Response to reviewer 1

This manuscript establishes a highly generalizable urban flood simulation model by combining ConvLSTM and CNN, which enables fast and accurate simulation of flood processes in the urban areas. The model structure is scientifically designed and can effectively learn the features of different types of data. The simulation results are promising, providing valuable insights and practical significance for understanding the spatiotemporal processes of urban flooding.

Although the manuscript is already quite comprehensive at its current stage, there are still several suggestions that could help make it even more polished.

The specific suggestions are as follows:

1. Training time

While the manuscript discusses the inference time of the trained model, it does not mention the time required for training the model.

Reply: Thank you for your comments. We appreciate this valuable comment. The training time of the proposed hybrid CNN–ConvLSTM model has now been added in Section 4.5. On an NVIDIA RTX 4090 GPU (24 GB VRAM), the model required approximately 5.2 hours to complete 60 epochs, whereas the CNN-only and ConvLSTM-only baselines took 4.7 hours and 4.8 hours, respectively. This information helps readers understand the computational efficiency of different architectures.

2. Typographical error

In “Figure 7. Comparison of water water depth processes in three flood-prone locations”, the word “water” appears twice, which seems to be a typographical error.

Reply: Thank you for your comments. The duplicated word “water” in Figure 7(now is Figure 8) has been corrected to “Comparison of water depth processes in three flood-prone locations.”

3. Relative error in Figure 10

The relative error shown in Figure 10 should be expressed with a percentage sign.

Reply: Thank you for your comments. We have already added the percentage sign in Figure 10(now is Figure 11).

4. Units in Figure 8

The RMSE and MAE values presented in Figure 8 are dimensional quantities, but the specific units are missing.

Reply: Thank you for your comments. The units of RMSE and MAE in Figure 8(now is Figure 9) have now been explicitly indicated as meters (m), since both metrics quantify the deviation between simulated and observed water depths. The updated figure caption and axis labels have been revised accordingly in the revised manuscript.

5. Description of stations

Section 4.1 analyzes the results of three stations, but the manuscript lacks descriptions, figures, or maps to indicate the locations of these stations.

Reply: Thank you for your comments. We have added the locations of the three analyzed stations (LHK, IHS, and LPM) in Figure 1 to provide a clearer understanding of their spatial distribution within the study area.

6. Data split methodology

In Section 3.3.1, it is stated that 80% of the dataset was used for training and 20% for testing. However, the manuscript does not clarify how this split was conducted. Was it an 8:2 split by rainfall events, or based on the constructed dataset as a whole? If it is the latter, there is a risk of data leakage from test events into the training set.

Reply: Thank you for your comments. In this study, the dataset was divided into 80% for training and 20% for testing based on rainfall events, please see line 283, rather than random sampling of the entire dataset, to prevent data leakage across temporal sequences. Each rainfall event was treated as an independent sample, and events selected for testing were completely excluded from the training process.

7. Performance under unseen scenarios

The manuscript does not discuss how the model performs when applied to unseen scenarios. It is suggested to include analysis related to underfitting or overfitting to better illustrate the model's reliability and limitations.

Reply: Thank you for your comments. The model adopts a tiling-based training strategy, which partially reflects the physical reality that individual urban drainage units function relatively independently during flood events, further improving the spatial generalization performance of the model. Due to data availability constraints, the model was applied and validated only within a relatively small study area in this research. However, the study area is a highly urbanized and topographically diverse region, encompassing both mountainous and low-lying plains as well as coastal boundaries. The external boundary conditions are also complex due to tidal and marine influences, making the study area highly representative of broader hydrological and hydrodynamic conditions. In addition, the model employed a tile size of 200x200m, and the selected study area was sufficiently large to provide an adequate number of tile samples. Therefore, using this area for model validation can be considered credible and representative. In future work, we will further collect flood datasets from cities with different characteristics to evaluate and validate the model's performance across diverse urban regions. In addition, benefiting from the parallel connection of the two encoders and the tiling-based training strategy, the proposed model can flexibly accommodate various types of dynamic and static data while exhibiting lower dependence on GPU memory. In the future, the model can be driven by real-time radar rainfall data and applied to urban flood nowcasting and real-time management tasks, providing technical support for intelligent and timely flood control operations.

Response to reviewer 2

Major comments

1. Study area and generalizability

The pilot area (4.06 km²) is small. Discuss scalability to larger, topographically complex basins and data requirements (computational cost, training data volume, transfer learning). Consider a short experiment or argument on tiling/patching strategies and edge effects.

Reply: Thank you for your comments. Due to data availability constraints, the model was applied and validated only within a relatively small study area in this research, which may limit the model's generalization ability. However, compared to prior ConvLSTM-based models in urban flood prediction, the proposed model employs two distinct encoders based on ConvLSTM and CNN architectures to process different types of input information. The ConvLSTM-based branch is designed to capture dynamic temporal features, such as rainfall time series, while the CNN-based branch extracts static geospatial characteristics, such as digital elevation model (DEM) data. The outputs of the two branches are then combined through a parallel concatenation operation, enabling the model to efficiently account for the joint influence of both dynamic and static factors on urban flood processes, thereby enhancing its spatiotemporal generalization capability. In addition, the model adopts a tiling-based training strategy, which partially reflects the physical reality that individual urban drainage units function relatively independently during flood events, further improving the spatial generalization performance of the model. Moreover, the study area is a highly urbanized and topographically diverse region, encompassing both mountainous and low-lying plains as well as coastal boundaries. The external boundary conditions are also complex due to tidal and marine influences, making the study area highly representative of broader hydrological and hydrodynamic conditions. The model employed a tile size of 200 x 200 m, and the selected study area was sufficiently large to provide an adequate number of tile samples. Therefore, using this area for model validation can be considered credible and representative. In future work, we will further collect flood datasets from cities with different characteristics to evaluate and validate the model's performance across diverse urban regions. (See lines 455-471.)

The supplementary experiment regarding the tiling strategy is presented as follows:
To enhance the spatial generalization capability of the model and reduce GPU memory requirements, the study area was divided into a series of square tiles, each of which was treated as an independent training sample. During the partitioning process, an overlap of less than 10% of the total extent of the study area was introduced in each respective direction to enhance the sample coverage and ensure smoother spatial continuity between adjacent tiles. The tiling strategy is physically reasonable, as urban 245 areas are typically organized into a series of drainage sub-catchments whose hydrological characteristics are primarily governed by internal factors such as local topographic deformations and drainage network

configuration. Different sub-catchments are relatively independent from one another, making this partitioning approach consistent with the physical structure of urban drainage systems. The tile size was determined by selecting the optimal configuration based on the performance of the model with different tile sizes. (See lines 244-252.)

Considering both the extent of the study area and the typical scale of urban drainage sub-catchments, we evaluated tile sizes ranging from 50 × 50 to 300 × 300 grid cells, with an interval of 50 cells between configurations. Each grid cell corresponds to a 2 × 2 m spatial resolution. As illustrated in Figure 7, this analysis examines how increasing the tile size influences model accuracy and stability. Overall, the simulations across all tile-size configurations performed well, achieving a mean NSE greater than 0.7, a MAE below 0.01, and a RMSE below 0.12. However, the average NSE exhibited a rising–falling trend, while the MAE showed the opposite declining–rising pattern, with both metricsreaching their optimal values when the tile size was 100 × 100 grid cells. Accordingly, this configuration was adopted for subsequent experiments. These results indicate that tile size exerts a measurable influence on model performance. The observed variation can be attributed to edge effects between adjacent tiles: When the tile size approximates the typical scale of the smallest urban drainage sub-catchment, each tile functions relatively independently, with limited intertile flux exchange. As a result, boundary effects are minimized and simulation accuracy is enhanced. (See lines 306-316.)

2. Model comparison

Consider including an ablation study: ConvLSTM-only vs. CNN-only vs. the proposed hybrid, to demonstrate the incremental benefit of CNN–ConvLSTM coupling.

Reply: Thank you for your comments. The ablation study has been supplemented as follows: To demonstrate the effectiveness of the CNN-ConvLSTM coupled architecture proposed in this study, we compared the simulation results of the CNN-only model, the ConvLSTM-only model and the hybrid model developed in this study. Both the CNN-only and ConvLSTM-only models underwent hyperparameter optimization to determine their respective optimal network configurations. Through hyperparameter optimization, the final CNN-only model was configured with a four-layer architecture, while the ConvLSTM-only model adopted a two-layer architecture. The mean NSE, mean RMSE, and mean MAE values of the three models within the study area are summarized in Table 4. As shown in the table, the hybrid model outperformed the other two models in all evaluation metrics, achieving a higher average NSE and lower average RMSE and MAE. (See lines 404-410).

3. Depth of analysis and scientific insight

In Results, please move beyond “figure + short captioned numbers.” For example, for those key evaluation metrics (e.g., NSE, RMSE, MAE), discuss why performance varies across rainfall events. Where performance is very high (e.g., NSE > 0.95) or notably lower (e.g., NSE < 0.80), probe the hypothesized mechanisms and provide in-depth explanations.

Reply: Thank you for your comments. The explanations have been added in lines 327-331, as detailed below:
The relatively lower NSE value (0.75) occurred under the combined condition of high tide levels and Pattern I heavy rainfall events. The training dataset primarily consisted of observed rainfall–tide combinations, supplemented by a small number of designed scenarios. Since the specific combination of high tide and Type I rainfall accounted for only a small proportion of the training samples, the model exhibited relatively poorer simulation performance under this condition.

Consider adding a brief uncertainty or robustness check (e.g., event-level cross-validation, bootstrapped confidence intervals, or sensitivity analysis).

Reply: Thank you for your comments. The k-fold event-level cross-validation has been added in lines 395-402, as detailed below:
The robustness of the proposed hybrid model was evaluated using k-fold event-level cross-validation. All rainfall–tide combination events were divided into ten subsets. In each iteration, nine subsets were used for training and one for validation, rotating sequentially until all subsets had served as the validation set once. The performance metrics obtained from the ten training iterations are summarized in Table 3. As shown in the table, the mean NSE was 0.76, with a maximum of 0.80, a minimum of 0.72, and a standard deviation of 0.03. These results indicate that the model can consistently and reliably capture the dynamic patterns of various inundation events across all grid cells within the study area. Similarly, both RMSE and MAE remained at consistently low levels with narrow variability (mean RMSE = 0.112 ± 0.013, mean MAE = 0.088 ± 0.013), suggesting that the model can reliably reproduce water depth variations across the study area.

In 4.1, the authors analyze five "randomly selected" rainfall events. I am curious about the other 7 events, maybe provide an overall table (mean/median NSE, RMSE, MAE; ranges) and show distribution across events.

Reply: Thank you for your comments. The new added robustness analysis indicates that the model performs well and consistently.

4. Deepen discussion

Discussing limitations of deep learning for flood modeling is beneficial; this section is more important to: (1) Articulate the paper's contributions relative to prior DL flood emulation/forecasting work (what's new about your integration?). (2) Position your work against related literature (including ConvLSTM-based flood studies such as Liao et al., 2025, WRR) and highlight similarities/differences and added value. (3) Offer practical implications (e.g., real-time forecasting potential, co-design with drainage management) and clear future directions (e.g., larger regions, real rainfall radar, DEM/land-use multi-scale features).

Reply: Thank you for your comments. The discussion has been expanded. The paper's

contributions are shown in lines 57-73 of the Introduction and lines 457-465 of the Discussion. The practical implications are discussed in lines 471-475.

lines 57-73:

Liao et al. (2025) proposed a ConvLSTM-based architecture that explicitly captures the spatiotemporal distribution of rainfall for flood prediction and compared its performance against that of a 3D CNN model. 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.

lines 457-465:

However, compared to prior ConvLSTM-based models in urban flood prediction (Liao et al., 2025), the proposed model employs two distinct encoders based on ConvLSTM and CNN architectures to process different types of input information. The ConvLSTM-based branch is designed to capture dynamic temporal features, such as rainfall time series, while the CNN-based branch extracts static geospatial characteristics, such as digital elevation model (DEM) data. The outputs of the two branches are then combined through a parallel concatenation operation, enabling the model to efficiently account for the joint influence of both dynamic and static factors on urban flood processes, thereby enhancing its spatiotemporal generalization capability. In addition, the model adopts a tiling-based training strategy, which partially reflects the physical reality that individual urban drainage units function relatively independently during flood events, further improving the spatial generalization performance of the model.

lines 471-475:

In addition, benefiting from the parallel connection of the two encoders and the tiling-based training strategy, the proposed model can flexibly accommodate various types of dynamic and static data while exhibiting lower dependence on GPU memory. In the future, the model can be driven by real-time radar rainfall data and applied to urban flood nowcasting and

real-time management tasks, providing technical support for intelligent and timely flood control operations.

Minor comments:

1. Figure 1b. Add elevation units to the legend.
Reply: Thank you for your comments. We have already added the elevation units ("m") in Figure 1b.

2. In 3.2.1, add citations for CNN and a bit more information about CNN.
Reply: Thank you for your comments. Additional background information about CNNs has been added to Section 3.2.1, along with representative references to foundational studies (LeCun et al., 1998; Krizhevsky et al., 2012) and recent hydrology-related applications (Sit et al., 2020). These additions clarify the theoretical basis and illustrate how CNNs capture spatial dependencies and heterogeneity in terrain-related features within flood-modeling contexts.

3. In 3.3.2, the authors mention the correlation coefficient (CC), but CC is absent from the Results. Either (i) report CC in the text/plots/tables, or (ii) remove it from 3.3.2 and justify its exclusion.
Reply: Thank you for your comments. We have decided to remove the correlation coefficient (CC) from Section 3.3.2, and the correlation coefficient (CC) has been replaced with the Nash–Sutcliffe Efficiency (NSE) in Section 3.3.2 for consistency with the evaluation metrics presented in the Results. This change also better aligns with common practices in hydrological model performance assessment.
4. Figure 4. The flowchart is not sufficiently informative. Clarify the concatenation between ConvLSTM and CNN components. Also, replace "…" with the complete set of data used to provide a more intuitive understanding for readers.
Reply: Thank you for your comments. The flowchart has been updated to include all data categories and to explicitly indicate that the ConvLSTM and CNN modules are connected via a concatenate operation.
5. Figure 7. Add column titles such as "Event 1" … "Event 5" for immediate readability. Clarify that "True" = physics-model simulation and state this consistently in the caption and text.
Reply: Thank you for your comments. Column titles ("Event 1" – "Event 5") have been added. The term "True" is now explicitly defined as "physics-model simulation" in the main text.
6. Figures 8 & 9. Consider combining them into a single figure: e.g., first row = former Fig. 8, second row = former Fig. 9, to streamline reading. Make titles explicit: indicate these are metrics for inundation water depth (e.g., "NSE of Inundation Depth," etc.). For the boxplot summaries, also split by locations (LHK, HIS, LPM) and all grids, yielding four boxplots for each metric (NSE, RMSE, MAE).
Reply: Thank you for your comments. We have combined the previous Figures 8 and 9 into a single figure for conciseness and easier comparison. The top row now presents the spatial distribution of the mean NSE, RMSE, and MAE for inundation depth, while the bottom row displays the corresponding boxplots at three representative stations (LHK, HIS, and LPM) and for all grids. Explicit titles ("NSE of Inundation Depth," "RMSE of Inundation Depth (m)," and

"MAE of Inundation Depth (m)") have also been added to clarify that these metrics refer to simulated inundation water depths.

7. In 4.3 & Figure 10 (bias vs. error). If the text uses "absolute and relative bias," define them in 3.3.2 and use the same terms in figures. If you actually plot errors, rename to "absolute error" and "relative error." Standardize relative error bins (e.g., 5%, 10%, ···) and include the % unit in axes/legends. Ensure caption and main text use the same terminology.

Reply: Thank you for your comments. Percentage signs (%) are now included on legends. We also renamed all occurrences of "bias" to "error" where maps plot errors, using "absolute error (m)" and "relative error (%)" consistently in Section 4.3 and Figure 10. The relative-error bins are standardized to 5%, 10%. The figure caption, axes, and main text now use identical terminology.