https://doi.org/10.5194/egusphere-2025-3171
Preprint. Discussion started: 1 August 2025

A highly generalizable data-driven model for spatiotemporal urban flood dynamics real-time forecasting based on coupled CNN and ConvLSTM

Wangqi Lou[1,2], Xichao Gao[1,2], Joseph Hun Wei Lee[3], Jiahong Liu[1,2], Jing Peng[1,2], Lirong Dong[3], and Kai Gao[1,2]

[1]State Key Laboratory of Water Cycle and Water Security, Beijing, 100038, Beijing, China
[2]China Institute of Water Resources and Hydropower Research, Yuyuantan South Road, Haidian District, Beijing, 100038, Beijing, China
[3]Macau University of Science and Technology, Avenida WaiLong,Taipa, Macau, 999078, Macau, China

Correspondence: Xichao Gao (gaoxc@iwhr.com)

Abstract. Flooding has become one of the most severe natural hazards in urban areas. Real-time and accurate prediction of flood processes is a crucial approach to mitigate urban flood disasters. Data-driven models based on machine learning methods offer significantly higher computational efficiency than physics-based models and have been widely applied in real-time urban flood simulation. However, most data-driven models target the temporal process of inundation depths at specific sites or the spatial distribution of peak inundation depths, while some models capable of simulating spatiotemporal urban flood inundation often lack spatial generalization capabilities. In this study, we proposed a novel data-driven model to predict the spatiotemporal distribution dynamics of urban inundation depths. The model integrates a ConvLSTM-based component alongside a CNN-based component via a concatenation process, facilitating the extraction of information from both temporal sequences and static geospatial features concurrently. A tiling approach that divides the study area into distinct spatial sub-regions, which serve as independent training samples, was employed during model training to enhance the model's generalization capability. The proposed model was applied to a flood-prone urban area in Macao and compared with a physics-based model. The results show that: (1) the proposed model effectively captures the inundation processes at specific sites, with NSE >0.80 for the majority events, as well as RMSE and MAE values <0.20. (2) The proposed data-driven model demonstrates robust generalization performance, with simulated inundation processes closely aligned with the results of the physics-based model in most regions (mean NSE >0.70, RMSE <0.10, MAE <0.10). Notable discrepancies persist only in localized zones of abrupt terrain variations, particularly near building edges.

1 Introduction

Urban flooding is a critical natural disaster that causes significant loss of life and property damage in urban areas and is expected to increase in both frequency and intensity as a result of global warming and rapid urbanization. In coastal cities, these challenges are intensified by storm surges and rising sea levels, which impose additional burdens on urban drainage systems. Rapid convergence of runoff in urban settings, compounded by intense short-duration rainfall, facilitates the rapid development of urban flooding, thus complicating emergency response efforts (Fu et al., 2023; Wang et al., 2022; Balaian et al., 2024). Consequently, flood forecasting using numerical models has emerged as an essential method to mitigate flood-related losses. In order to underpin effective disaster mitigation strategies, there exists a necessity for precise spatio-temporal processes of inundation depths, thus physics-based hydrodynamic models, which can simulate spatio-temporal flood dynamics in urban areas, have been developed and implemented. However, these models are computationally intensive, leading to low simulation efficiency. When deployed in extensive urban regions, the computation time required by such models may exceed the duration of the events they aim to simulate. Prolonged computation times significantly limit the utility of these models in real-time flood forecasting.

To mitigate the limitations associated with the inefficiency of high-precision, high spatiotemporal resolution flood simulations using physics-based hydrodynamic models, data-driven models have been devised in recent years. These models, characterized by machine learning (ML) or deep learning (DL) methods, infer the input-output relationships from historical data rather than relying on predefined equations or physical laws employed in process-based models for the purposes of prediction or comprehension of complex systems. Upon completion of the training phase, data-driven models are capable of executing a substantial number of simulations within a brief timeframe, all the while preserving a high level of accuracy. Moreover, data-driven models are able to leverage a broader spectrum of diverse datasets more comprehensively. Numerous data-driven models have been employed in the simulation of urban flooding in recent years. For example, Berkhahn et al. (2019) introduced an artificial neural network architecture to forecast peak water levels during flash flood incidents and subsequently evaluated the model in two urban locations. Löwe et al. (2021) introduced a model referred to as U-FlOOD, grounded in the U-NET methodology, to forecast the spatial distribution of the maximum flood depth. Gao et al. (2024) used a one-dimensional convolutional neural network (1D-CNN) to simulate the spatial distribution of the maximum inundation depth in the Tianhe district of Guangzhou City, China. Dai and Cai (2021) simulated water depth dynamics during typhoons in Macao, China, using a back-propagation neural network (BNPP). Zahura et al. (2020) predicted water depth over time in the roads segments during rainfall using a Random Forest (RF) model. These studies have shown that data-driven models are capable of effectively simulating urban flood events while also exhibiting significantly higher computational efficiency compared to traditional physics-based models. However, most studies (Hou et al., 2021; Aderyani et al., 2025; Piadeh et al., 2023) that employ data-driven models for urban flood modeling have focused predominantly on the spatial distribution of maximum depths of flooding or inundation processes at specific locations, while limited attention has been paid to the application of data-driven approaches to simulate spatiotemporal inundation dynamics throughout urban flood events. To simultaneously account for the temporal and spatial dependencies of the input data, Shi et al. (2015) integrated the LSTM and CNN models, proposing the convolutional LSTM model (ConvLSTM). They applied the model to precipitation nowcasting, demonstrating its ability to capture spatiotemporal correlations and perform effectively. ConvLSTM-based models have subsequently been widely employed in flood prediction applications due to their effective capacity to extract temporal and spatial information from input features. Specifically, Yang et al. (2024) proposed a ConvLSTM based model to simulate the spatiotemporal dynamics of inundation depths in urban areas and evaluated the model in Huangpu District, Guangzhou city, China. Wang et al. (2024b) introduced a time-guided convolutional neural network by integrating the target time matrix into the input features of the ConvLSTM-based model and evaluated

https://doi.org/10.5194/egusphere-2025-3171
Preprint. Discussion started: 1 August 2025

the model in the metropolitan area of Dalian, China. However, most studies that used ConvLSTM-based models to predict urban inundation depths did not consider static data, such as topography and pipe networks, particularly, or just incorporate static data into input features of ConvLSTM simply. Incorporating static data directly as inputs in ConvLSTM architectures may diminish their influence, as the model inherently prioritizes temporal dynamics over time-invariant attributes. Static data exert a critical influence on urban flooding processes, and neglecting their incorporation would lead to significant adverse impacts on the generalization capability of ConvLSTM-based models.

In order to address the generalization challenges associated with ConvLSTM in the context of urban flooding forecasting, this study proposes a deep learning framework that integrates ConvLSTM and CNN. The ConvLSTM component of the proposed model is utilized to capture the spatial and temporal dependencies inherent in input time series, while the CNN component addresses the spatial dependencies present in static geospatial inputs. To enhance the applicability of the model for real-time flood forecasting and facilitate the incorporation of observed flood data during model execution, an auto-regressive prediction framework is employed, wherein the inundation depth map predicted in the current timestep serves as the input for the subsequent timestep. Furthermore, considering the hydrodynamics characteristics of water flow, the target region is partitioned into multiple segments rather than treated as a singular entity during the training phase, thereby augmenting the model's capacity for generalization. The model was subsequently evaluated in Macao, China.

The organization of the paper is as follows. Section 2 describes the study area; Section 3 details the proposed methodology; Section 4 presents a comparative analysis of the simulation results between the proposed data-driven model and the physics-based model. Section 5 discusses the limitations of the proposed model, and Section 6 provides a concise conclusion.

2 Study Area and Data

2.1 Study Area

The research is concentrated in the western sector of the Macao Peninsula (Figure 1). This region is characterized by a subtropical climate and is influenced by an oceanic monsoon system, with an average annual precipitation of 1966.6 mm. It is highly urbanized and characterized by a low-lying topography with the lowest elevation only 1.4 meters above sea level and an average elevation of approximately 2 meters. Due to its climatic and geographical characteristics, this region suffers from floods induced by extreme precipitations and storm surges. The 4.06 km^2 region was ultimately selected as the focus of the study, based on the topographic distribution and drainage systems, as it is the site most prone to historical inundation events in Macao (Dong et al., 2024).

2.2 Data

2.2.1 Geospatial data

Digital Elevation Model (DEM) data, with a spatial resolution of 2 meters, along with the drainage network (Figure 1) and the building distribution information, were obtained from the Macao Cartography and Cadastre Bureau. The elevations in the

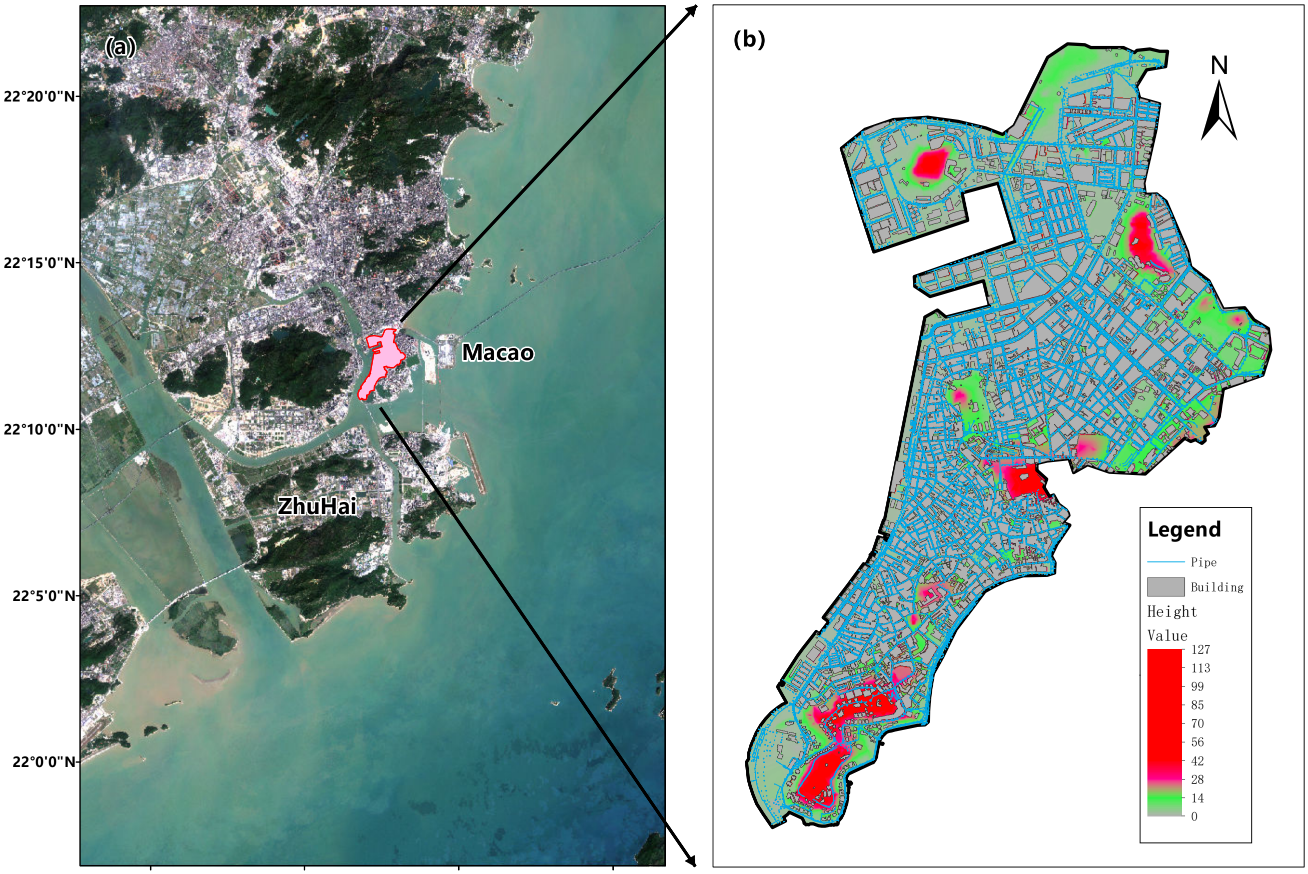

Figure 1. The study area. Satellite imagery from © Google Maps (© Google)

DEM at building sites increased by 5 meters to account for the impediment effect of buildings on surface water flow. All data were verified on the basis of satellite imagery and field investigation.

2.2.2 Rainfall

In this study, two types of rainfall data were used, including historical observed data and designed data, to consider more rainfall conditions. The hourly rainfall records for the Dapaotai station, covering the period from 2000 to 2022, were obtained from the Macao Meteorological and Geophysical Bureau. The designed rainfall was formulated by integrating rainfall patterns and intensities. Rainfall patterns were identified by classifying historical rainfall records into seven prototypical patterns, preserving the three most frequently occurring patterns. The seven typical rainfall patterns are shown in Figure 2. The patterns

I, II, and III exhibit a unimodal distribution, with peaks occurring in the early, middle, and late stages, respectively. The pattern IV is characterized by a uniform distribution, while patterns V, VI, and VII show a bimodal distribution (Chen et al., 2015). The predominant rainfall patterns within the study area are the pattern I, the pattern III and the pattern IV, contributing to 41.3%, 37.2%, and 13.2% of the total occurrences, respectively. Therefore, these three rainfall patterns were selected as the designed rainfall patterns. Rainfall intensities were computed based on equations provided by the Macao Meteorological and Geophysical Bureau. The equation is as follows.

$$I = at^b \tag{1}$$

where, I represents the intensity of the rainfall (mm/h); t represents the duration of the rainfall (min); a and b are experimental parameters, which can be obtained from the Macao Regulations on Water Supply and Drainage (Zhang et al., 2024).

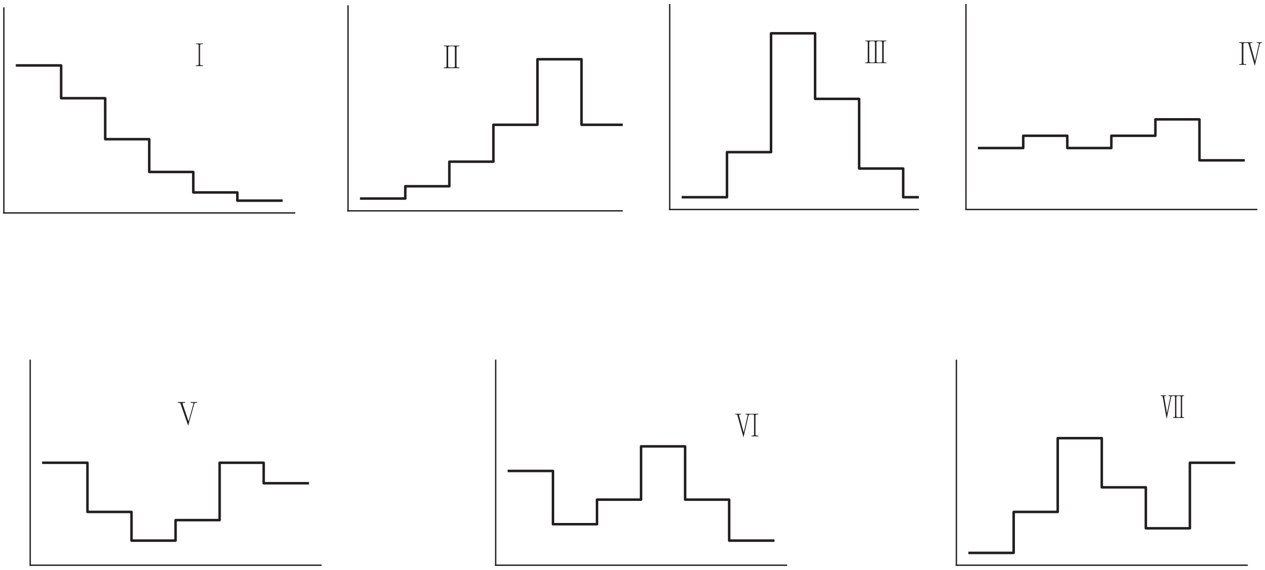

Figure 2. Seven typical rainfall patterns

Urban flooding is mainly caused by short, intense rainfall, so a 6-hour duration was chosen for the designed rainfall. The designed rainfall amounts for return periods of 10, 20, 50, and 100 years were used to cover most of the rainfall intensities observed in this region. As a result, a total of 12 rainfall scenarios were devised through the integration of three distinct rainfall patterns with four different return periods. Given the relatively small size of the study area, it was assumed that the rainfall would be uniformly distributed throughout the entire region.

2.2.3 Storm tide

110 Macao Peninsula is frequently affected by storm surges. Due to the low-lying topography, storm surges negatively impact the drainage capacity of the study area, thereby exacerbating flooding events when they coincide with heavy rainfall. Consequently, it is imperative to incorporate the tidal process in the analysis of flooding within the study area. In this study, the designed tidal process lines of 5 warning levels derived by Zhang et al. (2024) were used. The designed tidal process lines are shown in Figure 3.

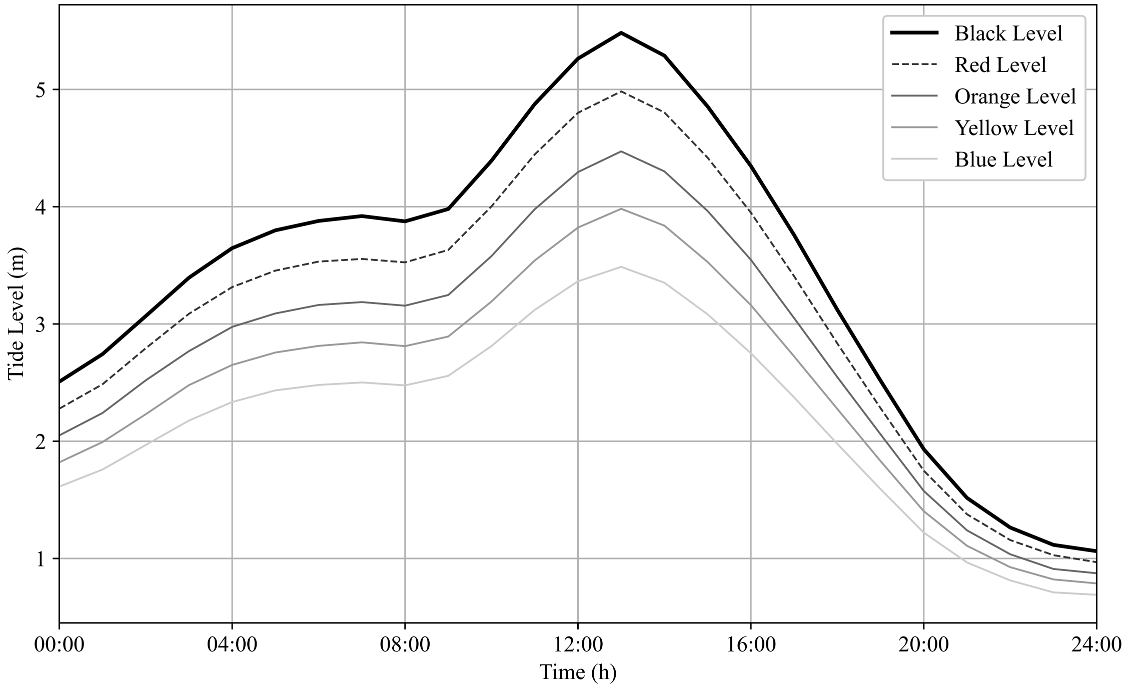

Figure 3. The designed tidal process lines

115 ### 2.2.4 Synthetic compound scenarios

The integration of rainfall events and tidal process lines is essential to accurately represent the combined impact of precipitation and storm surges. To integrate rainfall events with tidal process lines in different temporal phases and warning levels, this study proposes the following method for combination. For each rainfall event, a tidal process line is initially selected at random from among the five warning levels. Subsequently, a 6-hour interval is randomly determined from the chosen tidal process line to be

https://doi.org/10.5194/egusphere-2025-3171
Preprint. Discussion started: 1 August 2025

integrated with the rainfall event. The combination process was conducted thrice for each rainfall event to augment the sample variability.

3 Methodology

The data flow and workflow of this study are shown in Figure 4. Initially, the dataset was meticulously prepared. The input features for the proposed data-driven model were rigorously selected and classified into static and dynamic categories based on their temporal invariance. Urban flood inundation, which constitutes the output of the proposed data-driven model, was simulated using a physics-based hydrodynamic model, in light of the relative scarcity of inundation depth monitoring data. The dynamic and static input features, along with the corresponding simulated water depths, were paired and randomly divided into training and testing datasets. Subsequently, the proposed data-driven model was trained and its accuracy and computational efficiency was evaluated.

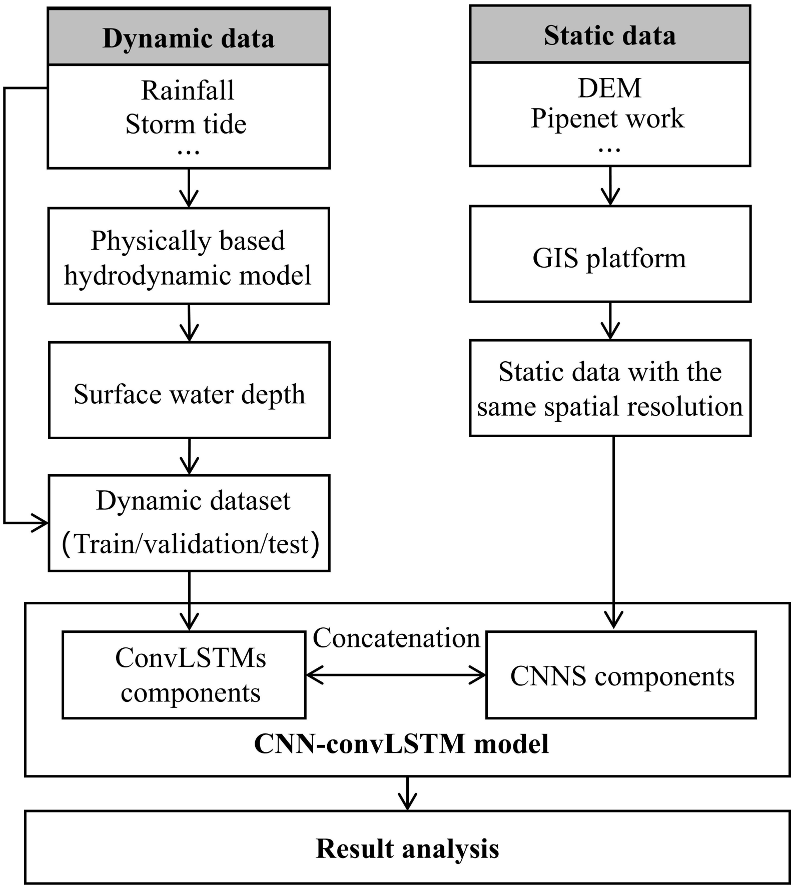

Figure 4. Data flow and workflow

3.1 physics-based Hydrodynamic Model

We developed a hydrodynamic model capable of simulating two-dimensional surface flows, one-dimensional flows from the pipe drainage network and the interactions between surface flows and flows from the pipe drainage network. The two-dimensional module solves the Saint-Venant equations using finite-volume methods based on triangular meshes (Anastasiou and Chan, 1997). The module effectively addresses dry-wet alternation, crucial in urban flooding. The EXTRAN module of the SWMM model was used to simulate flows in pipe drainage networks. It simulates the drainage system as links and nodes, enabling the simulation of parallel or looped pipe networks, as well as weirs, orifices, pumps, and system surcharges. The module assumes that the flow within a link is uniform and that the water surface at the node is continuous, resolves the one-dimensional Saint-Venant equations in link-node structures by employing the Predictor-Corrector Iterative method (Rossman and Huber, 2017). The interaction between surface flows and flows in pipe drainage networks is simulated using weir flow formulas (Wang et al., 2024a). The rainfall-runoff process is modeled using the Horton infiltration method. This empirical formula posits that, as the soil reaches saturation, the infiltration rate decreases exponentially from an initial maximum value to a steady minimum rate (Gülbaz et al., 2020; Beven, 2004).

The hydrodynamic model for the study area was built in our previous research. Detailed information on the model can be found in Dong et al. (2024). The performance of the hydrodynamic model was verified by comparing the simulated inundation depths with the waterlogging points observed during Typhoon Mangkhut. The Nash efficiency coefficients for the observed points exceed 0.75. Therefore, the model can be considered capable of accurately reflecting the relationship between rainfall and waterlogging in the area. Training the DL model with the simulation results from hydrodynamic models is a reasonable approach.

3.2 CNN-ConvLSTM Coupled Model

The proposed model is based on ConvLSTM, which is effective in capturing spatiotemporal correlations in multidimensional time series. However, ConvLSTM has limitations in handling static features such as geospatial information, which are essential to simulate the characteristics of urban floods. The LSTM components of the model, such as input and forget gates as well as memory cells, become superfluous in scenarios devoid of temporal dynamics, thereby introducing unwarranted computational overhead and increasing parameterization, which exacerbate the risk of overfitting. The necessity to process even a single-step input through temporally unfolded operations further results in resource inefficiency when compared to models focused solely on spatial data, like CNNs. Moreover, instability during training can occur, as gradient propagation within LSTM modules presents difficulties in adapting to static data absent of sequential dependencies. Consequently, the hybrid architecture of ConvLSTM is excessively complex for static contexts, where simpler models, such as CNNs or MLPs, demonstrate greater efficiency and performance by eliminating redundant temporal mechanisms. To enhance the efficiency of processing static data, we propose a novel hybrid architecture that integrates ConvLSTM and CNN in parallel. The temporal dynamic information and static features are separately processed by ConvLSTM cells and CNN cells, integrated through feature aggregation, and subsequently decoded to capture the spatiotemporal flood processes.

3.2.1 Convolutional neural network

Convolutional Neural Networks (CNNs) are deep learning models for grid-structured data like images. Widely used in fields such as computer vision and remote sensing, CNNs efficiently learn spatial feature hierarchies. They offer parameter efficiency through weight sharing in convolutional layers, reducing learnable parameters compared to fully connected networks. CNNs ensure translation invariance, enabling robust feature extraction despite data shifts. By automatically capturing local and global patterns, they excel in tasks like image classification and semantic segmentation. Techniques like pooling and dropout improve computational efficiency and reduce overfitting, enhancing generalization across datasets. In this study, two-dimensional CNNs are used to handle static features such as DEM and the spatial distribution of drainage systems.

3.2.2 Convolutional Long Short-Term Memory network

Convolutional Long Short-Term Memory (ConvLSTM) is a specialized recurrent neural network architecture designed to model spatiotemporal correlations in sequential data with spatial structure. Unlike traditional LSTM, which processes temporal dependencies in vectorized sequences, ConvLSTM replaces fully connected operations with convolutional gates, enabling it to simultaneously capture local spatial patterns and long-range temporal dynamics. This design inherently preserves the spatial topology while capturing temporal dependencies, enabling synergistic learning of localized spatial patterns and long-range temporal dynamics. Compared to architectures that separately stack CNNs and LSTMs, ConvLSTM achieves superior parameter efficiency through convolutional kernel sharing and supports hierarchical multiscale spatiotemporal feature extraction via deep stacking. Its state-of-the-art performance in applications such as precipitation nowcasting and traffic flow prediction underscores its capability to model complex spatiotemporal interactions, establishing it as a benchmark for grid-structured sequential data(Ahmad et al., 2023; Zhang et al., 2017; Lu et al., 2024). Critically, this integrated approach eliminates the need for hand-crafted feature engineering, enhancing generalization across diverse domains. In this study, Time series data, including precipitation and inundation depth, are incorporated as inputs to the ConvLSTM branch for spatiotemporal modeling of flood dynamics. The core operations of a ConvLSTM cell are defined as:

$$i_t = \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \circ c_{t-1} + b_i) \tag{2}$$

$$f_t = \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \circ c_{t-1} + b_f) \tag{3}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \tag{4}$$

$$o_t = \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \circ c_t + b_o) \tag{5}$$

$$h_t = o_t \circ \tanh(c_t) \tag{6}$$

Where $*$ denotes convolution, $\circ$ represents the Hadamard product, and σ is the sigmoid function. i_t, f_t, o_t represent input, forget, and output gates controlling information flow, respectively. X_t denotes the input tensor at time step t. H_t and C_t denote hidden state tensor and cell state tensor at time t. W_{x*} and W_{h*} denote convolutional kernels. W_{c*} denotes kernel for peephole connections. b_* denotes bias terms for each gate. $tanh$ denotes hyperbolic tangent activation function.

3.2.3 Inputs and outputs

The presented model's input features encompass static attributes, which correspond to fixed spatial characteristics, along with spatiotemporal dynamic attributes that signify meteorological and hydraulic variables.

According to the research of Gao et al. (2024), DEM, ASP, CURV, SLOPE, R_DEM, M_SLOPE, TWI, MANHOLE, and NETWORK are sensitive to urban inundation depths and were selected as static input features. The definitions of these static features are provided in Table 1. Among these static features, DEM, ASP, R_DEM, CURV, SLOPE, and M_SLOPE serve as indicators of local terrain deformations linked to surface flow convergence. TWI functions as an integrative index that reflects the potential for soil moisture saturation within the study area. Meanwhile, MANHOLE and NETWORK denote the distribution of drainage systems and suggest levels of urbanization to a certain degree. The correlation analysis conducted by Gao et al. (2024) revealed that R_DEM and TWI exhibit stronger correlations with inundation depth. The more pronounced correlation of R_DEM compared to DEM indicates that pluvial flooding is generally attributed to relative topographic depressions rather than absolute elevation. Furthermore, the selected variables, with the exception of CURV and SLOPE, exhibit no significant multicollinearity. Nevertheless, these two variables represent distinct topographic attributes. CURV reflects the characteristics of the accumulation of regional surface water, while SLOPE refers to the characteristics of regional surface drainage. Consequently, both variables were incorporated into the study. Furthermore, a building mask (MASK) was incorporated into the input features, acknowledging the significant impact of buildings on hydrological flow within urban environments. The analysis of skewness and kurtosis revealed that certain datasets present right-skewed, long-tailed distributions. Therefore, suitable transformations were implemented to mitigate the effects of outliers. The selection of transformation methods was based on skewness, kurtosis, and the occurrence of negative values in the datasets. A square root transformation was applied to DEM, TWI, M_SLOPE, MANHOLE, and NETWORK, whereas a cube root transformation was implemented for R_DEM, SLOPE, and CURV. The selected variables were normalized to a range of 0 to 1 to ensure consistent feature scaling, stabilize training, speed up convergence, and enhance model generalization. The static input features can be expressed as

$$\begin{aligned} X_{sta} =&(DEM, ASP, CURV, SLOPE, R\_DEM, \\ &M\_SLOPE, MANHOLE, NETWORK, MASK) \end{aligned} \tag{7}$$

No.	Input feature	Meaning
1	DEM	Terrain elevation data augmented with architectural structural attributes.
2	ASP	Terrain flow direction. Decomposed into two orthogonal raster layers (ASP_COS and ASP_SIN).
3	CURV	Plan curvature, quantifies the lateral convexity or concavity of terrain surfaces along contour lines. It characterizes the divergence or convergence of surface flow perpendicular to the slope direction.
4	SLOPE	Terrain slope, quantifies the steepness of a topographic surface by measuring the maximum rate of elevation change across a spatial domain, determining the general direction of water flow.
5	R_DEM	Relative Digital Elevation Model (relative DEM) refers to a representation of terrain elevation where elevations are expressed relative to a specific local reference (focal mean elevation within 50 m radius used in this study) rather than an absolute global datum, such as mean sea level.
6	M_SLOPE	Local mean terrain slope, represents the mean gradient of elevation changes over a specified spatial scale (within 50 m radius used in this study) and used to capture surface runoff characteristics in a specified area.
7	MANHOLE	Kernel density of manhole, indicates the concentration of manhole points in the study area.
8	NETWORK	Kernel densities of pipe network length (DONL) and pipe network length area (DONA), respectively. Indicates the spatial distribution of the drainage network.
9	MASK	Matrix with 0 illustrating buildings and 1 illustrating other areas.

Table 1. Definition of static input features.

Rainfall, tide level, and inundation dynamics are incorporated as time-varying forcing input to the model. To improve the efficacy of the model in real-time urban flood forecasting and to effectively incorporate historical information, rainfall and tide level from the three previous hours to the following 1 hour, in conjunction with the inundation depth of the last three hours, are employed to predict the inundation distribution for the following hour. The rainfall distribution is obtained using the inverse distance weighting interpolation based on observed station data, while the inundation distribution is extracted from the model's output of the preceding step. The dynamic input features can be expressed as

$$X_d = (P^{t-s:t+1}, T^{t-s:t+1}, D^{t-s:t}) \tag{8}$$

where P, T, and D represent rainfall, tide level, and inundation water depths, respectively. The number of historical temporal intervals (s) can be considered a hyperparameter within the model, as it is potentially influenced by the scale of the study area

https://doi.org/10.5194/egusphere-2025-3171
Preprint. Discussion started: 1 August 2025

and the prevailing climate conditions. Considering that the study area is relatively small and its rainfall-runoff process is rapid, in this study the value of s was set to 3 hours.

The study area comprises a total of 1976 by 1321 grid cells. Given the complex structure of the model and numerous input variables, training it for the entire study area as a whole requires substantial computer memory. Consequently, the study area was partitioned into blocks of 100 by 100 grid cells, with each block regarded as an independent sample for model input, thereby mitigating memory demands during batch training. The partitioning is physically reasonable, since the primary geographical factor affecting urban pluvial flooding is local topographic deformations. An overlap length of less than ten percent of the total extent of the study area in the respective direction was allowed during the partitioning process to increase the sample size. The implementation of a spatial tiling strategy enables the model to intrinsically capture the influence of static geospatial inputs on flood inundation dynamics, thereby improving spatial generalization by systematically decoupling localized terrain-flood interactions from basin-scale hydrological patterns.

The model outputs the estimated water depth at time $t+1$. The output is used to update the dynamic input feature D^{t+1} for the subsequent iteration, which directs the model to forecast the inundation distribution in a recursive way. This framework enables real-time data assimilation of inundation depths during model execution by integrating observational measurements.

3.2.4 Model structure

The structure of the CNN-ConvLSTM coupled model is shown in Figure 5. The model has two components, including the ConvLSTM-based component and the CNN-based component. The ConvLSTM-based component employs an encoder-decoder architecture. The encoder consists of a single ConvLSTM2D layer with 32 convolutional kernels (3×3) and LeakyReLU activation, yielding a 100×100×32 spatiotemporal feature tensor. The decoder mirrors the encoder's structure to reconstruct spatiotemporal dependencies. A batch normalization layer is incorporated between the encoder and the decoder to accelerate the convergence of the model by stabilizing the propagation of the gradient. The input of the ConvLSTM-based component is a temporal sequence comprising rainfall, tidal level, and depth of inundation in four continuous time intervals, structured as a $4\times100\times100\times3$ tensor (timesteps $\times$ height $\times$ width $\times$ channels). The CNN-based component comprises three 2D convolutional layers. The initial layer employs 64 kernels (3×3) to extract primary spatial features, while the subsequent two layers utilize 32 kernels (3×3) each to compress channels and enhance discriminative power through hierarchical feature refinement. The CNN module processes a single-sample input tensor of dimensions 1×100×100×9, comprising nine categories of geographic information data (e.g., digital elevation models [DEM], drainage networks), and generates an output tensor of size 1×100×100×32, which maintains spatial alignment with the ConvLSTM branch for subsequent feature fusion. The outputs of the two branches are merged by a concatenation operation, forming a fused tensor of dimensions $1\times100\times100\times64$. Finally, a 3D convolutional layer employing $3\times3\times3$ kernels compresses the channel dimension to generate the water depth prediction map ($1\times100\times100\times1$).

https://doi.org/10.5194/egusphere-2025-3171
Preprint. Discussion started: 1 August 2025

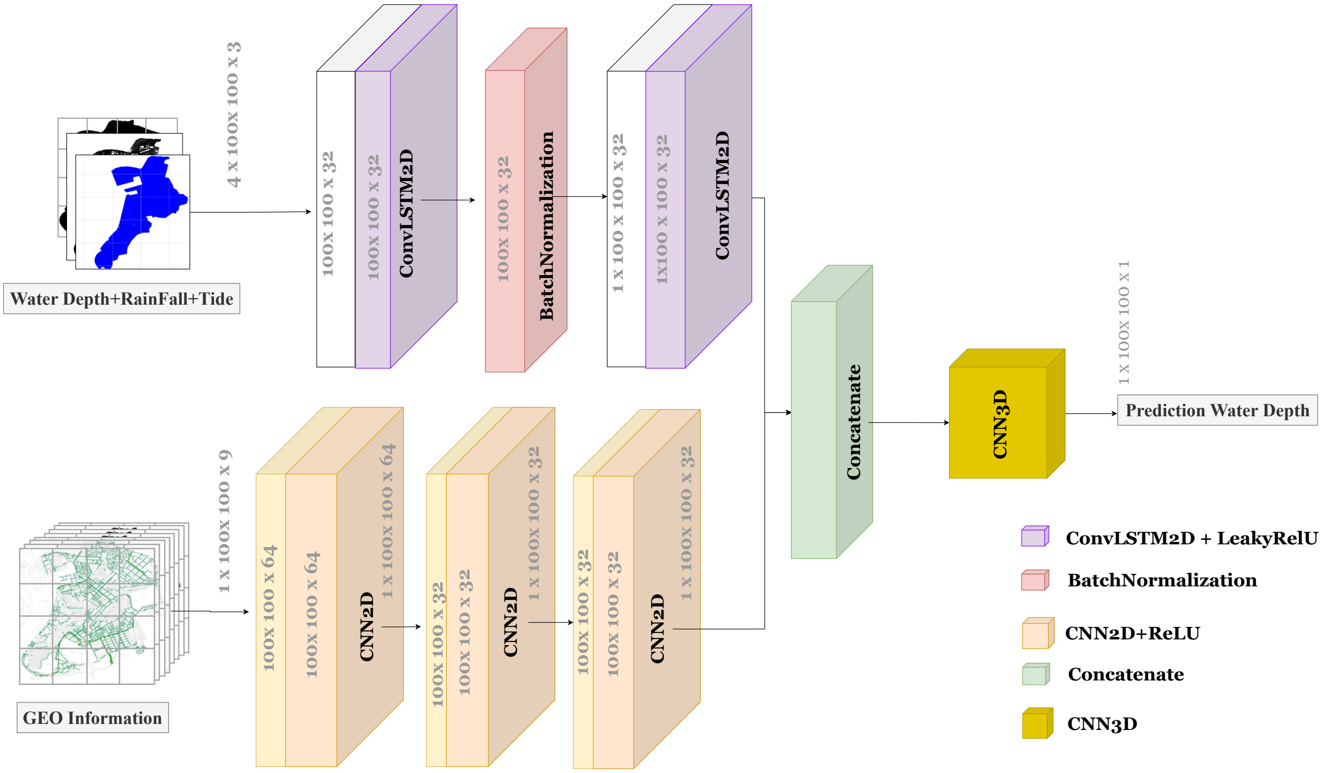

Figure 5. The architecture of the CNN-ConvLSTM coupled model.

3.3 Data-driven Model Setup

3.3.1 Training strategy

A multistep-ahead loss function $\mathcal{L}$ was used to stabilize the model output, measuring the accumulated error over consecutive time steps. The function is defined as

$$\mathcal{L} = \frac{1}{H}\sum_{\tau=1}^{H} \|\hat{D}^{t+\tau} - D^{t+\tau}\|_2 \tag{9}$$

where H refers to the number of consecutive prediction time instants; $\hat{D}^{t+\tau}$ and $D^{t+\tau}$ are the estimated and observed inundation depths, respectively. The loss function computes the average root mean squared error (RMSE) across all prediction iterations. This process enables the model to refine its predictions autonomously and enhances its capability to produce accurate output even when initial predictions are slightly inaccurate, thus enhancing its robustness. To improve training speed and stability, we employ a progressive training strategy (curriculum learning strategy), initially calibrating the model over a restricted set of forecast horizons and incrementally expanding the prediction window to H (Bentivoglio et al., 2023). The progress of the training strategy is shown in Figure 6.

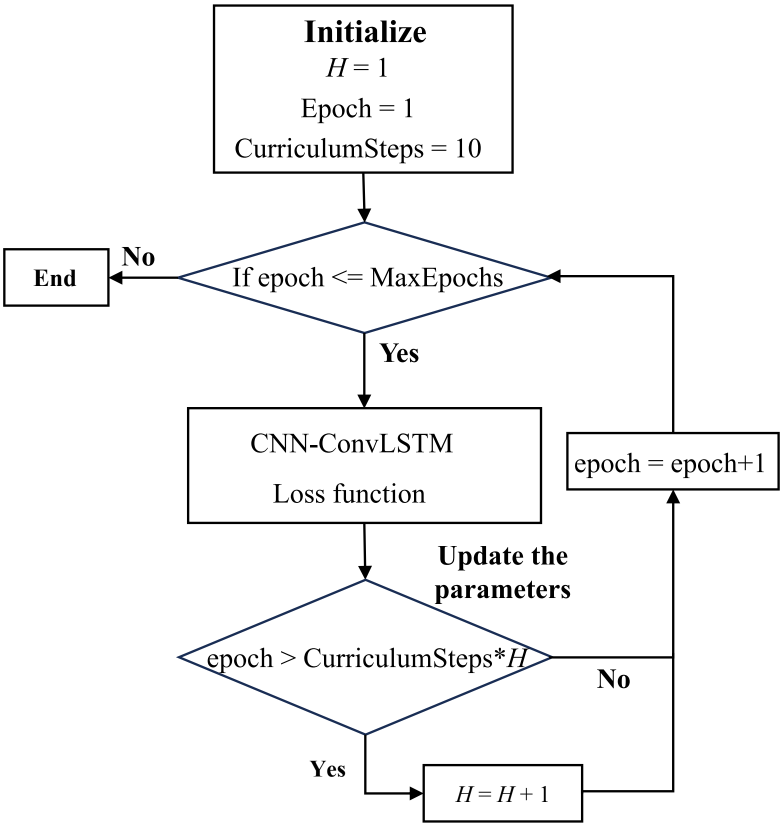

Figure 6. Workflow of curriculum learning strategy

As described in the section on inputs and outputs, the data pairs employed to train the data-driven model were extracted from rainfall events and their corresponding inundation simulations of the physics-based hydrodynamic model using a fixed time window. Therefore, it is unnecessary to account for the effects of rainfall patterns and return periods particularly, when preparing the training and testing datasets. 80% of the dataset was randomly allocated for model training, while the remaining 20% was reserved for model testing. To ensure the model's generalization capability, the test set was strictly excluded from the training process, and 15% of the training set was allocated as a validation set for hyperparameter tuning and early stopping.

3.3.2 Evaluation metrics

The efficacy of the proposed CNN-ConLSTM model was assessed by comparing the water depths forecasted by the CNN-ConLSTM model with those simulated by the physics-based hydrodynamic model. The assessment used various performance metrics, including the correlation coefficient (CC), the root mean square error (RMSE), the mean absolute error (MAE), and the critical success index (CSI). CC indicates the linear correlation between predictions and observations to assess trend consistency. RMSE and MAE measure average prediction errors; RMSE highlights larger errors, while MAE shows the overall error. CSI assesses the model's ability to accurately distinguish between flooded and non-flooded areas. The ranges of values and the optimal criteria for these evaluation metrics are summarized in Table 2. The definitions of these metrics are as follows.

$$\text{NSE} = 1 - \frac{\sum_{t=1}^{T}\left(d^t - \hat{d}^t\right)^2}{\sum_{t=1}^{T}\left(d^t - \overline{d}\right)^2} \tag{10}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - \hat{d}_i)^2} \tag{11}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|d_i - \hat{d}_i| \tag{12}$$

where d_i and $\hat{d}_i$ represent observed and predicted water depth on the i-th grid; d^t and $\hat{d}^t$ represent observed and predicted water depth of time t; n denotes the number of predicted values.

$$\text{CSI} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{13}$$

where TP (True Positive) signifies the count of wet grids accurately predicted by the proposed data-driven model; FP (False Positive) refers to the count of dry grids mistakenly identified as wet; FN (False Negative) represents the number of grids mistakenly predicted to be dry.

Evaluation Metrics	**Range**	**Best Value**
NSE	$-\infty \sim 1$	1
MAE	$0 \sim \infty$	0
RMSE	$0 \sim \infty$	0
CSI	$0 \sim 1$	1

Table 2. Range and best values of evaluation metrics.

4 Results

4.1 Performance of data-driven model in water depth simulation at flood-prone locations

Water depth predictions from the CNN-ConvLSTM model were compared with the physics-based model in three flood-prone locations, LHK, IHS, and LPM, which were used to calibrate the physics-based model in our previous study (Dong et al., 2024). Inundation processes corresponding to five rainfall events at each location were randomly selected to assess the capability of the proposed data-driven model in replicating the inundation dynamics simulated by the physics-based model. As shown in Figure 7, the water depth processes predicted by the data-driven model for the selected rainfall events exhibited a strong consistency with the simulations generated by the physics-based model. In five randomly selected rainfall events, the NSE values at stations IHS and LPM consistently exceeded 0.80, while station A also showed NSE values above 0.80 in all but one event, which had 295 an NSE of 0.75. Among the 15 rainfall events evaluated, 13 exhibited both RMSE and MAE below 0.20, with more than half exhibiting values below 0.10 for both metrics. This demonstrates that the proposed data-driven model effectively captures the dynamics of water depth in flood-prone locations, capturing key temporal patterns of inundation processes.

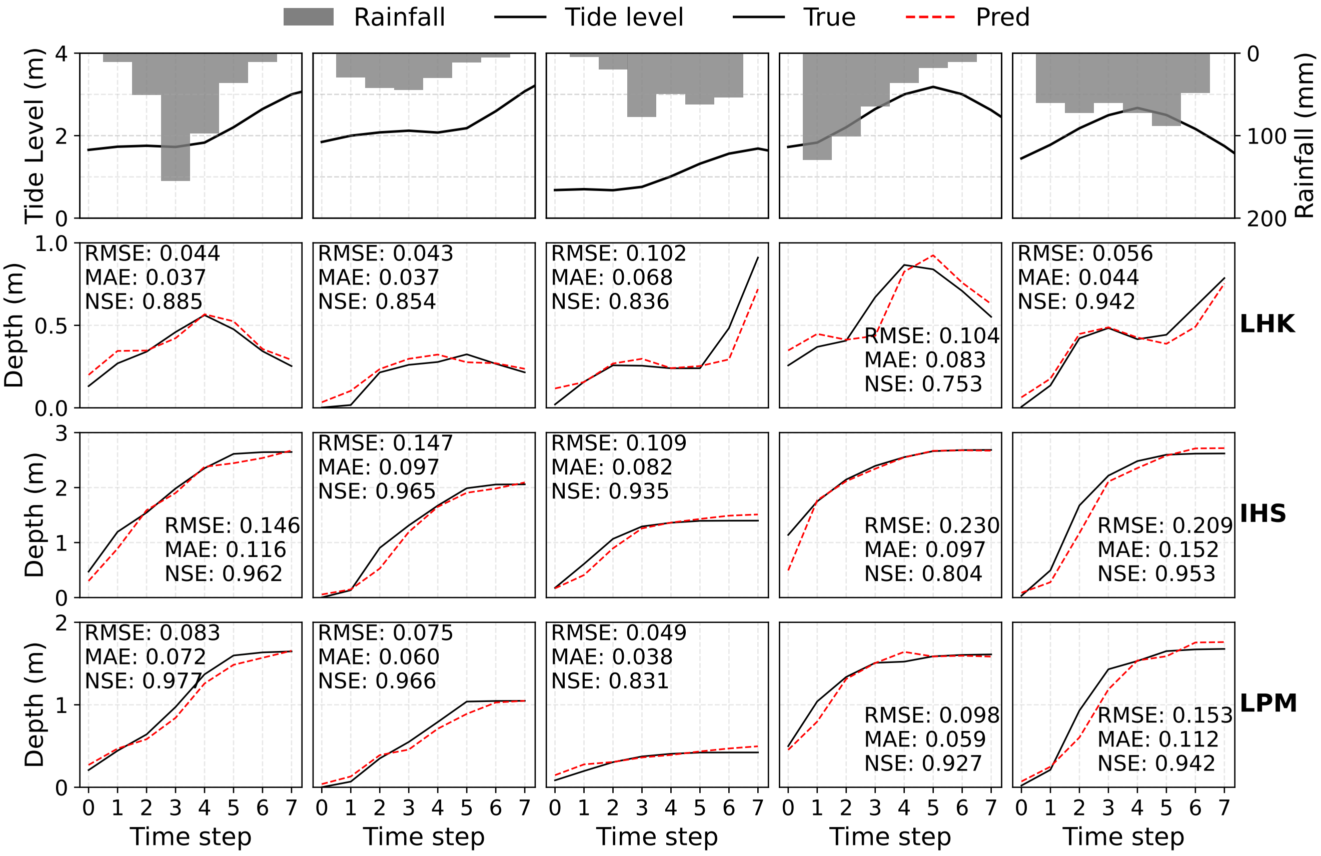

Figure 7. Comparison of water water depth processes in three flood-prone locations.

4.2 Performance of data-driven model in simulating water depth spatiotemporal dynamics

The mean values of NSE, RMSE, and MAE across the study area were 0.83, 0.08, and 0.05, respectively, demonstrating the 300 efficacy of the proposed data-driven model in simulating inundation processes from a basin-wide perspective. In addition, CSI was recorded as 0.83, indicating that the model detects the presence of flooding in the study area efficiently. In order to exhibit the model's efficacy in capturing the spatiotemporal dynamics of inundation water depths, the mean NSE, RMSE, and MAE of the simulated water depth for all rainfall events, along with the box-plot of these three metrics across each grid cell within the study area, are depicted in Figure 8 and Figure 9. Figure 8 demonstrates that the majority of regions exhibited mean NSE values exceeding 0.7, with RMSE and MAE consistently below 0.10, indicating robust model performance in simulating spatiotemporal inundation depths. Regions exhibiting significant discrepancies between the proposed CNN-ConLSTM model and the physics-based model are predominantly localized in narrow inter-building zones, highlighting the model's limited capability to resolve inundation processes in areas with abrupt topographic gradients or complex urban microtopography. Figure 9 demonstrates that the proposed data-driven model attained 75% of the NSE values exceeding 0.80, along with 75% of the RMSE and MAE values remaining below 0.10 for all simulated inundation events. This demonstrates the robust capability of the model in simulating inundation dynamics across diverse locations and rainfall events. It should be noted that grids with water depths less than 0.20 m were ignored when evaluating the model, since the proposed model focuses on flooding simulation.

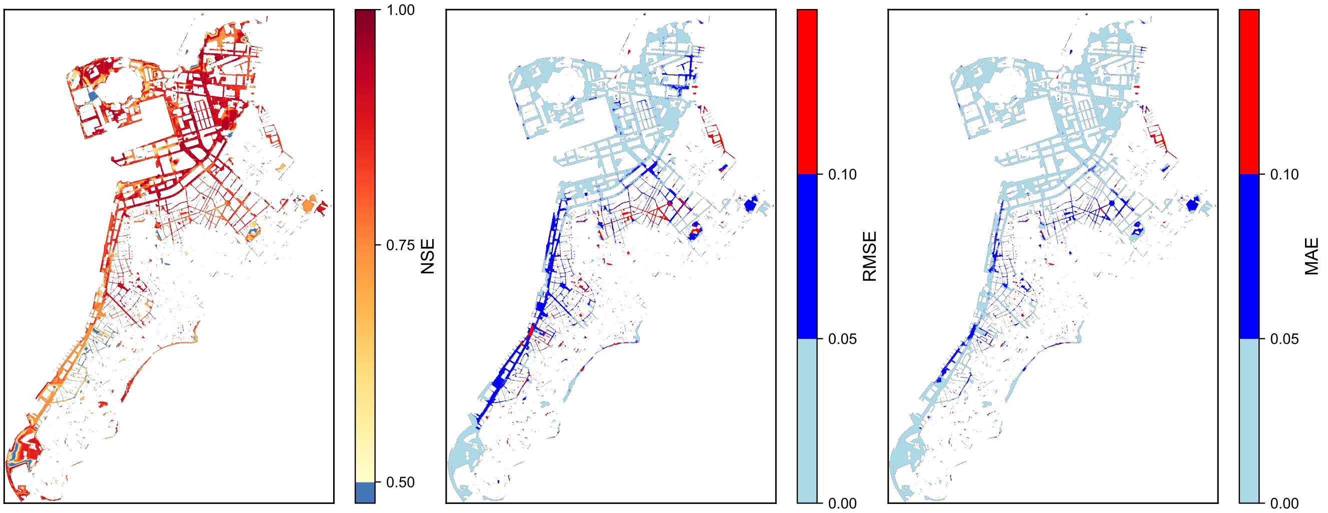

Figure 8. The spatial distribution of the mean NSE, RMSE, and MAE. The blank areas on the map indicate regions with no water depth.

4.3 Performance of data-driven model in maximum inundation water depth simulation

The maximum inundation depth is acknowledged as a crucial metric to assess the severity of urban flooding. The absolute and relative bias between the maximum inundation depth predicted from the proposed data-driven model and the physic-based model is shown in Figure 10. The figure illustrates that the majority of regions demonstrated an absolute bias in maximum inundation depth of less than 0.10 meters, with the corresponding relative bias remaining under 5%. This indicates that the model effectively captures flood peaks and therefore can be applied to predict extreme urban flooding events. The distribution of bias in maximum inundation depth aligns with other evaluative metrics such as NSE and RMSE. In particular, a greater bias

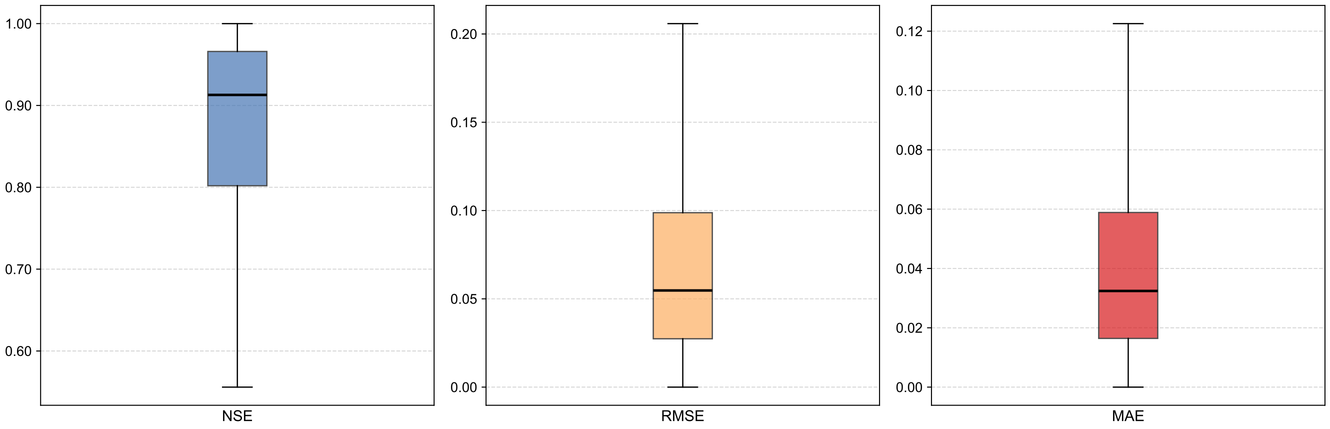

Figure 9. The boxplots of NSE, RMSE, and MAE for all grids across the study area under all rainfall events.

is observed in regions characterized by sudden topographic variations, such as the peripheries of structures and zones of terrain transition. To explicitly visualize the maximum depth discrepancies in inundation depth, Figure 10 presents absolute values of absolute and relative bias. Across all grid cells and inundation events, 41.3% of the bias exhibited positive deviations while 58.7% showed negative deviations. This relative balanced distribution suggests that the proposed data-driven model does not have a systematic bias toward overestimation or underestimation when compared to physics-based models.

4.4 Computational efficiency

The assessment of computational efficiency between the proposed data-driven model and the physics-based model was carried out on an identical computational platform equipped with an Intel Core i9-14900K CPU (24 core, 32 thread) and an NVIDIA RTX 4090 GPU (24 GB VRAM). To minimize the influence of stochastic errors, the mean computation time was compared for both the data-driven model and the physics-based model in all rainfall scenarios. The physics-based model was executed utilizing the CPU, whereas the data-driven model benefited from GPU acceleration. The physics-based model required an average runtime of 16,200 seconds per simulation, while the pre-trained data-driven model achieved GPU-accelerated inference times of 4 seconds per prediction, demonstrating a 4,000× speed advantage post-training. Despite the significant initial investment associated with the high efficiency of the data-driven model, the findings suggest that it is more suitable for real-time prediction, as the training phase can be carried out during dry periods. While physics-based models are capable of obtaining computational acceleration through GPU-based parallel processing, the practical implementation of such optimizations in 1D-2D coupled hydrodynamic models continues to pose significant challenges. Present research on the utilization of GPU acceleration within the realm of physics-based models primarily concentrates on standalone 2D hydrodynamic modules, where speedups reported typically achieve an order of magnitude (e.g., 100×). This represents a substantially lower level of computational efficiency compared to the gains evidenced by data-driven methodologies.

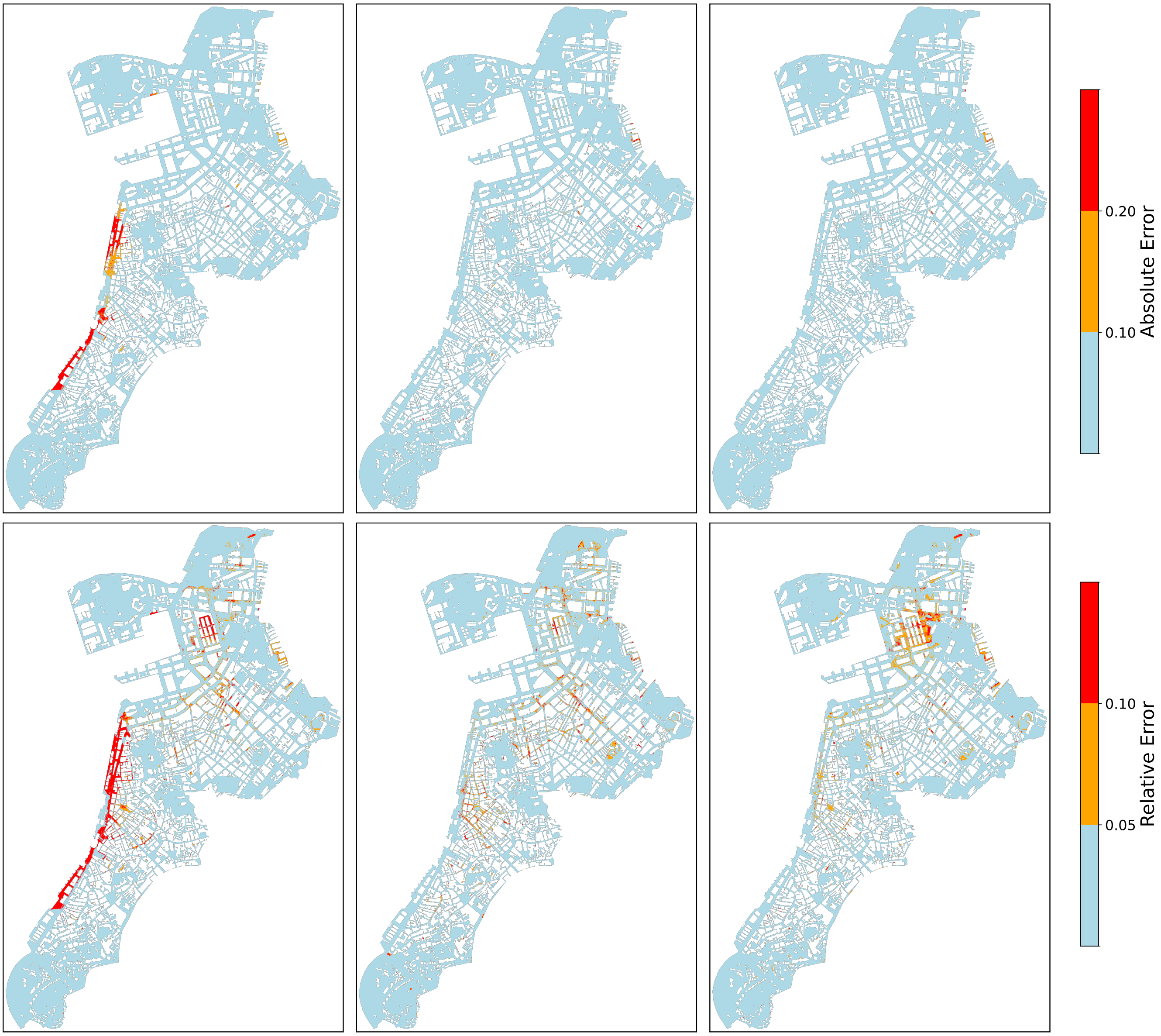

Figure 10. The absolute and relative bias of maximum inundation depths.

5 Discussion

Although the proposed data-driven model demonstrates close alignment with the physics-based counterpart in urban inundation simulations, discrepancies persist in regions characterized by abrupt topographical variations, such as the peripheries of buildings. Several factors may account for this issue. The buildings in the study area are densely distributed, with many areas exhibiting small gaps between structures. The spatial fidelity of the terrain representation of the study area is constrained by 345

DEM resolution limitations, where abrupt elevation transitions spanning merely 1-2 pixels induce high-frequency terrain variations. This undersampling effect creates insufficient spatial gradients in input features, preventing the model from resolving critical discontinuities and systematically concentrating prediction errors in topographically complex zones. In addition, both CNN and ConvLSTM architectures exhibit inherent smoothing effects, as their localized convolutional kernels prioritize continuous spatial or temporal feature extraction. This mechanism suppresses abrupt transitions in edge zones or high-frequency variation regions by outputting compromised intermediate values rather than preserving sharp discontinuities, resulting in over-smoothed predictions. Furthermore, the proposed data-driven model integrates the ConvLSTM-based branch with the CNN-based branch through a straightforward concatenation operation at the final layers. This configuration may result in less effective integration between the static spatial features extracted by the CNN branch and the dynamic temporal processes modeled by the ConvLSTM branch, thus leading to the potential neglect or attenuation of critical spatial information during temporal modeling. Moreover, the physics-based model incorporates localized topographic reconstruction at the peripheries of buildings to address microscale flow discontinuities, a refinement mechanism absent in machine learning methodologies. This deficiency in assimilating such localized terrain optimizations inherently limits the capacity of data-driven models to accurately resolve hydrodynamic processes in critical edge zones, where microscale elevation variations dominantly govern flow bifurcations. The loss function used further exacerbated significant errors near building peripheries, as it exhibits greater sensitivity to spatially extensive low-magnitude errors compared to localized extreme errors at boundaries. This intrinsic bias inherently prioritizes optimization of the global averaged performance during training, thus compromising accuracy in the localized edge regions. In the further study, the proposed data-driven model can be enhanced through three key optimizations: (1) increasing the sampling density of building-edge regions in training datasets to address underrepresented topographical discontinuities, (2) integrating edge-aware convolutional architectures (e.g., boundary-attention modules) to explicitly prioritize high-gradient zones, and (3) embedding physics-based constraints (e.g., mass conservation principles) to regularize predictions near artificial structures.

Although deep learning models demonstrate considerable capabilities across various fields, they are frequently regarded as black-box models due to their complex architectures, which impede the direct interpretation of the relationships between inputs and outputs. Compared to physical models, their reliability in practical applications is subject to more frequent scrutiny. In recent years, the analysis of the relationships between inputs and outputs of deep learning models using mathematical methods has become an important research direction in the field of deep learning-based hydrological analysis (Liu et al., 2024; Huang et al., 2023). However, due to the multidimensional nature of the inputs and outputs of this study, as well as the complexity of the model architecture, there is currently no effective method for conducting a global interpretability analysis of the proposed model. In this study, we just analyzed the impacts of dynamic inputs on predicted water depths at a specific site LHK with the SHAP method. The mean absolute SHAP value of the dynamic features is shown in Figure 11. The figure illustrates that the inundation depth from the preceding timestep exerts the most substantial influence on the forecasted inundation depth. Additionally, the rainfall from the two preceding timesteps and the tidal level from the preceding timestep also play a significant role in influencing the predicted inundation depth. While the interpretability of static features was not specifically analyzed in this study, the selection of input variables was grounded in their demonstrated physical and statistical relevance to inundation processes, as identified in prior research (Gao et al., 2024). This approach seeks to mitigate interpretability issues that may arise from redundant feature dimensions to some extent. Future research should prioritize advancements in interpretability assessment frameworks for multidimensional input problems to enhance the credibility of data-driven models in hydrological applications.

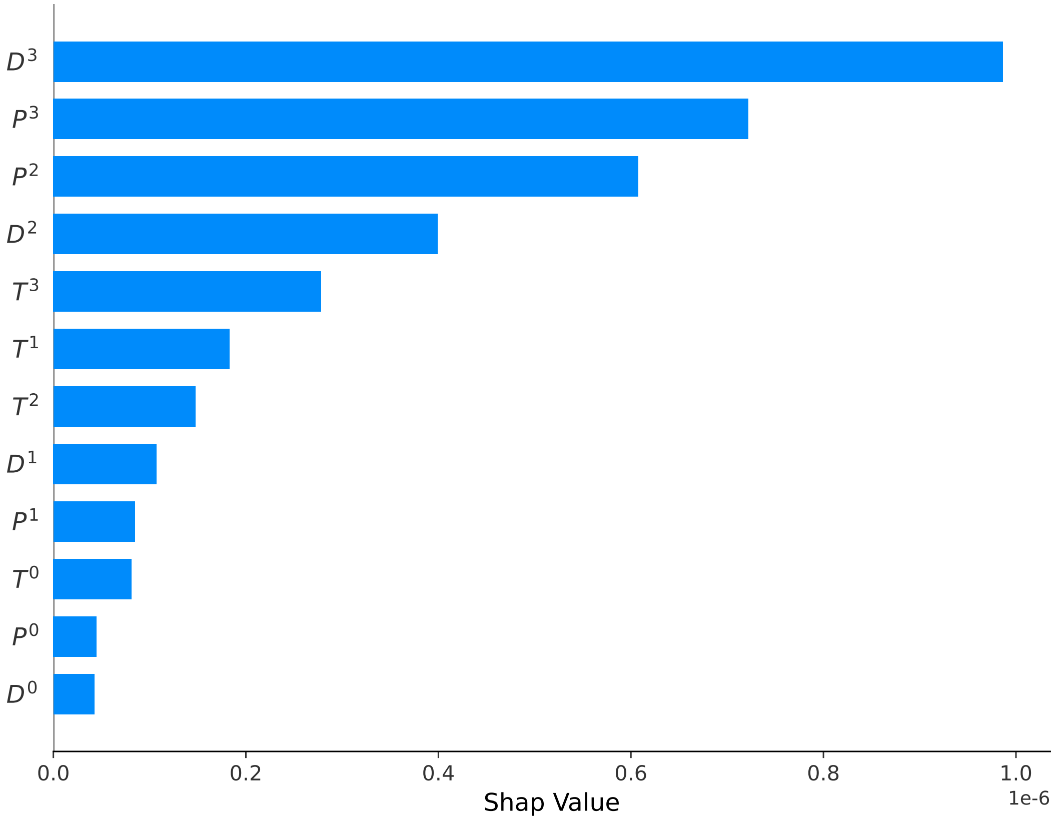

Figure 11. SHAP values of dynamic features. P^t, D^t, and T^t represent rainfall, inundation depth, and tide level at time step t, respectively.

385 The capacity for generalization serves as the primary criterion for evaluating the efficacy of data-driven models. The generalization of the proposed CNN-ConvLSTM model was enhanced by the efficient integration of static geospatial data through the use of a parallel concatenation operation and a tiling training strategy. In this research, the generalizability of the proposed model was substantiated by conducting training on small tiles and subsequently evaluating the model on the entire study area. In the future, the model will be used in diverse regions to further assess and validate its generalizability.

6 Conclusions

In this study, we proposed a novel deep learning model to predict the spatiotemporal distribution dynamics of urban inundation depths. The model comprises two distinct branches: a ConvLSTM-based branch and a CNN-based branch, which are amalgamated through a concatenation operation. The ConvLSTM-based branch extracts information from temporal input sequences, while the CNN-based branch captures static geospatial features. A tiling strategy was implemented during model training, partitioning the study area into spatially discrete sub-regions to serve as independent training samples, thereby enhancing generalization capability across heterogeneous terrain configurations. The proposed model was applied in a flood prone area of Macao and compared with a physics-based model. The results show that: (1) the proposed model effectively captures the dynamics of water depth in flood-prone locations, with NSE >0.80 for the majority events, as well as RMSE and MAE values <0.20. (2) The model demonstrates a high degree of efficiency in detecting flooding within the study area, as evidenced by a CSI value of 0.83. (3) The proposed data-driven model demonstrates robust generalization performance, with simulated inundation processes closely aligned with the results of the physics-based model in most regions (mean NSE >0.70, RMSE <0.10, MAE <0.10). Notable discrepancies persist only in localized zones of abrupt terrain variations, particularly near building edges.

Author contributions. LWQ and GXC were involved in the conceptualisation and methodology of the project. LWQ was responsible for model development, with guidance from GXC, LWH, DLR, and GK. LWQ also ran the model simulations and analysed the results under the supervision of GXC and LJH. Data visualisation was carried out by LWQ. DLR provided the original model data. The original draft of the paper was prepared by GXC, with contributions from LWQ, PJ and DLR. All authors were involved in reviewing and editing the manuscript.

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

Acknowledgements. The research is financially supported by the National Key R&D Program of China (2022YFE0205200) and the National Natural Science Foundation of China (52209044).

References

Aderyani, F. R., Jafarzadegan, K., and Moradkhani, H.: A surrogate machine learning modeling approach for enhancing the efficiency of urban flood modeling at metropolitan scales, Sustainable Cities and Society, 123, 106 277, 2025.

Ahmad, R., Yang, B., Ettlin, G., Berger, A., and Rodríguez-Bocca, P.: A machine-learning based ConvLSTM architecture for NDVI forecasting, International Transactions in Operational Research, 30, 2025–2048, 2023.

Anastasiou, K. and Chan, C.: Solution of the 2D shallow water equations using the finite volume method on unstructured triangular meshes, International Journal for Numerical Methods in Fluids, 24, 1225–1245, 1997.

Balaian, S. K., Sanders, B. F., and Abdolhosseini Qomi, M. J.: How urban form impacts flooding, Nature Communications, 15, 6911, 2024.

Bentivoglio, R., Isufi, E., Jonkman, S. N., and Taormina, R.: Rapid spatio-temporal flood modelling via hydraulics-based graph neural networks, Hydrology and Earth System Sciences, 27, 4227–4246, 2023.

Berkhahn, S., Fuchs, L., and Neuweiler, I.: An ensemble neural network model for real-time prediction of urban floods, Journal of Hydrology, 575, 743–754, https://doi.org/10.1016/j.jhydrol.2019.05.066, 2019.

Beven, K.: Robert E. Horton's perceptual model of infiltration processes, Hydrological processes, 18, 3447–3460, 2004.

Chen, Z., Yin, L., Chen, X., Wei, S., and Zhu, Z.: Research on the characteristics of urban rainstorm pattern in the humid area of Southern China: A case study of Guangzhou City, International Journal of Climatology, 35, 4370–4386, https://doi.org/10.1002/joc.4294, 2015.

Dai, W. and Cai, Z.: Predicting coastal urban floods using artificial neural network: The case study of Macau, China, Applied Water Science, 11, https://doi.org/10.1007/s13201-021-01448-8, 2021.

Dong, L., Liu, J., Zhou, J., Mei, C., Wang, H., Wang, J., Shi, H., and Nazli, S.: The influence of astronomical tide phases on urban flooding during rainstorms: Application to Macau, Journal of Hydrology: Regional Studies, 56, https://doi.org/10.1016/j.ejrh.2024.101998, 2024.

Fu, G., Zhang, C., Hall, J. W., and Butler, D.: Are sponge cities the solution to China's growing urban flooding problems?, Wiley Interdisciplinary Reviews: Water, 10, e1613, 2023.

Gao, W., Liao, Y., Chen, Y., Lai, C., He, S., and Wang, Z.: Enhancing transparency in data-driven urban pluvial flood prediction using an explainable CNN model, Journal of Hydrology, 645, https://doi.org/10.1016/j.jhydrol.2024.132228, 2024.

Gülbaz, S., Boyraz, U., and Kazezyılmaz-Alhan, C. M.: Investigation of overland flow by incorporating different infiltration methods into flood routing equations, Urban Water Journal, 17, 109–121, 2020.

Hou, J., Zhou, N., Chen, G., Huang, M., and Bai, G.: Rapid forecasting of urban flood inundation using multiple machine learning models, Natural Hazards, 108, 2335–2356, 2021.

Huang, F., Zhang, Y., Zhang, Y., Shangguan, W., Li, Q., Li, L., and Jiang, S.: Interpreting Conv-LSTM for spatio-temporal soil moisture prediction in China, Agriculture, 13, 971, 2023.

Liu, L., Liang, X., Xu, Y.-P., Guo, Y., Wang, Q. J., and Gu, H.: Enhanced rainfall nowcasting of tropical cyclone by an interpretable deep learning model and its application in real-time flood forecasting, Journal of Hydrology, 644, 131 993, 2024.

Lu, M., Jin, C., Yu, M., Zhang, Q., Liu, H., Huang, Z., and Dong, T.: MCGLN: A multimodal ConvLSTM-GAN framework for lightning nowcasting utilizing multi-source spatiotemporal data, Atmospheric Research, 297, 107 093, 2024.

Löwe, R., Böhm, J., Jensen, D. G., Leandro, J., and Rasmussen, S. H.: U-FLOOD – Topographic deep learning for predicting urban pluvial flood water depth, Journal of Hydrology, 603, https://doi.org/10.1016/j.jhydrol.2021.126898, 2021.

Piadeh, F., Behzadian, K., Chen, A. S., Campos, L. C., Rizzuto, J. P., and Kapelan, Z.: Event-based decision support algorithm for real-time flood forecasting in urban drainage systems using machine learning modelling, Environmental Modelling & Software, 167, 105 772, 2023.

Rossman, L. A. and Huber, W.: Storm water management model reference manual volume II–hydraulics, US Environmental Protection Agency: Washington, DC, USA, 2, 190, 2017.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting, Advances in neural information processing systems, 28, 2015.

Wang, Y., Li, C., Liu, M., Cui, Q., Wang, H., Lv, J., Li, B., Xiong, Z., and Hu, Y.: Spatial characteristics and driving factors of urban flooding in Chinese megacities, Journal of Hydrology, 613, 128 464, 2022.

Wang, Z., Chen, Y., Zeng, Z., Chen, X., Li, X., Jiang, X., and Lai, C.: A tight coupling model for urban flood simulation based on SWMM and TELEMAC-2D and the uncertainty analysis, Sustainable Cities and Society, 114, 105 794, 2024a.

Wang, Z., Lyu, H., Fu, G., and Zhang, C.: Time-guided convolutional neural networks for spatiotemporal urban flood modelling, Journal of Hydrology, 645, 132 250, 2024b.

Yang, F., Ding, W., Zhao, J., Song, L., Yang, D., and Li, X.: Rapid urban flood inundation forecasting using a physics-informed deep learning approach, Journal of Hydrology, 643, 131 998, 2024.

Zahura, F. T., Goodall, J. L., Sadler, J. M., Shen, Y., Morsy, M. M., and Behl, M.: Training Machine Learning Surrogate Models From a High-Fidelity Physics-Based Model: Application for Real-Time Street-Scale Flood Prediction in an Urban Coastal Community, Water Resources Research, 56, https://doi.org/10.1029/2019WR027038, 2020.

Zhang, J., Zheng, Y., and Qi, D.: Deep spatio-temporal residual networks for citywide crowd flows prediction, in: Proceedings of the AAAI conference on artificial intelligence, vol. 31, 2017.

Zhang, R., Li, Y., Chen, T., and Zhou, L.: Flood risk identification in high-density urban areas of Macau based on disaster scenario simulation, International Journal of Disaster Risk Reduction, 107, https://doi.org/10.1016/j.ijdrr.2024.104485, tide level data macau, 2024.