https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

Threshold Effects and Generalization Bias in AI-based Urban Pluvial Flood Prediction: Insights from a Dataset Design Perspective

Hao Hu[1], Fei Xiao[2], Peng Xu[1], Yuxuan Gao[3], Dongfang Liang[3], Yizi Shang[1,3,*]

1) Yellow River Conservancy Technical University, Kaifeng 475004, China
2) North China University of Water Resources and Electric Power, Zhengzhou 450045, China
3) Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK

Correspondence to: Yizi Shang (yzshang@foxmail.com)

Highlights:

- Proposes a three-dimensional dataset design framework (length, feature combination, rainfall distribution) for AI-based urban pluvial flood prediction.
- Identifies a threshold effect of data length (~14,400 samples) where model performance significantly improves and then saturates.
- Reveals that rainfall distribution dominates model generalization and bias, with mixed-intensity training achieving the best robustness.
- Shows that the effectiveness of multi-feature inputs (P+I+D) depends on dataset size, improving stability only when sufficient data are available.
- Integrates a hydrological–hydrodynamic model with machine learning, enabling reliable training data generation in data-scarce urban areas.

Abstract: Reliable urban flood prediction hinges on how datasets are designed, yet most existing research disproportionately emphasizes network architectures over data foundations. This study systematically investigates how dataset characteristics—scale, feature composition, and rainfall-event distribution—govern predictive performance and generalization in AI-based pluvial flood modeling. A physically calibrated hydrological–hydrodynamic model was employed to generate synthetic datasets with varied temporal lengths, input feature combinations (rainfall, infiltration, drainage), and rainfall-intensity distributions. A long short-term memory (LSTM) network, chosen for its widespread adoption and proven performance in hydrology, was applied as a representative benchmark to assess accuracy, computational cost, and bias under controlled conditions. Results identify: (1) a threshold effect of dataset length (~14,400 samples), beyond which performance gains plateau; (2) rainfall-intensity distribution as the dominant driver of generalization—training solely on light or extreme events induces systematic bias, whereas mixed-intensity datasets substantially enhance robustness; (3) ancillary features (infiltration and drainage) improve stability only when data are sufficiently abundant. These findings quantify trade-offs and pinpoint actionable design levers, offering general insights into dataset design for machine learning models in flood prediction and beyond. By clarifying critical dataset requirements, this study provides transferable guidance for building efficient and balanced datasets in hydrology and broader Earth system sciences.

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

Keywords: Urban pluvial flooding; Machine learning dataset design; Threshold effect; Generalization bias; Rainfall intensity distribution; Hydrological–hydrodynamic modeling

1. Introduction

Global climate change and rapid urbanization have escalated urban pluvial flooding into a global crisis, threatening public safety, mobility, and economic stability (Qi et al., 2021; Wilhelm et al., 2022; Won et al., 2022). High-density cities are particularly vulnerable, as short-duration, high-intensity rainfall events exhibit strong suddenness and spatial heterogeneity, introducing complex spatiotemporal nonlinearities into forecasting tasks (Zhang et al., 2016). Improving the accuracy and timeliness of urban flood prediction has therefore become a central concern in hydrological modeling and urban resilience research.

Physically based rainfall–runoff and flood propagation models—such as SWMM, HEC-RAS, and their coupled hydrological–hydrodynamic extensions—have substantially advanced the representation of drainage dynamics and inundation processes (Chen et al., 2016; Gomes et al., 2021; Chitwatkulsiri and Miyamoto, 2023). Techniques including GIS integration for estuarine systems (Cardoso et al., 2020), real-time storm warning from 2D models (Hofmann and Schüttrumpf, 2020), and bidirectional coupling strategies (Jamali et al., 2020; Barreiro et al., 2022) have enriched practical forecasting pathways. Nevertheless, these models remain highly dependent on detailed parameters such as pipe network topology and surface roughness (Fu et al., 2022) and demand laborious calibration (Liu et al., 2017; Hattermann et al., 2018; Her et al., 2019). Such requirements limit their transferability to real-time operations, especially in small and medium-sized cities where data scarcity and limited computational capacity are pervasive obstacles (Yang et al., 2020; Chen et al., 2023).

Parallel to these advances, data-driven approaches have gained prominence. Deep learning models have shown remarkable capability in capturing nonlinear hydrological processes and have been widely applied in rainfall–runoff and flood forecasting (Ahani et al., 2018; Pollard et al., 2018; Kim and Han, 2020). Among them, Long Short-Term Memory (LSTM) networks and their variants have become especially prominent in urban hydrology (Zhang et al., 2018; Abbasimehr and Paki, 2022; Zheng et al., 2024). Recent developments include ES-LSTM with exponential smoothing (Hayder et al., 2023), lightweight architectures via knowledge distillation (Ma et al., 2022), swarm-intelligence optimization (Mahmoodzadeh et al., 2022), attention-based mechanisms (Xu et al., 2022; Jhong et al., 2024; Li et al., 2025), and encoder–decoder frameworks that accelerate large-scale simulations (Wei et al., 2024; Ni et al., 2024). Hybrid physics–AI systems that correct the errors of conventional models further demonstrate the promise of deep learning in hydrology (Wenchuan et al., 2024; Zhou et al., 2023). Collectively, these advances confirm that algorithmic innovation continues apace. However, relatively little attention has been devoted to systematic analysis of dataset design—how the scale, feature composition, and distribution of events influence predictive performance and generalization.

Urban pluvial flood data differ fundamentally from riverine hydrological time series: they are often sparse, intermittent, and strongly constrained by monitoring infrastructure (Nearing et al., 2021; Liu et al., 2024). Scarcity, imbalance, and uncertainty in rainfall records have been shown to directly impair model generalization, with notable performance deterioration once data uncertainty exceeds certain thresholds (Dong et al., 2020; Huang et al., 2021; Ghaith et al., 2022; Chen et al., 2024). To alleviate these challenges, researchers have generated synthetic rainfall–runoff data using physically based models, and such augmentation has been shown to improve learning (Kilsdonk et al., 2022; Chen et al., 2023). Yet most existing efforts focus narrowly on enlarging dataset size or applying local corrections, while the fundamental questions of dataset construction—such as the optimal sequence length, the role of

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

feature combinations, and the influence of rainfall-intensity distribution—remain insufficiently addressed (Tikhamarine et al., 2020). Evidence indicates that dataset effects are nonlinear: improvements plateau beyond certain lengths (He et al., 2019; Śliwowski et al., 2023), diversity often outweighs sheer volume (Fang et al., 2019; Hou et al., 2022; Kratzert et al., 2021; Gupta, 2024), and feature combinations contribute only when supported by sufficient samples (Paz et al., 2018; Son et al., 2020; Wang and Ying, 2023). Moreover, rainfall intensity distribution strongly influences generalization, particularly for extremes (Zheng et al., 2024), yet systematic evaluation of this factor is largely absent.

This study therefore investigates how dataset length, feature composition, and rainfall-intensity distribution influence predictive performance and generalization in urban flood prediction. Using synthetic datasets generated from a calibrated hydrological–hydrodynamic model, we establish a controlled experimental framework to examine threshold effects, feature interactions, and distributional biases. While LSTM is employed as a representative benchmark due to its widespread use in hydrology, the insights obtained are intended to be transferable across a broad range of machine learning models. In doing so, the study provides a data-centric perspective on AI-based flood forecasting and offers guidance for designing efficient and balanced datasets in hydrology and Earth system sciences.

The remainder of this paper is organized as follows. Section 2 introduces the study design, including the hydrological–hydrodynamic model setup, synthetic dataset generation, and the benchmark ML framework. Section 3 presents the experimental results, focusing on dataset length, rainfall-intensity distribution, and feature composition. Section 4 discusses the underlying mechanisms, compares the findings with previous studies, and outlines broader implications. Finally, Section 5 summarizes the main conclusions and provides directions for future research.

2. Methodology

2.1 Technical Roadmap

This study proposes a structured technical roadmap to investigate how different dataset design strategies influence the predictive performance of deep learning models in urban pluvial flood scenarios, as illustrated in Figure 1.

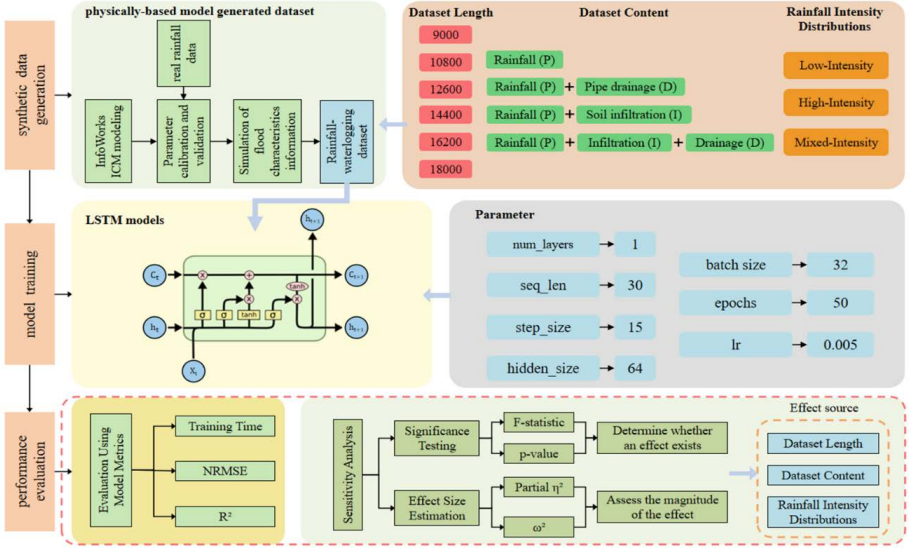


Figure 1: Technical roadmap for dataset generation, model training, and performance evaluation.

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

The methodology is organized into three sequential phases, encompassing synthetic data generation, model training, and performance evaluation.

The first phase involves the construction of rainfall–inundation time-series datasets using InfoWorks ICM 2021.4, a physically based hydrodynamic modeling platform. The model is pre-calibrated with observed urban waterlogging data, incorporating adjustments to parameters such as pipe roughness, surface flow coefficients, and land cover types to ensure simulation reliability. To represent a range of rainfall conditions, synthetic storm events are generated based on intensity-duration-frequency (IDF) curves for 1- to 10-year return periods. Each event spans 24 hours, followed by a 6-hour recession period to ensure temporal independence. By aggregating varying numbers of these events, datasets of different sequence lengths (ranging from 5 to 10 events) are constructed. Moreover, input features are organized into four distinct combinations—including rainfall, infiltration, and drainage flow—and rainfall intensities are stratified into light, heavy, and mixed categories to facilitate a multi-dimensional analysis of model behavior.

Next, model training is conducted under a standardized LSTM architecture to ensure consistency across experimental settings. The model employs a single-layer LSTM network with 64 hidden units, a 30-minute input sequence length, and a 15-minute step size, allowing it to effectively capture short-term rainfall–runoff dynamics. A 50% overlapping sliding window is applied during data segmentation to maximize feature retention and improve learning stability. All input features are normalized, and the model is trained for 50 epochs using a batch size of 32 and a learning rate of 0.005. The training is implemented in PyTorch and executed on an Intel i9 processor with an NVIDIA RTX 3090 GPU. Each dataset configuration is trained and validated independently using an 80:20 split, enabling comparative evaluation of convergence speed and stability under different data conditions.

Finally, a multi-metric evaluation scheme is employed to assess model performance from the perspectives of predictive accuracy, training efficiency, and generalization capacity. Predictive accuracy is quantified using the Normalized Root Mean Square Error (NRMSE), which accounts for scale sensitivity in runoff predictions. The coefficient of determination (R^2) is used to measure how well the model captures variance in the observed data. In addition, total training time is recorded to evaluate computational efficiency. To facilitate multi-objective comparison, radar charts are used to visualize performance across dataset configurations, revealing trade-offs and optimal strategies in a clear and interpretable manner.

2.2 Rainfall – Runoff Data Generation Method

Due to the high cost and limited spatiotemporal coverage of observed data for urban pluvial flood events, this study employs physically based simulation to generate controlled, high-fidelity datasets. A hydrological–hydrodynamic model is developed in InfoWorks ICM 2021.4, incorporating calibrated parameters such as pipe roughness, land cover types, and surface flow coefficients. The model is calibrated using measured inundation data, and its simulation performance is evaluated using the Nash–Sutcliffe Efficiency (NSE), with values consistently exceeding 0.5, indicating reliable accuracy for synthetic data generation.

The rainfall input used in simulation is derived from a regional design storm intensity–duration–frequency (IDF) formula, expressed in the following form:

$$q = \frac{a(1+b\times\log_{10}p)}{(t+c)^d} \quad (1)$$

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

where q is the rainfall intensity (L/s·ha), t is the rainfall duration (minutes), and p is the return period (years). The coefficients a, b, c, d are empirical constants fitted to local rainfall statistics.

Based on this formula, design rainfall events for 1-, 2-, ..., up to 10-year return periods are generated, each lasting 24 hours, with a peak factor set to 0.4 to simulate realistic storm profiles. Each event is followed by a 6-hour drainage period to ensure temporal independence. The generated rainfall curves are illustrated in Figure 2.

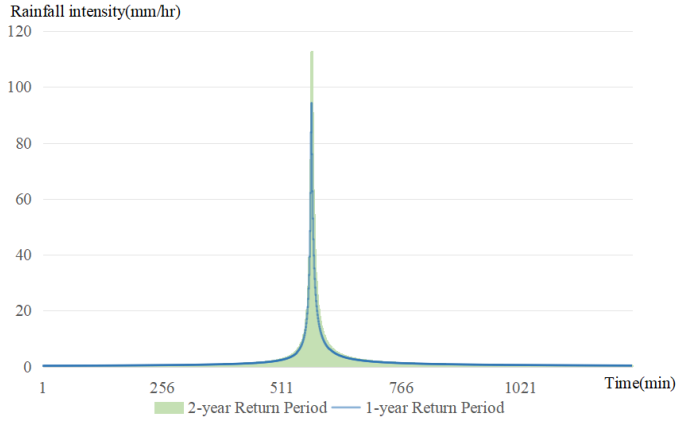


Figure 2: Rainfall curves generated from intensity-duration-frequency (IDF) formula for different return periods.

By progressively aggregating these events, time-series datasets of varying lengths are created. This allows for controlled evaluation of how data scale impacts model learning. The corresponding runoff response—represented as inundation area—is calculated using a simplified water balance model:

$$Q = P - I - D \tag{2}$$

where P is total precipitation, III is infiltration estimated via the Horton method, and D is the volume drained through the sewer system. The residual Q serves as the model's predictive target, representing surface water accumulation over time.

2.3 Dataset Design Strategies

This section outlines the design logic for constructing datasets across three key dimensions: input feature configurations, sequence length, and rainfall intensity distribution. These configurations form the foundation for a factorial experimental setup that enables systematic evaluation of model behavior under varying data conditions.

2.3.1 Input Feature Configurations

Accurate urban flood prediction requires capturing the complex interplay among rainfall generation, infiltration processes, and drainage dynamics. To represent these factors, we construct four types of input feature configurations:

1. Configuration 1: Rainfall (P) only
2. Configuration 2: Rainfall (P) + Pipe drainage (D)
3. Configuration 3: Rainfall (P) + Soil infiltration (I)
4. Configuration 4: Rainfall (P) + Infiltration (I) + Drainage (D)

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

These configurations reflect varying degrees of hydrological process coupling. Configuration 1 represents the minimal input case, relying solely on external forcing. Configurations 2 and 3 incorporate key physical subsystems—engineered drainage and soil infiltration—individually. Configuration 4 integrates both, offering the most physically complete scenario. This structure enables a comparative analysis of the effect of hydrological complexity on model performance and interpretability.

2.3.2 Sequence Length Design

To evaluate the sensitivity of LSTM model performance to time-series span, six different sequence lengths are defined, each composed of multiple synthetic rainfall events. A single rainfall event consists of 24 hours of precipitation followed by a 6-hour recession period. The training sets are then constructed by stacking the following sequence lengths:

1. Length 1: 9000 samples
2. Length 2: 10800 samples
3. Length 3: 12600 samples
4. Length 4: 14400 samples
5. Length 5: 16200 samples
6. Length 6: 18000 samples

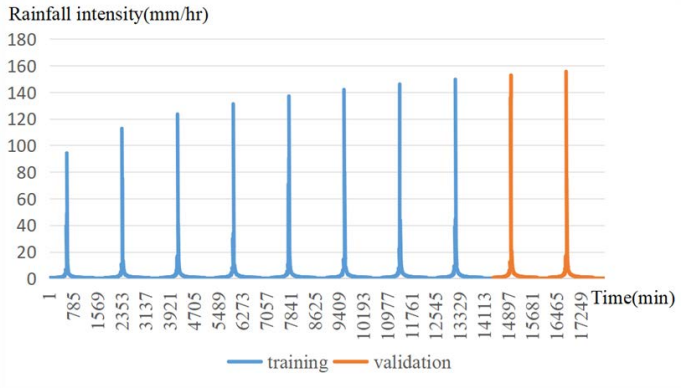
All datasets are split into training, validation, and test sets with a fixed ratio of 6:2:2. This design allows us to investigate how the amount of temporal information affects learning efficiency, model stability, and prediction accuracy under different data scales.

2.3.3 Rainfall Intensity Distributions

This experimental dimension is designed to investigate whether the inclusion of extreme rainfall events in the training dataset is essential for achieving accurate flood predictions, and whether representative samples selected from the broader rainfall spectrum can support robust model generalization. It also facilitates the examination of potential predictive biases—such as the tendency of models trained predominantly on heavy rainfall to systematically overestimate inundation levels.

To explore these aspects, three distinct rainfall classification schemes are defined based on intensity distribution within the training set:

1. **Low-Intensity Training Set:** Includes only low-intensity rainfall events in the training set, while validation contains extreme rainfall scenarios, as shown in Figure 3.



https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

Figure 3: Rainfall intensity classification for low-intensity training set.

2. **High-Intensity Training Set:** Contains primarily high-intensity rainfall events in the training set, with validation samples representing mild to moderate scenarios, as illustrated in Figure 4.

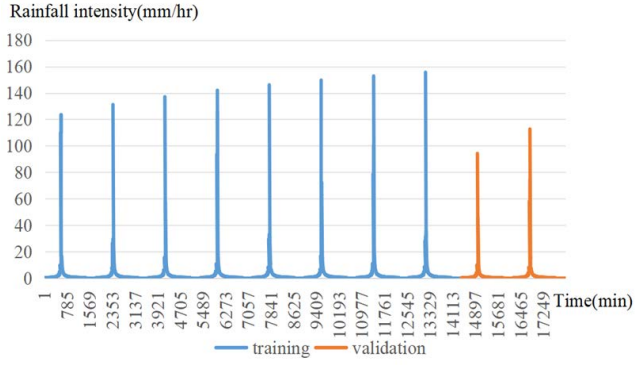


Figure 4: Rainfall intensity classification for high-intensity training set.

3. **Mixed-Intensity Training Set:** Combines low and high rainfall events in the training set and includes moderate events in validation, as presented in Figure 5.

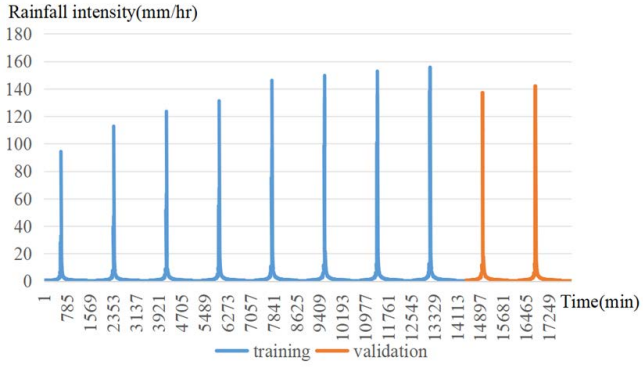


Figure 5: Rainfall intensity classification for mixed-intensity training set.

2.4 LSTM Model Configuration and Parameters

A unified LSTM neural network structure is adopted in this study to ensure consistency across experiments. The model is designed with a sequence length of 30 minutes and a step size of 15 minutes, enabling it to effectively capture the dynamic response cycle of short-duration rainfall–runoff processes. To enhance data utilization and reduce the risk of missing critical hydrological features, a 50% overlapping sliding window strategy is applied.

The network architecture consists of a single-layer LSTM with 64 hidden units, providing a balance between model complexity and computational efficiency—particularly suited for extracting temporal patterns from minute-

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

level hydrological data. During training, the batch size is set to 32 and the learning rate to 0.005, ensuring stable gradient updates while allowing the model to converge fully within 50 epochs. This helps mitigate overfitting, especially when training on limited datasets.

The model is implemented using the PyTorch framework and trained on a computing environment equipped with an Intel i9 processor and NVIDIA RTX 3090 GPU. Detailed model parameters are listed in Table 1.

Table 1: LSTM Model Configuration and Parameters.

Parameter	Value
Model layers (num_layers)	1
Sequence length (seq_len)	30
Step size (step_size)	15
Number of neurons (hidden_size)	64
Batch size	32
Number of epochs (epochs)	50
Learning rate (lr)	0.005

2.5 Evaluation Metrics

To comprehensively evaluate how different dataset construction strategies affect model performance, this study adopts three representative metrics: prediction accuracy, explanatory power, and computational efficiency. These are detailed as follows:

2.5.1 Normalized Root Mean Square Error (NRMSE)

NRMSE quantifies relative prediction error and is suitable for comparing performance across datasets with varying rainfall magnitudes. It is defined as:

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \tilde{Y}_i)^2}}{\overline{X}} \quad (3)$$

where n is the number of data points, Y_i is the observed value, $\tilde{Y}_i$ is the predicted value, and $\overline{X}$ is the mean of observed values. Its normalized nature makes it robust when comparing models trained on light versus heavy rainfall samples.

2.5.2 Coefficient of Determination (R^2)

R^2 reflects the proportion of variance in observed data that is explained by the model. It is a widely used indicator of model fitting quality and is defined as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (4)$$

where SS_{res} is the sum of squared residuals and SS_{tot} is the total sum of squares. A value closer to 1 indicates better alignment between predicted and actual water depth values.

2.5.3 Training Time

Training time, measured in seconds, serves as a proxy for model efficiency. Given the identical hardware and software environments across all experiments, this metric offers a fair basis to compare convergence behavior and computational overhead across datasets with different sequence lengths and feature complexities.

2.6 Sensitivity Analysis

To further investigate how different dataset construction dimensions influence model performance, this study applies Multivariate Analysis of Variance (MANOVA) to assess the statistical significance and effect strength of three key factors: sequence length, input feature combination, and rainfall intensity level. Unlike single-factor tests,

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

MANOVA enables a systematic assessment of main effects and interaction effects under multi-condition experiments, offering quantitative insights into the extent to which each factor contributes to prediction variability.

This section consists of two components: significance testing, which determines whether a factor has a statistically significant influence on model metrics (e.g., NRMSE and R^2), and effect size evaluation, which quantifies the magnitude of such influence and guides future dataset design strategies.

2.6.1 Significance Testing: F-statistic and p-value

At the core of variance analysis is the F-statistic, which compares the variance between groups to the variance within groups to determine the presence of significant effects (Becher et al., 2025). It is calculated as:

$$F = \frac{MS_{effect}}{MS_{error}} \tag{5}$$

where MS_{effect} and MS_{error} are the mean squares of the effect and residual error, respectively. These are derived by dividing the corresponding sum of squares (*SS*) by their degrees of freedom (*df*). The total sum of squares (SS_{total}) captures the overall variance from the global mean, while the error sum of squares (SS_{error}) represents unexplained random variation. Factor-specific sums of squares—such as SS_{length}, $SS_{combination}$, and $SS_{rainfall}$—capture variance uniquely attributed to each design factor.

Interaction terms (e.g., length × combination, length × rainfall, combination × rainfall, and the three-way interaction) are also included to assess whether the combined influence of multiple factors significantly affects the model's behavior. Each term's degrees of freedom are defined based on the factorial structure of the experiment. The significance of each effect is determined by computing the corresponding p-value using the F-distribution:

$$p = P(F \geq F_{observed} \mid H_0) \tag{6}$$

where $F_{observed}$ is the computed F-statistic, and H_0 denotes the null hypothesis of no effect.

2.6.2 Effect Size Estimation: Partial η² and ω²

While significance testing reveals whether an effect exists, it does not indicate how substantial that effect is. To address this limitation, the analysis incorporates two effect size measures: partial η² and ω², which together offer a more nuanced interpretation.

Partial η² represents the proportion of variance in the dependent variable uniquely explained by a factor, relative to the unexplained variance. It is computed as:

$$\eta^2 = \frac{SS_{effect}}{SS_{effect}+SS_{error}} \tag{7}$$

In contrast, ω² introduces a degrees-of-freedom correction, making it more robust for small samples or multi-factor models [66]. Its formula is:

$$\omega^2 = \frac{SS_{effect}-df_{effect}\times MS_{error}}{SS_{total}+MS_{error}} \tag{8}$$

These two metrics complement each other: partial η² provides a direct interpretation of explained variance, while ω² offers a more conservative estimate of generalizable effect strength. Together, they enhance the interpretability of MANOVA results and support evidence-based model optimization decisions.

3. Dataset Generation and Experimental Setup

The organization of training data is a critical factor influencing the performance and generalization capability of LSTM-based models. In response to the complexities of urban pluvial flood forecasting, this section presents the construction of multiple structurally diverse datasets, designed along three key dimensions: input feature combinations,

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

time series length, and rainfall intensity distribution. These datasets form the basis for the systematic experiments presented in Section 4, enabling a controlled analysis of how each design factor affects model behavior. All data are derived from physically validated flood simulations to ensure scientific rigor and experimental reliability.

3.1 Study Area and Model Validation

This case study investigates a representative residential neighborhood located in a plain city in China. The study area covers approximately 6,500 m², with the land surface predominantly occupied by buildings and grassland; buildings account for 44.14% of the total area. Stormwater is conveyed through a municipal drainage network that is densely distributed in low-lying zones. The system consists of uniformly spaced inspection wells and standardized pipeline structures, which reflect the typical configuration of urban stormwater infrastructure. An overview of the study area is shown in Figure 6.

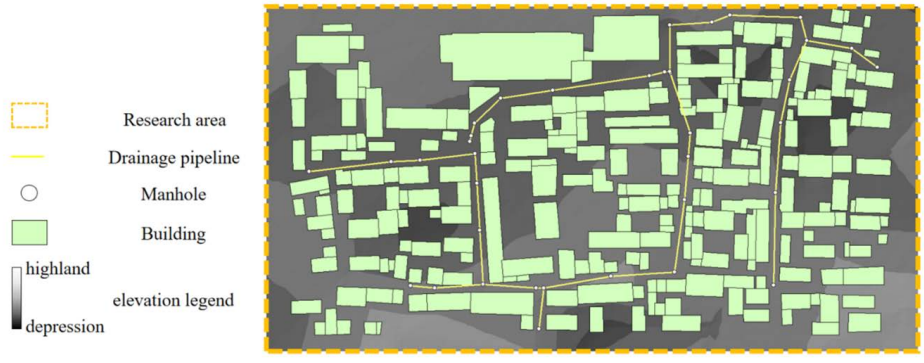


Figure 6: Overview of the study area: a typical residential neighborhood in a plain city.

A physics-based urban flood model was developed using InfoWorks ICM, incorporating detailed representations of surface properties, sewer topology, and boundary conditions. Key parameters such as surface roughness, slope, and initial losses were calibrated using field observations. Model validation was performed using recorded rainfall events to assess predictive reliability. As illustrated in Figure 7, the simulated inundation process aligns closely with observed data in terms of peak water level, total runoff volume, and temporal response. The calculated Nash–Sutcliffe Efficiency (NSE) exceeds 0.5, a widely accepted threshold for reliable hydrological simulations, confirming that the model exhibits adequate accuracy and generalizability.

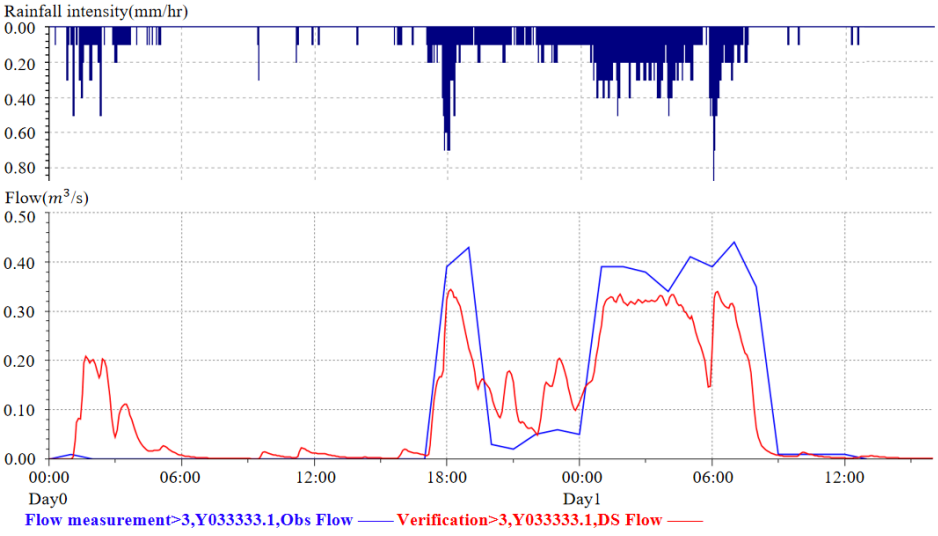
These results support the use of the validated model as a reliable data generation engine for constructing high-quality training datasets used in the subsequent deep learning experiments.

3.2 Dataset Construction Based on Feature Combinations

To evaluate how different types of physical information affect model performance, this study designs four input configurations based on three core hydrological variables involved in urban flood processes:

1. P: Rainfall intensity (external driving force)
2. I: Soil infiltration volume (controls surface retention and loss)
3. D: Pipe drainage flow (reflects internal drainage capacity and network response)

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025



	Obs Flow			DS Flow		
Flow measurement>3,Cond3026	Min(m^3/s)	Min(m^3/s)	Min(m^3/s)	Min(m^3/s)	Min(m^3/s)	Min(m^3/s)
Verification>3,Y033333.1	0.000	0.440	15012.000			
				0.000	0.345	13662.832

Figure 7: Simulated inundation process aligning with observed data for validation.

The prediction target is the inundation area (*Y*). Each dataset represents a different level of input complexity and physical completeness. The combinations are detailed in Table 2.

Table 2: Different Input Feature Combinations Design.

Dataset Configuration	Dataset Content	Dataset Width
Configuration 1	Rainfall (P) → Inundation Area (Y)	1
Configuration 2	Rainfall (P) + Pipe Drainage (D) → Inundation Area (Y)	2
Configuration 3	Rainfall (P) + Soil Infiltration (I) → Inundation Area (Y)	2
Configuration 4	Rainfall (P) + Soil Infiltration (I) + Pipe Drainage (D) → Inundation Area (Y)	3

These configurations form a progressively enriched feature space, allowing us to investigate the impact of additional physical information on both training efficiency and predictive accuracy.

Taking Combination 4 as an example, the input and output data at time step *t* are structured as:

$$\mathbf{Input} = \left\{P^{(t)}, I^{(t)}, D^{(t)}\right\} \rightarrow Y^{(t)} \tag{9}$$

The corresponding LSTM prediction structure is formulated as:

$$\boldsymbol{Y^{(t)}} = \boldsymbol{sigmoid}\left(\mathbf{w}\left[\boldsymbol{P^{(t)}}, \boldsymbol{I^{(t)}}, \boldsymbol{D^{(t)}}, \boldsymbol{Y^{(t-1)}}\right] + \mathbf{b}\right) \tag{10}$$

where $Y^{(t)}$ denotes the predicted inundation area at time step *t*; $P^{(t)}$, $I^{(t)}$, and $D^{(t)}$ represent rainfall intensity, soil infiltration, and pipe drainage flow at time *t*, respectively; $Y^{(t-1)}$ is the previous-step output fed back into the model to capture temporal dependency; sigmoid denotes the nonlinear activation function; **w** is the weight matrix, and **b** is the bias term.

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

3.3 Dataset Construction with Varying Sequence Lengths

To assess the effect of temporal sequence length on model learning capacity and convergence behavior, this study constructs six datasets of different durations by incrementally aggregating design storm events. Each rainfall event is generated based on return periods ranging from 1 to 10 years and includes 24 hours of rainfall followed by a 6-hour recession phase, totaling 30 hours per event. This ensures that each sample captures the full flood response process, from initiation to dissipation.

Longer datasets are constructed by sequentially stacking multiple storm events, thereby simulating varying levels of historical data availability. This approach reflects practical scenarios where models are trained on datasets of different completeness depending on data collection infrastructure or simulation budget. The sample size associated with each dataset is summarized in Table 3.

Table 3: Dataset Time Length Settings.

Dataset Length	Number of Samples
Length 1	9000
Length 2	10800
Length 3	12600
Length 4	14400
Length 5	16200
Length 6	18000

All datasets are processed using a 15-minute time step and overlapping sliding windows to improve data efficiency. The total samples reflect both the number of storm events and the internal segmentation strategy. Each dataset is then split into training, validation, and test subsets in a 6:2:2 ratio to ensure consistent model evaluation.

This design enables systematic investigation into the trade-offs between dataset length and model performance. Longer sequences offer more comprehensive temporal information, potentially enhancing the model's ability to capture long-range dependencies. However, excessive length may introduce noise or redundancy and increase training cost. Understanding this balance is critical for optimizing LSTM applications in urban flood forecasting, especially under data-limited conditions.

3.4 Model Performance under Different Rainfall Intensity Compositions

Building on the rainfall intensity classification described in Section 2.3.3 (see Figures 3–5), this section analyzes how varying the composition of rainfall intensities in the training dataset influences model generalization, particularly in unseen conditions.

Three dataset configurations were used to represent distinct training regimes:

1. Light-rain training, which includes only low-intensity rainfall events;
2. Heavy-rain training, primarily composed of extreme storms; and
3. Mixed-rain training, combining both low and high rainfall intensities.

All three setups were evaluated using a common validation set to ensure comparability.

Experimental results reveal distinct behavioral patterns. Models trained exclusively on low-intensity rainfall exhibit high accuracy under frequent, mild conditions but tend to underestimate peak inundation during rare events. In contrast, heavy-rain-trained models demonstrate strong performance under extreme rainfall but frequently overpredict flooding when tested on moderate or low-intensity scenarios. The mixed-rain training set strikes a balance, achieving stable performance across the spectrum of rainfall intensities and minimizing both overestimation and underestimation tendencies.

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

These results underscore the importance of training data diversity. Overexposure to one rainfall category can result in systematic predictive bias. In urban flood forecasting—where both frequent and extreme events carry practical importance—balanced or strategically composed training sets are crucial for robust model generalization.

4. Model Results and Comparative Analysis

This section presents a comprehensive evaluation of how different dataset configurations affect the predictive performance of the LSTM model in urban flood forecasting. The analysis focuses on three core variables: input feature combinations, dataset length, and rainfall intensity distribution. These factors were selected for their practical relevance in real-world flood scenarios and their potential to influence both model generalization and training efficiency.

To ensure consistency across experiments, all models were trained using the same network architecture and hyperparameters, as described in Section 2.4. Model performance was evaluated using three metrics introduced in Section 2.5: Normalized Root Mean Square Error (NRMSE), coefficient of determination (R^2), and training time. In addition to comparing predictive accuracy, this section includes generalization tests and multifactor sensitivity analysis to assess the robustness of model performance under unseen or variable conditions. Together, these analyses aim to provide insights into how data design choices affect the stability, effectiveness, and computational efficiency of LSTM-based flood forecasting models.

4.1 Evaluation Using Model Metrics

4.1.1 Training Time Analysis

To quantify how different dataset configurations influence training efficiency, we begin by analyzing the model training time. As shown in Figure 8, training durations exhibit a clear stepwise increase with longer dataset lengths. The average training times for the six sequence lengths are: 437.972, 525.75, 627.055, 717.861, 804.555, and 896.444 seconds. The relative increases between each level are 20.04%, 19.27%, 14.48%, 12.08%, and 11.42%, respectively.

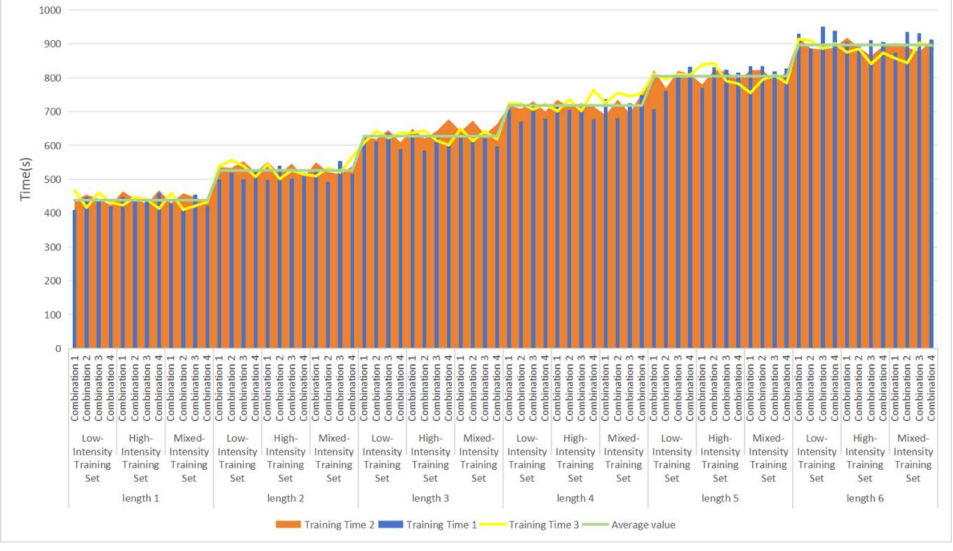


Figure 8: Model training time across different dataset lengths.

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

This trend demonstrates that increased data volume significantly impacts computational cost, although the rate of increase tapers off as datasets grow. In contrast, differences in rainfall intensity distribution and input feature combinations have comparatively minor effects on training time, suggesting that data volume—rather than data diversity—is the primary factor influencing training efficiency.

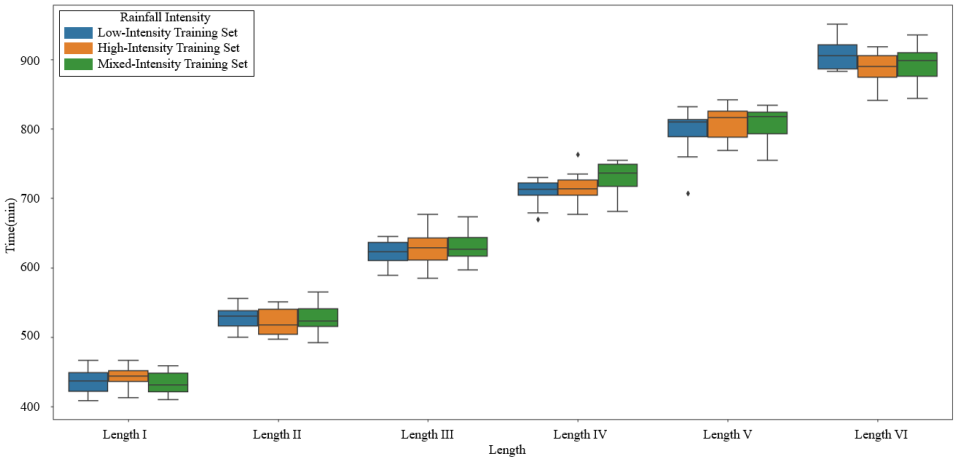


Figure 9: Comparison of training durations across different feature combinations.

Figures 9 and 10 further illustrate training time trends from two perspectives. Figure 10 compares training durations across feature combinations under the same rainfall category. While training time increases by over 100% from the shortest to longest datasets, the variation between feature combinations remains moderate (e.g., from 50 to 200 seconds). Mixed-intensity training sets consistently show the lowest standard deviation (8–15% lower than others), indicating that rainfall diversity improves training stability.

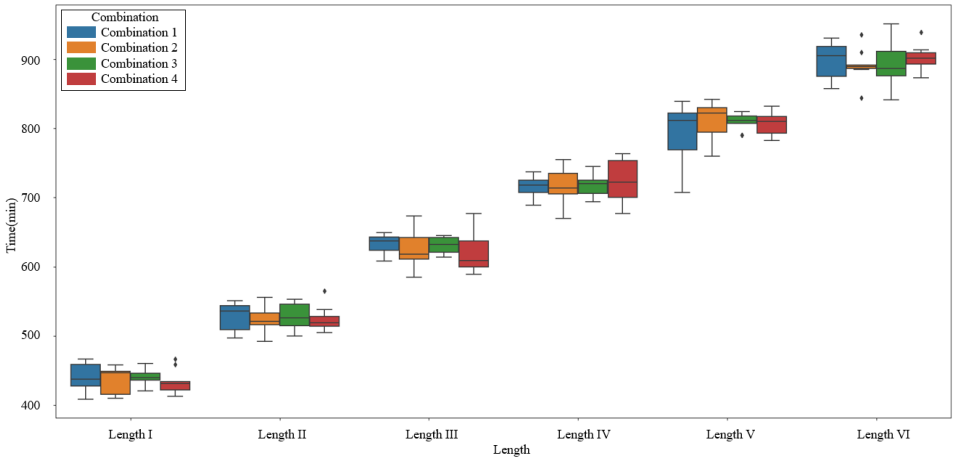


Figure 10: Training time comparisons across rainfall structures within each feature combination.

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

Figure 10 presents training time comparisons across rainfall structures within each feature combination. All combinations maintain stable scalability, confirming their robustness under complex and heterogeneous data configurations.

4.1.2 NRMSE Analysis

Figure 11 presents the normalized root mean square error (NRMSE) values for all dataset variants. Results indicate that model performance is significantly influenced by dataset length, rainfall intensity category, and feature combination.

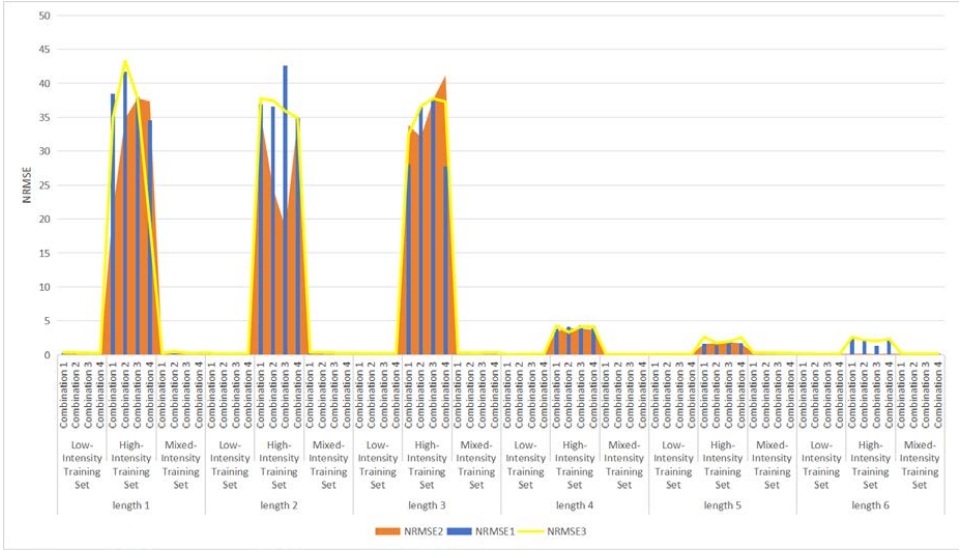


Figure 11: NRMSE values across different dataset configurations.

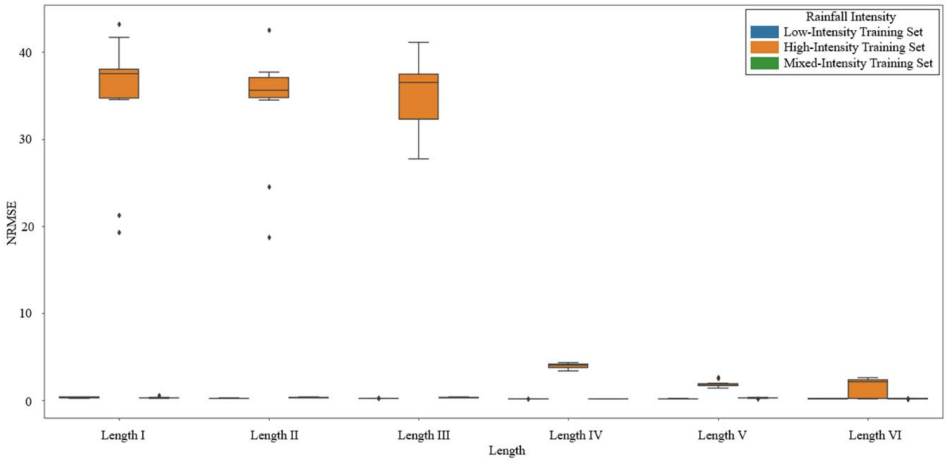


Figure 12: NRMSE for different rainfall intensity categories.

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

From the rainfall intensity perspective (Figure 12), the low-intensity and mixed-intensity training sets exhibit marked reductions in NRMSE when the dataset length reaches Level 4. For instance, the NRMSE for the low-intensity dataset decreases by more than 50% and stabilizes thereafter. In contrast, the high-intensity dataset performs poorly under short sequences (e.g., NRMSE of 41.689 for Combination 2 at Length 1), followed by sharp improvement beyond Length 4, with values dropping to between 1.6 and 4.3. This highlights the importance of data volume for accurately capturing extreme rainfall behavior.

The mixed-intensity datasets demonstrate strong generalization across scales, maintaining minimal NRMSE variation across combinations (e.g., a range of just 0.028 at Length 4), suggesting that diversified rainfall input enhances model stability.

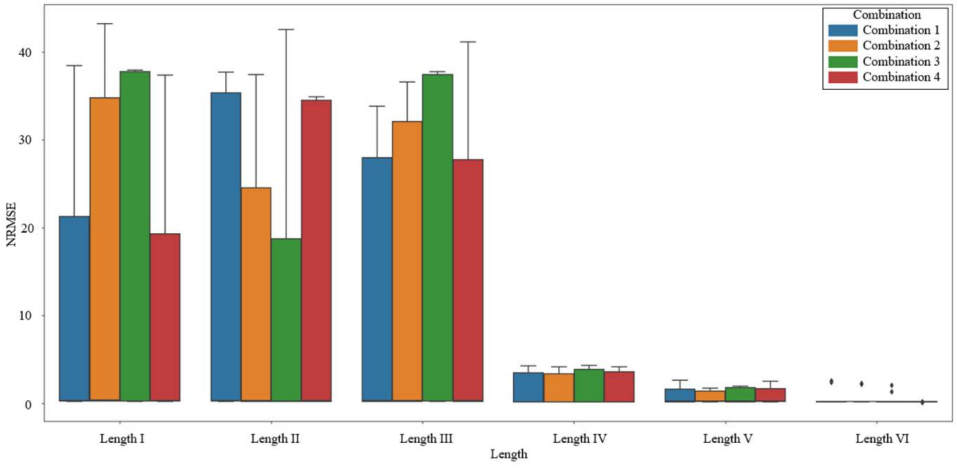


Figure 13: NRMSE for different feature combinations.

From the perspective of feature combinations (Figure 13), the high-intensity group exhibits substantial variability under shorter sequences (e.g., an inter-combination range of 23.945 at Length 1). This variance diminishes with longer datasets, and at Length 6, Combination 3 achieves an exceptionally low NRMSE of 0.139.

Combination 4 consistently performs well across all rainfall categories. For low-intensity rainfall at Length 4, it achieves an average NRMSE of 0.161 (within 6.3% of the best-performing setting); for high-intensity rainfall at Length 5, the NRMSE is 2.307 (14.6% deviation); and for mixed-intensity rainfall at Length 6, the NRMSE is 0.173 (4.8% deviation). Notably, Combination 4 achieves the overall best performance of 0.117 under the mixed-intensity configuration at Length 6.

Some nonlinear anomalies are also observed. For instance, in the mixed-intensity group, the NRMSE drops abruptly from 0.361 to the 0.132–0.163 range at Length 4, followed by unexpected fluctuations at Length 5. The underlying causes of this nonlinearity remain unclear and require further investigation through controlled experiments.

4.1.3 R^2 Analysis

Figure 14 shows the coefficient of determination (R^2) values for the LSTM model across different dataset configurations. The low-intensity and mixed-intensity training sets perform well even with short sequences, achieving R^2 values above 0.95 at Length 1. As the dataset length increases, their performance stabilizes further—for example, the mixed-intensity training set reaches an R^2 of 0.992 at Length 4.

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

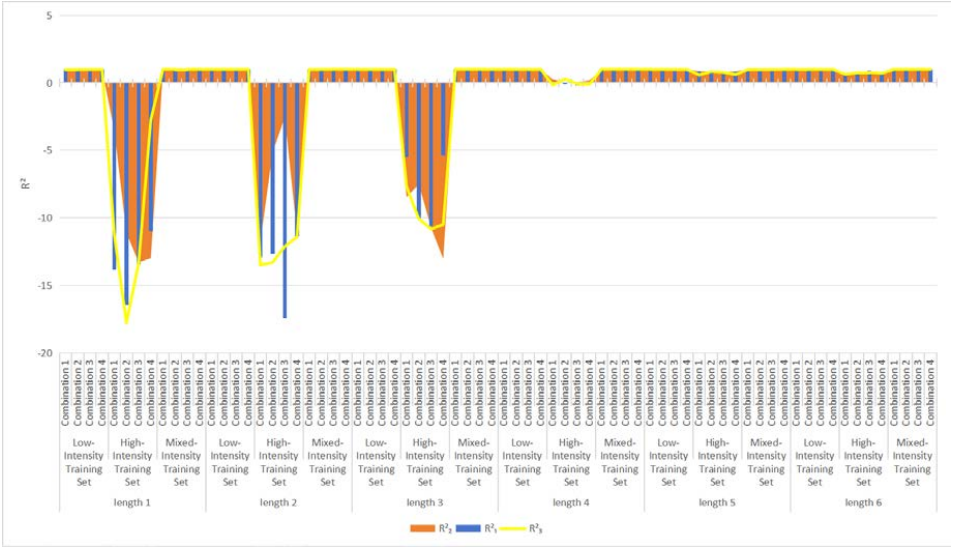


Figure 14: Coefficient of determination (R^2) values for LSTM models across configurations.

In contrast, the high-intensity training set exhibits a two-stage behavior. When trained on short sequences, the model fails to generalize, with R^2 values falling into the negative range, such as –17.758 at Length 1. However, once the dataset length exceeds Level 4, model performance improves sharply, with R^2 values rising to between 0.071 and 0.289, and continuing upward to between 0.6 and 0.893 at Lengths 5 and 6. These results emphasize the critical role of sufficient training volume in enabling the model to learn from extreme rainfall events.

Further analysis from two perspectives—rainfall intensity (Figure 15) and input feature combinations (Figure 16)—reveals more nuanced trends.

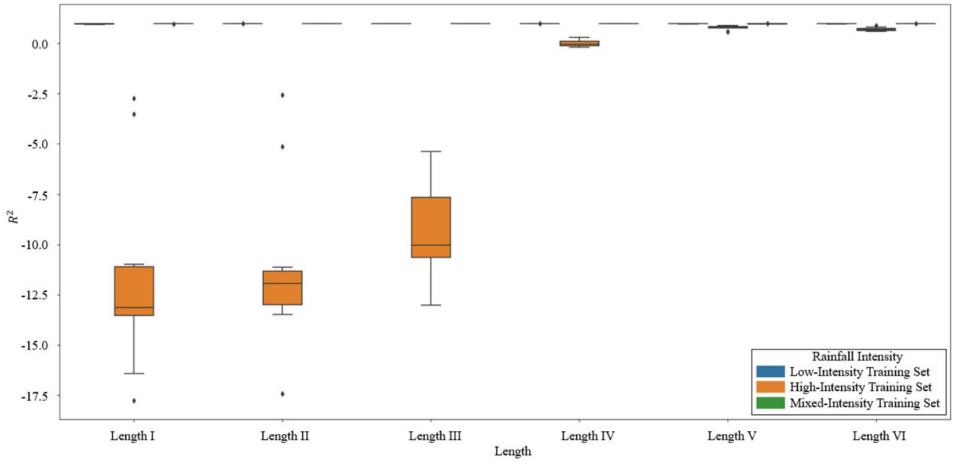


Figure 15: R^2 values for low-intensity rainfall scenarios.

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

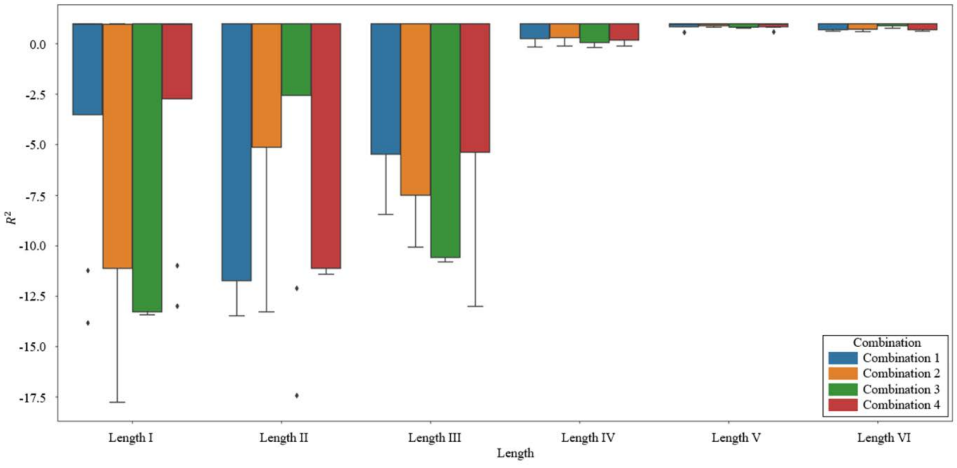


Figure 16: R^2 values for high-intensity rainfall scenarios.

Figure 15 illustrates that in low-intensity scenarios, all feature combinations consistently achieve high R^2 values across different sequence lengths. The highest performance is observed at Length 4 using Combination 1, where the R^2 reaches 0.992. The difference between combinations also narrows significantly as dataset length increases, with the inter-combination range shrinking from 0.036 at Length 1 to just 0.015 at Length 4. This indicates that the model effectively captures low-intensity flood dynamics regardless of the feature configuration.

Figure 16 shows that in high-intensity scenarios, short datasets lead to unstable and inaccurate predictions. Most R^2 values at Length 1 are negative—for example, Combination 2 records –16.424—indicating that the model fails to learn meaningful patterns from limited data in the presence of extreme variability. From Length 4 onward, however, R^2 values improve significantly. For instance, the average R^2 for Combination 4 rises from –9.635 at Length 3 to values between –0.189 and 0.289 at Length 4, and continues to increase to between 0.572 and 0.876 at Length 5. At Length 6, Combination 3 reaches a peak R^2 of 0.893, which represents an improvement of over 1100% compared to its Length 1 performance.

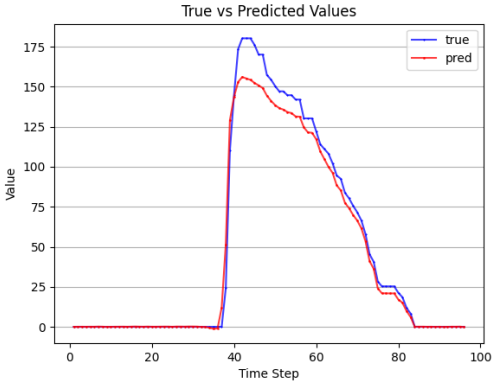
The mixed-intensity training set once again demonstrates the most stable and robust results. At Length 4, the R^2 range across all combinations is just 0.028—significantly lower than that of the low-intensity set (0.036) and the high-intensity set (0.828) at the same length. Combination 4, in particular, exhibits cross-scenario robustness. It consistently approaches or exceeds optimal performance in all conditions: R^2 values are near 0.893 for mixed-intensity at Length 6, while maintaining competitive accuracy for low- and high-intensity cases at Lengths 4 and 5.

Although these patterns highlight the advantages of longer datasets and diverse rainfall conditions, the mechanisms behind the sharp improvements in the high-intensity training set remain insufficiently understood. Further controlled studies are needed to quantify the relationship between data volume and model performance, and to determine the threshold at which model generalization behavior transitions from failure to success.

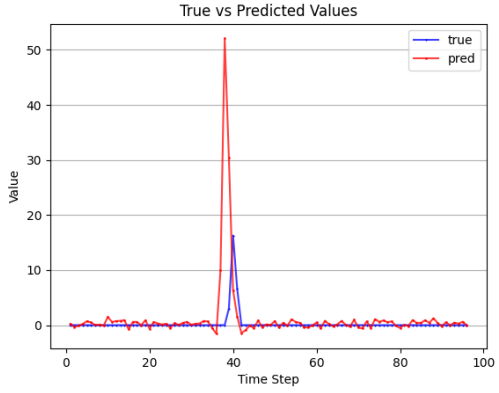
4.2 Impact of Rainfall Intensity on Generalization

To investigate how different rainfall intensity distributions in the training data affect model generalization, three training strategies were designed with a fixed total dataset size: low-intensity, high-intensity, and mixed-intensity training sets. The corresponding model performances are illustrated in Figure 17.

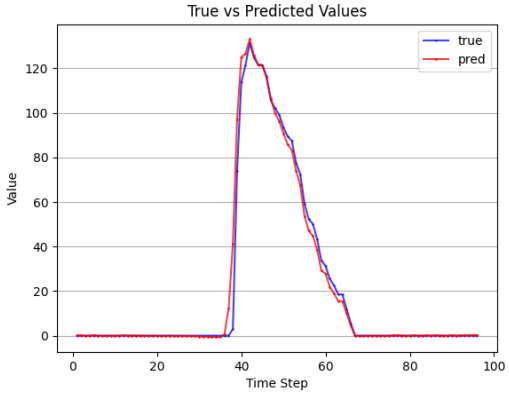
https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025



(a)Low-Intensity Training Set



(b)High-Intensity Training Set



(c)Mixed-Intensity Training Set

Figure 17: Impact of rainfall intensity distribution on model generalization.

Figure 17 clearly demonstrates the following patterns:

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

1. When the training set includes only low-intensity rainfall events, the model consistently underestimates inundation levels during high-intensity rainfall scenarios.
2. Conversely, when the training set includes only high-intensity rainfall events, the model tends to overestimate flooding under low-intensity conditions.
3. The mixed-intensity training set substantially improves model fitting across moderate rainfall test samples, achieving the best overall generalization performance.

These results underscore the importance of incorporating a diverse range of rainfall intensities during model training. A lack of variability in the training set—particularly the exclusion of either low or high extremes—can introduce significant predictive bias. Therefore, it is recommended that training datasets include representative samples spanning multiple rainfall categories to enhance the model's robustness and adaptability across different hydrological conditions.

4.3 Sensitivity Analysis of Evaluation Metrics

This section employs Multi-factor Analysis of Variance (ANOVA) to systematically evaluate the influence of three independent variables—dataset length, rainfall intensity distribution, and input feature combination—on key model performance metrics: training time, normalized root mean square error (NRMSE), and coefficient of determination (R^2). The goal is to quantify both the individual effects and interaction effects of these factors to inform data design and model optimization strategies.

4.3.1 Sensitivity of Training Time to Input Variables

Figure 18 summarizes the ANOVA results for the effects of dataset length, rainfall intensity, and feature combination on model training time.

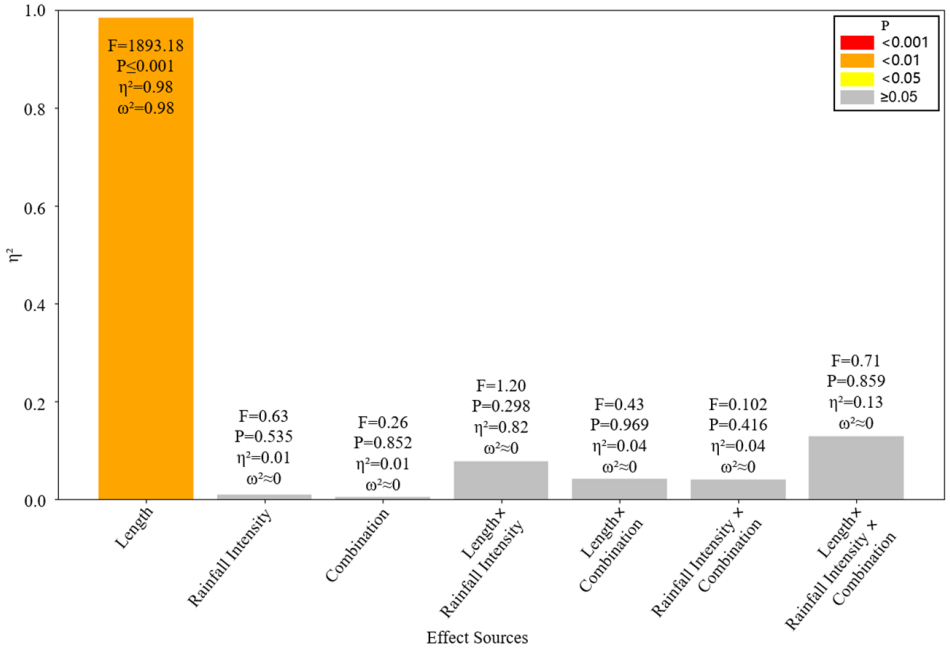


Figure 18: ANOVA Results for Training Time Sensitivity.

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

The results indicate that dataset length is the dominant factor affecting training time, with extremely high statistical significance ($F = 1893.18$, $p < 0.001$) and a very large effect size ($\eta^2 = 0.985$). This implies that increased sequence length directly drives higher computational costs. In contrast, rainfall intensity and input feature combination have negligible effects on training duration, with very low η^2 and non-significant p-values ($p > 0.5$), suggesting that their impact on computational efficiency can be reasonably ignored under the current experimental setup.

4.3.2 Sensitivity of NRMSE to Input Variables

Figure 19 presents the sensitivity results for NRMSE across the three experimental factors.

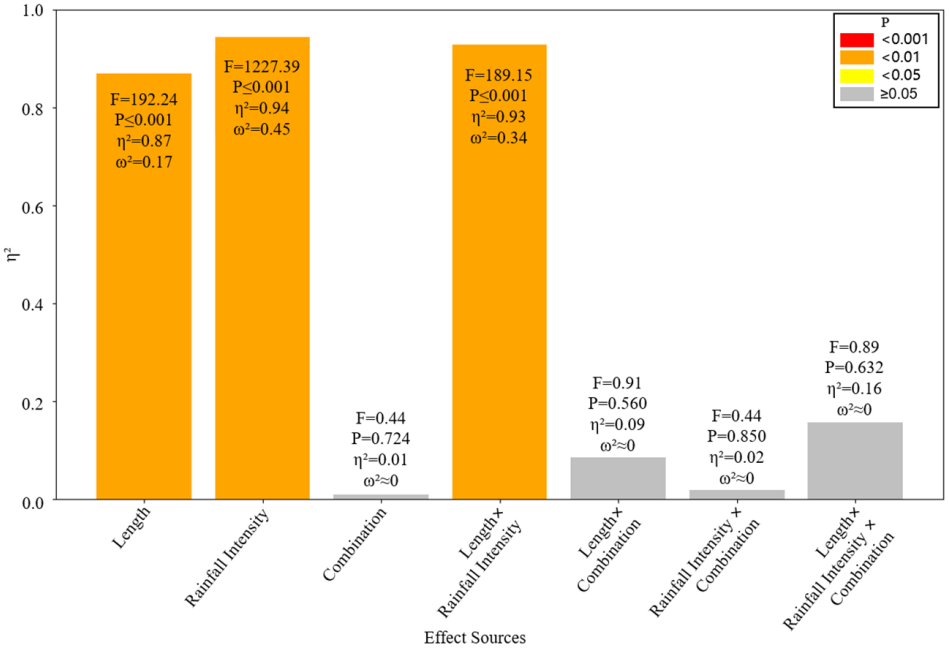


Figure 19: Sensitivity of NRMSE to Input Variables.

Both rainfall intensity ($\eta^2 = 0.945$) and dataset length ($\eta^2 = 0.870$) show strong influence on model error. Among them, rainfall intensity contributes more significantly to the variance in NRMSE, as reflected by its higher adjusted effect size ($\omega^2 = 0.446$ vs. 0.174). Their interaction is also statistically significant ($F = 189.15$, $p < 0.001$), suggesting that the performance impact of dataset length varies considerably depending on rainfall conditions. Conversely, feature combination has no statistically meaningful impact ($p > 0.7$, $\eta^2 < 0.02$), with negligible contribution to prediction error under current configurations.

4.3.3 Sensitivity of R^2 to Input Variables

Figure 20 reports the ANOVA results for the coefficient of determination (R^2). Rainfall intensity again emerges as the most influential factor on model fit, with a high F value ($F = 381.26$, $p < 0.001$) and the largest effect size ($\omega^2 = 0.391$). Dataset length is also significant ($F = 65.96$, $\eta^2 = 0.696$), indicating that longer time series enhance the model's explanatory power. Their interaction is notably significant as well ($F = 65.98$, $\eta^2 = 0.821$), demonstrating that rainfall conditions can amplify the sensitivity of R^2 to sequence length. In contrast, feature combination remains statistically irrelevant, with consistently low F values and near-zero effect sizes across all interaction terms.

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

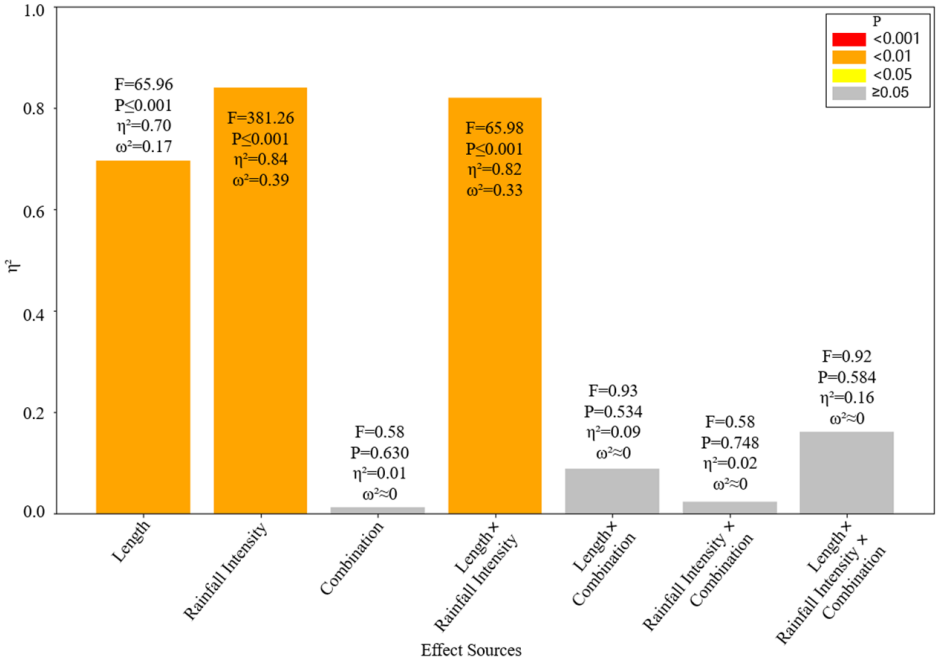


Figure 20: Sensitivity of R^2 to Input Variables

5. Discussion

The results of this study demonstrate that the way datasets are constructed plays a decisive role in determining the performance and generalization of machine learning models for urban flood forecasting. Three dimensions in particular—dataset length, rainfall-intensity distribution, and feature composition—emerged as critical levers shaping predictive outcomes. Although LSTM was adopted as the benchmark model due to its prominence in hydrological applications, the observed patterns are not limited to a single algorithm but reflect more general properties of data-driven modeling.

The effect of dataset length followed a nonlinear trajectory. Expanding the number of samples initially produced significant improvements in predictive accuracy, but beyond approximately 14,400 sequences, the gains plateaued. This saturation indicates that once the essential temporal variability is captured, additional data primarily reinforce already-learned dynamics rather than introduce new information. The fluctuations observed at certain lengths further suggest that the interaction between sequence size and rainfall variability can create instability, reflecting overfitting to recurrent patterns. For practical applications, this implies that indiscriminately increasing dataset size is not always efficient. More effective strategies may include pre-training, transfer learning across basins, or adaptive sequence-length adjustment, which can yield comparable improvements while reducing computational cost.

Rainfall-intensity distribution proved to be the dominant factor governing generalization. Models trained on imbalanced datasets—whether dominated by light or extreme rainfall—exhibited systematic biases, underestimating peaks or exaggerating minor events depending on the skew. In contrast, datasets that incorporated a balanced mix of intensities consistently delivered more robust predictions across diverse scenarios. This highlights the necessity of representativeness in dataset design: rare but high-impact rainfall events cannot be ignored or treated as statistical

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

outliers. Deliberate stratification, targeted augmentation of extremes, or probabilistic weighting may be required to ensure sufficient coverage of critical events, particularly in regions with limited observational records.

The contribution of additional hydrological features was more conditional. Incorporating infiltration and drainage information enhanced model stability and reduced variance when data availability was adequate, but in smaller datasets, these inputs increased the risk of overfitting. This outcome reflects the trade-off between richer input dimensionality and the statistical support required to train it. Effective feature engineering should therefore be closely linked to dataset scale and coupled with appropriate regularization. Hydrological knowledge provides an additional safeguard, guiding the selection of features with clear process-based relevance rather than relying solely on statistical correlations.

Taken together, the findings redirect attention from network complexity toward data curation. Sophisticated model architectures cannot compensate for insufficient or poorly balanced datasets, whereas carefully constructed data can enable even relatively simple algorithms to perform reliably. Extending this analysis to other neural architectures such as GRU, Transformer, or graph-based networks would help test the generality of these patterns, while cross-city applications could assess the scalability of dataset design strategies under varying hydrological and infrastructural conditions. Embedding process-based knowledge—through rainfall stratification, infiltration dynamics, or drainage topology—directly into dataset construction represents a promising path forward. Collectively, the evidence clarifies three actionable levers—length thresholds, rainfall distribution balance, and conditional feature enrichment—that should guide the design of balanced datasets for reliable, generalizable applications of machine learning in urban flood prediction.

6. Conclusion

This study reframes urban flood forecasting as a data-design challenge. Controlled experiments with high-fidelity synthetic rainfall–inundation datasets reveal that three factors—dataset length, rainfall-intensity distribution, and feature composition—systematically shape predictive accuracy and generalization. While LSTM served as the benchmark, the patterns observed here reflect broader, architecture-agnostic properties of data-driven hydrological modeling.

Dataset length exhibits a clear saturation threshold. Performance improves steeply as the number of training sequences increases, but beyond approximately 14,400 samples gains plateau, while computational costs continue to rise almost linearly. This “sufficiency frontier” suggests that resources are better invested in transfer learning, active sampling, or multi-basin pre-training rather than brute-force expansion. In short: respect the 14k-sample ceiling.

Rainfall-intensity distribution emerged as the dominant driver of generalization. Models trained on skewed datasets—whether dominated by light or extreme events—developed systematic biases, either muting peaks or inflating minor floods. Mixed-intensity datasets, by contrast, produced robust skill across the full spectrum of rainfall conditions. The implication is clear: balanced representation of rare, high-impact storms must be treated as a design principle, not an afterthought.

Feature enrichment was found to be conditional. Supplementing rainfall with infiltration and drainage inputs improved stability only when the dataset exceeded the sufficiency frontier; under smaller sample budgets, the added complexity exacerbated overfitting. A pragmatic strategy is therefore to start lean with rainfall-only inputs for rapid prototyping and enrich features only once the data budget can support it—start lean, enrich later.

These patterns extend beyond LSTM. Preliminary experiments with GRU and temporal convolutional networks yielded similar saturation and bias signatures, underscoring that the identified principles are not architecture-specific.

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

Future work should extend these analyses to graph-based networks that incorporate sewer topology, and to multi-city catchments with heterogeneous drainage systems.

In summary, the results establish a transferable blueprint for data-centric urban flood forecasting: balance rainfall extremes, respect the sufficiency frontier in dataset length, and enrich features only when statistically supported. Redirecting innovation from increasingly complex models to hydrologically informed data curation provides a pathway toward scalable, reliable, and trustworthy AI in urban flood management.

Author Contribution: HH and YS conceived the study. HH and FX designed the dataset framework and carried out the hydrological–hydrodynamic simulations. PX and YG performed the machine learning experiments and statistical analyses. DL provided methodological guidance on hydrodynamic modeling and contributed to the interpretation of results. HH drafted the manuscript with input from all co-authors. YS revised and finalized the paper. All authors discussed the results and approved the final version.

Competing interests: The authors declare that they have no conflict of interest.

Acknowledgement: This work was jointly supported by the Water Resources Talent Development Fund of the Ministry of Water Resources of China (Grant No. YC202303), the Research and Development Program of China State Construction Engineering Corporation (Grant No. CSCEC-2024-7-41-5), the Key Research and Development Program of Henan Province (Grant No. 241111210300), and the Henan Provincial Science and Technology Development Fund Guided by the Central Government (Grant No. Z20241471035).

7. Reference:

[1] Abbasimehr, H., Paki, R., 2022. Improving time series forecasting using LSTM and attention models. J. Ambient Intell. Hum. Comput. 13, 673-691. https://doi.org/10.1007/s12652-020-02761-x

[2] Ahani, A., Shourian, M., Rahimi rad, P., 2018. Performance Assessment of the Linear, Nonlinear and Nonparametric Data Driven Models in River Flow Forecasting. Water Resour. Manag. 32, 383-399. https://doi.org/10.1007/s11269-017-1792-5

[3] Barreiro, J., Santos, F., Ferreira, F., Neves, R., Matos, J.S., 2022. Development of a 1D/2D Urban Flood Model Using the Open-Source Models SWMM and MOHID Land. Sustain. 15, 707. https://doi.org/10.3390/su15010707

[4] Becher, M., Hothorn, L.A., Konietschke, F., 2025. Analysis of Covariance in General Factorial Designs Through Multiple Contrast Tests Under Variance Heteroscedasticity. Stat. Med. 44. https://doi.org/10.1002/sim.70018

[5] Berkhahn, S., Fuchs, L., Neuweiler, I., 2019. An ensemble neural network model for real-time prediction of urban floods. J. Hydrol. 575, 743-754. https://doi.org/10.1016/j.jhydrol.2019.05.066

[6] Botto, A., Belluco, E., Camporese, M., 2018. Multi-source data assimilation for physically based hydrological modeling of an experimental hillslope. Hydrol. Earth Syst. Sci. 22, 4251-4266. https://doi.org/10.5194/hess-22-

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

4251-2018

[7] Cardoso, M.A., Almeida, M.C., Brito, R.S., Gomes, J.L., Beceiro, P., Oliveira, A., 2020. 1D/2D stormwater modelling to support urban flood risk management in estuarine areas: Hazard assessment in the Dafundo case study. Wiley. 13. http://dx.doi.org/10.1111/jfr3.12663

[8] Chen, J., Li, Y., Zhang, C., Tian, Y., Guo, Z., 2023. Urban Flooding Prediction Method Based on the Combination of LSTM Neural Network and Numerical Model. Int. J. Environ. Res. Public Health. 20, 1043. https://doi.org/10.3390/ijerph20021043

[9] Chen, J., Li, Y., Zhang, S., 2023. Fast Prediction of Urban Flooding Water Depth Based on CNN−LSTM. Water. 15, 1397. https://doi.org/10.3390/w15071397

[10] Chen, Y., Li, J., Xu, H., 2016. Improving flood forecasting capability of physically based distributed hydrological models by parameter optimization. Hydrol. Earth Syst. Sci. 20, 375-392. https://doi.org/10.5194/hess-20-375-2016

[11] Chen, Y., Wang, C., Yang, Q., Lei, X., Wang, H., Jiang, S., Wang, Z., 2024. Model predictive control and rainfall Uncertainties: Performance and risk analysis for drainage systems. J. Hydrol. 630, 130779. https://doi.org/10.1016/j.jhydrol.2024.130779

[12] Chitwatkulsiri, D., Miyamoto, H., 2023. Real-Time Urban Flood Forecasting Systems for Southeast Asia—A Review of Present Modelling and Its Future Prospects. Water. 15, 178. https://doi.org/10.3390/w15010178

[13] Darabi, H., Rahmati, O., Naghibi, S.A., Mohammadi, F., Ahmadisharaf, E., Kalantari, Z., Haghighi, A.T., Soleimanpour, S.M., Tiefenbacher, J.P., Bui, D.T., 2021. Development of a novel hybrid multi-boosting neural network model for spatial prediction of urban flood. Informa UK Ltd. 37, 5716-5741. http://dx.doi.org/10.1080/10106049.2021.1920629

[14] Dong, S., Yu, T., Farahmand, H., Mostafavi, A., 2020. A hybrid deep learning model for predictive flood warning and situation awareness using channel network sensors data. Wiley. 36, 402-420. http://dx.doi.org/10.1111/mice.12629

[15] Fang, Z., Cao, Y.A., Li, Q., Zhang, D.J., Zhang, Z.Y., Liu, Y.B., 2019. Joint Entity Linking with Deep Reinforcement Learning. The Web Conf. 438-447. https://doi.org/10.1145/3308558.3313517

[16] Fu, G., Jin, Y., Sun, S., Yuan, Z., Butler, D., 2022. The role of deep learning in urban water management: A critical review. Water Res. 223, 118973. https://doi.org/10.1016/j.watres.2022.118973

[17] Ghaith, M., Yosri, A., El-dakhakhni, W., 2022. Synchronization-Enhanced Deep Learning Early Flood Risk Predictions: The Core of Data-Driven City Digital Twins for Climate Resilience Planning. Water. 14, 3619. https://doi.org/10.3390/w14223619

[18] Gomes, M.M.D.A., Verçosa, L.F.D.M., Cirilo, J.A., 2021. Hydrologic models coupled with 2D hydrodynamic model for high-resolution urban flood simulation. Springer Sci. Bus. Media LLC. 108, 3121-3157. http://dx.doi.org/10.1007/s11069-021-04817-3

[19] Gupta, A., 2024. Information and disinformation in hydrological data across space: The case of streamflow predictions using machine learning. J. Hydrol. Reg. Stud. 51, 101607. https://doi.org/10.1016/j.ejrh.2023.101607

[20] Hattermann, F., Vetter, T., Breuer, L., Su, B., Daggupati, P., Donnelly, C., Fekete, B., Flörke, F., Gosling, S., Hoffmann, P., Liersch, S., Masaki, Y., Motovilov, Y., Mü ller, C., Samaniego, L., Stacke, T., Wada, Y., Yang, T., Krysnaova, V., 2018. Sources of uncertainty in hydrological climate impact assessment: a cross-scale study. Environ. Res. Lett. 13, 015006. https://doi.org/10.1088/1748-9326/aa9938

[21] Hayder, I.M., Al-amiedy, T.A., Ghaban, W., Saeed, F., Nasser, M., Al-ali, G.A., Younis, H.A., 2023. An

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

Intelligent Early Flood Forecasting and Prediction Leveraging Machine and Deep Learning Algorithms with Advanced Alert System. MDPI AG. 11, 481. http://dx.doi.org/10.3390/pr11020481

[22] He, T.X., Yu, S.C., Wang, Z.Y., Li, J.Q., Chen, Z.Y., 2019. From Data Quality to Model Quality. Proc. 11th Asia - Pac. Symp. Internetware. https://doi.org/10.1145/3361242.3361260

[23] Her, Y., Yoo, S., Cho, J., Hwang, S., Jeong, J., Seong, C., 2019. Uncertainty in hydrological analysis of climate change: multi-parameter vs. multi-GCM ensemble predictions. Sci. Rep. 9. https://doi.org/10.1038/s41598-019-41334-7

[24] Hofmann, J., Schüttrumpf, H., 2020. Risk-Based and Hydrodynamic Pluvial Flood Forecasts in Real Time. MDPI AG. 12, 1895. http://dx.doi.org/10.3390/w12071895

[25] Hofmann, J., Schüttrumpf, H., 2021. floodGAN: Using Deep Adversarial Learning to Predict Pluvial Flooding in Real Time. MDPI AG. 13, 2255. http://dx.doi.org/10.3390/w13162255

[26] Hou, L., Zhang, J., Wu, O., Yu, T., Wang, Z., Li, Z., Gao, J., Ye, Y., Yao, R., 2022. Method and dataset entity mining in scientific literature: A CNN + BiLSTM model with self-attention. Knowl. Based Syst. 235, 107621. https://doi.org/10.1016/j.knosys.2021.107621

[27] Hu, R., Fang, F., Pain, C., Navon, I., 2019. Rapid spatio-temporal flood prediction and uncertainty quantification using a deep learning method. J. Hydrol. 575, 911-920. https://doi.org/10.1016/j.jhydrol.2019.05.087

[28] Huang, X., Li, Y., Tian, Z., Ye, Q., Ke, Q., Fan, D., Mao, G., Chen, A., Liu, J., 2021. Evaluation of short-term streamflow prediction methods in Urban river basins. Phys. Chem. Earth, Parts A/b/c. 123, 103027. https://doi.org/10.1016/j.pce.2021.103027

[29] Hussain, F., Wu, R., Wang, J., 2021. Comparative study of very short-term flood forecasting using physics-based numerical model and data-driven prediction model. Springer Sci. Bus. Media LLC. 107, 249-284. http://dx.doi.org/10.1007/s11069-021-04582-3

[30] Jamali, B., Bach, P.M., Deletic, A., 2020. Rainwater harvesting for urban flood management – An integrated modelling framework. Water Res. 171, 115372. https://doi.org/10.1016/j.watres.2019.115372

[31] Jha, M., Afreen, S., 2020. Flooding Urban Landscapes: Analysis Using Combined Hydrodynamic and Hydrologic Modeling Approaches. MDPI AG. 12, 1986. http://dx.doi.org/10.3390/w12071986

[32] Jhong, Y., Chen, C., Jhong, B., Tsai, C., Yang, S., 2024. Optimization of LSTM Parameters for Flash Flood Forecasting Using Genetic Algorithm. Water Resour. Manag. 38, 1141-1164. https://doi.org/10.1007/s11269-023-03713-8

[33] 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, 105. https://doi.org/10.3390/hydrology9060105

[34] Kim, H., Han, K., 2020. Urban Flood Prediction Using Deep Neural Network with Data Augmentation. MDPI AG. 12, 899. http://dx.doi.org/10.3390/w12030899

[35] Kratzert, F., Klotz, D., Hochreiter, S., Nearing, G.S., 2021. A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling. Hydrol. Earth Syst. Sci. 25, 2685-2703. https://doi.org/10.5194/hess-25-2685-2021

[36] Li, H., Zhang, L., Yao, Y., Zhang, Y., 2025. Prediction of water levels in large reservoirs base on optimization of deep learning algorithms. Earth Sci. Inform. 18. https://doi.org/10.1007/s12145-024-01670-3

[37] Liu, Y., Li, Y., Huang, G., Zhang, J., Fan, Y., 2017. A Bayesian-based multilevel factorial analysis method for analyzing parameter uncertainty of hydrological model. J. Hydrol. 553, 750-762.

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

https://doi.org/10.1016/j.jhydrol.2017.08.048

[38] Liu, Z., Zhou, J., Yang, X., Zhao, Z., Lv, Y., 2024. Research on Water Resource Modeling Based on Machine Learning Technologies. Water. 16, 472. https://doi.org/10.3390/w16030472

[39] Ma, H., Yang, S., Wu, R., Hao, X., Long, H., He, G., 2022. Knowledge distillation-based performance transferring for LSTM-RNN model acceleration. Sig. Image Video Process. 16, 1541-1548. https://doi.org/10.1007/s11760-021-02108-9

[40] Mahmoodzadeh, A., Mohammadi, M., Ghafoor salim, S., Farid hama ali, H., Hashim ibrahim, H., Nariman abdulhamid, S., Nejati, H.R., Rashidi, S., 2022. Machine Learning Techniques to Predict Rock Strength Parameters. Rock Mech. Rock Eng. 55, 1721-1741. https://doi.org/10.1007/s00603-021-02747-x

[41] Nearing, G.S., Kratzert, F., Sampson, A.K., Pelissier, C.S., Klotz, D., Frame, J.M., Prieto, C., Gupta, H.V., 2021. What Role Does Hydrological Science Play in the Age of Machine Learning?. Water Resour. Res. 57. https://doi.org/10.1029/2020wr028091

[42] Nguyen, T., Nguyen, H., Ahmadi, Z., Hoang, T., Doan, T., 2021. On the Impact of Dataset Size:A Twitter Classification Case Study. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol. 210-217. https://doi.org/10.1145/3486622.3493960

[43] Ni, C., Fam, P.S., Marsani, M.F., 2024. A Data-Driven Method and Hybrid Deep Learning Model for Flood Risk Prediction. Int. J. Intell. Syst. 2024, 1-20. https://doi.org/10.1155/2024/3562709

[44] Paz, I., Willinger, B., Gires, A., Ichiba, A., Monier, L., Zobrist, C., Tisserand, B., Tchiguirinskaia, I., Schertzer, D., 2018. Multifractal Comparison of Reflectivity and Polarimetric Rainfall Data from C- and X-Band Radars and Respective Hydrological Responses of a Complex Catchment Model. Water. 10, 269. https://doi.org/10.3390/w10030269

[45] Pollard, J.A., Spencer, T., Jude, S., 2018. Big Data Approaches for coastal flood risk assessment and emergency response. WIREs Clim. Change. 9. https://doi.org/10.1002/wcc.543

[46] Qi, W., Ma, C., Xu, H., Chen, Z., Zhao, K., Han, H., 2021. A review on applications of urban flood models in flood mitigation strategies. Nat. Hazards.. 108, 31-62. https://doi.org/10.1007/s11069-021-04715-8

[47] Śliwowski, M., Martin, M., Souloumiac, A., Blanchart, P., Aksenova, T., 2023. Impact of dataset size and long-term ECoG-based BCI usage on deep learning decoders performance. Front. Hum. Neurosci. 17. https://doi.org/10.3389/fnhum.2023.1111645

[48] Son, H., Kim, D., Kim, S., 2020. Vehicle-Level Traffic Accident Detection on Vehicle-Mounted Camera Based on Cascade Bi-LSTM. J. Adv. Inf. Technol. Converg. 10, 167-175. https://doi.org/10.14801/jaitc.2020.10.2.167

[49] Tikhamarine, Y., Souag-gamane, D., Ahmed, A.N., Sammen, S.S., Kisi, O., Huang, Y.F., El-shafie, A., 2020. Rainfall-runoff modelling using improved machine learning methods: Harris hawks optimizer vs. particle swarm optimization. J. Hydrol. 589, 125133. https://doi.org/10.1016/j.jhydrol.2020.125133

[50] Vitry, M.M.D., Kramer, S., Wegner, J.D., Leitão, J.P., 2019. Scalable flood level trend monitoring with surveillance cameras using a deep convolutional neural network. Copernicus Gmbh. 23, 4621-4634. http://dx.doi.org/10.5194/hess-23-4621-2019

[51] Wang, M., Ying, F., 2023. Point and interval prediction for significant wave height based on LSTM-GRU and KDE. Ocean Eng. 289, 116247. https://doi.org/10.1016/j.oceaneng.2023.116247

[52] Wei, G., Xia, W., He, B., Shoemaker, C., 2024. Quick large-scale spatiotemporal flood inundation computation using integrated Encoder-Decoder LSTM with time distributed spatial output models. J. Hydrol. 634, 130993. https://doi.org/10.1016/j.jhydrol.2024.130993

[53] Wenchuan, W., Yanwei, Z., Dongmei, X., Yanghao, H., 2024. Error correction method based on deep learning

https://doi.org/10.5194/egusphere-2025-4125
Preprint. Discussion started: 14 October 2025

for improving the accuracy of conceptual rainfall-runoff model. J. Hydrol. 643, 131992. https://doi.org/10.1016/j.jhydrol.2024.131992

[54] Wilhelm, B., Rapuc, W., Amann, B., Anselmetti, F.S., Arnaud, F., Blanchet, J., Brauer, A., Czymzik, M., Giguet-covex, C., Gilli, A., Glur, L., Grosjean, M., Irmler, R., Nicolle, M., Sabatier, P., Swierczynski, T., Wirth, S.B., 2022. Impact of warmer climate periods on flood hazard in the European Alps. Nat. Geosci. 15, 118-123. https://doi.org/10.1038/s41561-021-00878-y

[55] Won, Y., Lee, J., Moon, H., Moon, Y., 2022. Development and Application of an Urban Flood Forecasting and Warning Process to Reduce Urban Flood Damage: A Case Study of Dorim River Basin, Seoul. Water. 14, 187. https://doi.org/10.3390/w14020187

[56] Xu, Y., Hu, C., Wu, Q., Jian, S., Li, Z., Chen, Y., Zhang, G., Zhang, Z., Wang, S., 2022. Research on particle swarm optimization in LSTM neural networks for rainfall-runoff simulation. J. Hydrol. 608, 127553. https://doi.org/10.1016/j.jhydrol.2022.127553

[57] Yang, S., Yang, D., Chen, J., Santisirisomboon, J., Lu, W., Zhao, B., 2020. A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data. J. Hydrol. 590, 125206. https://doi.org/10.1016/j.jhydrol.2020.125206

[58] Zhang, D., Martinez, N., Lindholm, G., Ratnaweera, H., 2018. Manage Sewer In-Line Storage Control Using Hydraulic Model and Recurrent Neural Network. Water Resour. Manag. 32, 2079-2098. https://doi.org/10.1007/s11269-018-1919-3

[59] Zhang, J.Y., Wang, Y.T., He, R.M., Hu, Q.F., Song, X.M., 2016. Discussion on the urban flood and waterlogging and causes analysis in China (in Chinese: 中国城市洪涝及其成因分析探讨). Adv. Water Sci. , 485-491. https://kns.cnki.net/kcms2/article/abstract?v=ZZIl2iqmIcQBMEpk-w34cNDtyEAX4i6FmMSjp2mf0TDk5ZzFz9CYk5bL-vpODYo7AGOlinhAN_aQvGyqsrDq4MYn6oKUvP9uz_aXAqITiMDqQ30gOhTUlM4Or9kfodHI3Wm-NtqD0aN0jjR9mkCitFDdcbqg06D7FNjf8Yy7ic-XRMV2SvPXirMFwS1St0oTHNoqKzL06No2GjaCf9d3xEkD-rhO6m9f&uniplatform=NZKPT&language=CHS

[60] Zheng, Y., Jing, X., Lin, Y., Shen, D., Zhang, Y., Yu, M., Zhou, Y., 2024. Research on nowcasting prediction technology for flooding scenarios based on data-driven and real-time monitoring. Water Sci. Technol. 89, 2894-2906. https://doi.org/10.2166/wst.2024.174

[61] Zhou, Q., Teng, S., Situ, Z., Liao, X., Feng, J., Chen, G., Zhang, J., Lu, Z., 2023. A deep-learning-technique-based data-driven model for accurate and rapid flood predictions in temporal and spatial dimensions. Hydrol. Earth Syst. Sci. 27, 1791-1808. https://doi.org/10.5194/hess-27-1791-2023