of urban flood models. *Water*, 15(4), 622. <https://doi.org/10.3390/w15040622>.
- [5] Cao, X., Wang, B., Yao, Y., et al. (2025). U-RNN high-resolution spatiotemporal nowcasting of urban flooding. *Journal of Hydrology*, 659, 133117. <https://doi.org/10.1016/j.jhydrol.2025.133117>.
- [6] do Lago, C. A. F., Giacomoni, M. H., Bentivoglio, R., et al. (2023). Generalizing rapid flood predictions to unseen urban catchments with conditional generative adversarial networks. *Journal of Hydrology*, 618, 129276. <https://doi.org/10.1016/j.jhydrol.2022.129276>.
- [7] Guo, Z., Leitão, J. P., Simões, N. E., et al. (2021). Data-driven flood emulation: Speeding up urban flood predictions by deep convolutional neural networks. *Journal of Flood Risk*

- Management, 14(1), e12684. <https://doi.org/10.1111/jfr3.12684>.
- [8] Löwe, R., Böhm, J., Jensen, D. G., et al. (2021). U-FLOOD – Topographic deep learning for predicting urban pluvial flood water depth. *Journal of Hydrology*, 603, 126898. <https://doi.org/10.1016/j.jhydrol.2021.126898>.
- [9] He, J., Zhang, L., Xiao, T., et al. (2023). Deep learning enables super-resolution hydrodynamic flooding process modeling under spatiotemporally varying rainstorms. *Water Research*, 239, 120057. <https://doi.org/10.1016/j.watres.2023.120057>.
- [10] Rözer, V., Peché, A., Berkhahn, S., et al. (2021). Impact-based forecasting for pluvial floods. *Earth's Future*, 9(2), e2020EF001851. <https://doi.org/10.1029/2020EF001851>.