Dear Reviewers,

Thank you for reviewing our manuscript. The comments have helped us strengthen the methodology presentation and clarify several points in the original text. We address each concern below; all corresponding revisions have been incorporated into the updated manuscript.

Below are our detailed, point-by-point responses to your major concerns (Part I) and specific comments (Part II).

Part I: Response to Major Comments

Comment 1. **The physics-based model that generated the hydrological responses is not fully presented or validated. This is important because if the physical model does not adequately represent the complexity of the case study, the study loses much of its meaning.**

Response: We agree — the synthetic data is only as reliable as the physical model behind it. In the updated manuscript, Section 2.2 now presents the InfoWorks ICM model with expanded detail and validation. We note that a physically based model of this catchment necessarily involves simplifications, but the validation results below demonstrate that these do not compromise the fidelity of the generated synthetic dataset.

1) The revised Section 2.2 includes a detailed overview of the 6,500 m^2 study area, covering buildings, squares, green spaces, roads, and drainage pipe networks.

2) We supplemented the initial calibration (Fig. 3) with a new validation phase covering three additional typical observed rainfall events (Table 2).

3) These supplementary validations yielded Nash-Sutcliffe Efficiency (NSE) values of 0.82, 0.75, and 0.88, demonstrating that the model accurately captures the dynamics of the hydrosystem.

4) To directly validate the LSTM model against physical reality, we added Fig. 13, which compares the inundation area hydrographs simulated by the InfoWorks ICM model and the LSTM model under real rainfall events.

5) Quantitative analysis of these real events reveals peak error percentages of approximately 3.8%, NSE values exceeding 0.8, and R^2 values above 0.9.

Comment 2. The simulated rainfall is very unrealistic (100 mm/h vs. 1 mm/h observed), and events all have the same duration.

Response: We apologize for the confusion regarding the rainfall units in the original manuscript. The observed rainfall measurements were presented in mm/min, which led to the apparent discrepancy when compared to the simulated design storms in mm/h. The updated manuscript now uses consistent units throughout. We recognize that these intensities are high, but they reflect the short-duration, extreme storm patterns characteristic of urban pluvial flooding in this region.

1) The revised text clarifies that the observed storm intensities are consistent with the high-intensity nature of urban pluvial floods in the study area.
2) The simulated storms utilize the Chicago design storm pattern (Eqs. 2–5), generating hydrographs for 1–10 year return periods specific to the region.
3) Regarding the fixed duration: Each rainfall event is strictly set to a 24-hour duration followed by a 6-hour recession period (30 hours total).
4) This controlled design isolates the specific effects of data length and rainfall intensity distribution on model performance. It prevents duration variations from introducing confounding factors into our statistical evaluation.

Comment 3. Justification for the LSTM model over a simpler Perceptron model.

Response: While a multilayer perceptron (MLP) is computationally lighter, short-duration, high-intensity urban flooding is a complex spatiotemporal nonlinear problem.

1) We selected the LSTM network because it is designed to decode temporal dependencies and sequential rainfall-inundation responses, which are critical for minute-level high-frequency hydrological data.
2) Our architecture uses a 30-minute sequence length and a 15-minute sliding window, capturing the dynamic response cycle of short-term rainfall-runoff.
3) Furthermore, we constrained the model to a single-layer LSTM with 64 hidden neurons.
4) This keeps training times efficient, averaging approximately 896 seconds on our hardware setup, addressing concerns about excessive energy consumption.

Comment 4. Lack of cross-validation for the LSTM model.

Response: This is an important point. We have restructured the evaluation framework to address it, as described below.

1) The updated manuscript now employs a 5-fold cross-validation strategy across all dataset configurations (Section 2.4, Figs. 8–9).
2) Data partitioning is conducted independently for each dataset configuration to prevent bias.
3) The results are reported as the mean and standard deviation across the five folds, providing robust performance metrics that rule out anomalies caused by random validation splits.
4) The corresponding box plots (Figs. 10–12) and data tables (Tables 6–8) confirm stable model performance across the folds.

Part II: Response to Specific Comments

1. The variability in maximum precipitation intensities is not representative (90mm/h to 140mm/h). Could start at 10 mm/h.

Response: The revised manuscript clarifies the categorization. In this study, the low-intensity training set consists of intensities of 90–120 mm/h, the high-intensity set spans 120–170 mm/h, and the mixed-intensity set spans 90–170 mm/h. These classifications are based on the historical threshold criteria for urban flooding in this catchment. Events starting at 10 mm/h in this area do not trigger meaningful surface inundation due to pipeline drainage capacity, which is why they are excluded from flood forecasting training. We agree that this range is narrow, but it reflects the actual flood-producing rainfall spectrum for this specific urban setting.

2.L128-135: Lack of information on the rainfall-runoff model, response time, and output hydrographs. Nash 0.5 may conceal mediocre simulations. Is there saturation for intense events?

Response: Supplementary validation for the physical model has been added to the revised manuscript. Table 2 now displays the observed vs. simulated peak flows, peak time errors, and NSE values for three independent events.

We also plotted the comparison of inundation area hydrographs (Fig. 13) under real rainfall events, demonstrating that the model accurately captures the rising limb,

peak phase, and recession limb, proving the system is not oversimplified.

3. Confusion between sequence length L171 and the LSTM sequence of 2 time steps (30 min).

Response: This nomenclature has been corrected throughout the revised manuscript. The fundamental sample unit is a 45-minute continuous sliding window sequence. The LSTM internally uses a 30-minute sequence length with a 15-minute step size. The "Dataset Length" variables (L1 to L6) refer to the *total count* of these sequence samples, ranging from 598 to 1,198.

4. L137: q is the intensity of the rain, poorly chosen; L139: specify figure number; L148: specify 'total precipitation'.

Response: The variables have been rewritten to align with standard hydrological nomenclature in the revised manuscript. Precipitation is now P_t, infiltration is I_t, drainage is D_t, and inundation area is Y_t.

Cumulative surface runoff volume is now defined mathematically via integration over time (Eq. 7) in the updated manuscript. $V(t) = A \int_0^t Q(\tau)\, d\tau$ All figure references have been corrected.

5.Problem with Eq 2: time does not appear.

Response: The equations have been updated in the revised manuscript to reflect time dependencies. The water balance equation for the net surface runoff rate is now defined at time t as $Q(t) = P(t) - I(t) - D(t)$.

6. L149: The output of the hydrological model called 'rainfall-runoff' is in fact a flooded area. This is an approximation.

Response: We agree and have included the exact mechanism mapping runoff to flooded area in the updated manuscript. Eq. 8 demonstrates how the instantaneous runoff rate is integrated into volume, then mapped to the inundation area Y(t) using the hydraulic conversion function established by the shallow water equations, DEM, surface slope, and depression storage. We note that this mapping, while an approximation, is standard practice in urban flood modeling and has been validated against observed events as shown in Fig. 13.

7. L156-166: We would like to have the Nash criteria for the model by category of experiments.

Response: The Nash-Sutcliffe Efficiency (NSE) values for the physical model validation have been added in Table 2 of the revised manuscript, ranging from 0.75 to 0.88.

8. L172-177: Specify duration in minutes and number of floods. At 90 mm per hour, it is difficult to consider the rain as 'light'.

Response: 1) We have specified in the revised manuscript that the total duration configurations encompass 5 to 10 independent storm events, corresponding to cumulative durations of 150 to 300 hours. 2) We have also refined our terminology; we now refer to these events as "low-intensity" relative to the extreme storm design thresholds of the study area, rather than universally "light" rain.

9. L178: How is the database divided into learning, testing and validation?

Response: The revised manuscript now employs a 5-fold cross-validation mechanism via stratified random sampling. For each fold, the remaining four folds serve as the training set, ensuring no data leakage. We chose stratified sampling over simple random splits because it better preserves the distribution of rainfall intensities across training and test sets.

10. L202-203: I do not understand how a sequence of two time steps can help approximate long-term memory.

Response: This was an error in the previous draft. The corrected manuscript now states that the sequence length is 30 time steps (representing 30 minutes, with a sampling frequency of 1 minute).

11. L209: Specify what helps avoid overfitting? There should be no overfitting on a synthetic system.

Response: Although the data are synthetic, the system output still presents complex nonlinear mappings and varied temporal dynamics depending on the design storm. In the updated manuscript we describe using a learning rate of 0.005, 50 epochs, and the Adam optimizer with weight decay to control variance. We recognize that overfitting is less common with synthetic data, but our goal was to ensure

reproducibility and avoid fitting to specific hydrograph shapes.

12. Table 1: Specify units of time step, sequence, and batch size.

Response: Table 1 has been renumbered as Table 5 in the revised manuscript. Table 5 now includes a batch size of 32.

2.5.1. L220: Why mention rainfall when target is flooded area? Ymean not X.

Response: The NRMSE formula (Eq. 11) has been corrected in the revised manuscript to reflect the variables: predicted Y_i, observed $\widehat{Y}_i$, and mean observed $\bar{Y}$ representing the inundation area.

14. 2.5.2: Express coefficient of determination using the same notation as NRMSE. Why is water depth mentioned?

Response: Eq. 12 has been updated in the revised manuscript to standardize the notation for R^2, and references to water depth have been replaced with "inundation area" to match the target variable.

2.6.1: Why are we talking about an effect? The concept should be defined.

Response: We rewrote this section in the revised manuscript to introduce a formal Multi-factor Analysis of Variance (MANOVA) framework. The updated text defines statistical main effects, interaction effects, significance testing (F-value, p-value), and effect size estimation (partial η^2 and ω^2).

3.1: Move entire section to the chapter where the model is discussed.

Response: Agreed. The physical model calibration and validation content has been moved to Section 2.2 in the revised manuscript.

Fig 6, 7, 8, 11, 12, 17 formatting and readability.

Response: All figures have been revised in the updated manuscript. We added proper scales and clearer colors. The confusing performance variance charts (formerly Fig 11/12) have been replaced with box plots showing 5-fold CV performance (Figs. 10–12). Figure 17 now labels axes and shows the linear growth of training time against data scale.

Line 561: "Controlled experiments with high-fidelity...". The physics-based model is poorly evaluated; the conclusion is unproven.

Response: With the additional physical model validation (Table 2 and the direct

comparison against real events in Fig. 13, yielding NSE > 0.8), the model's fidelity is now demonstrated in the revised manuscript. We recognize that no physical model is perfect, but these metrics support the use of the synthetic dataset for the subsequent LSTM experiments. The conclusions drawn from the synthetic data are therefore based on a more thoroughly characterized physical baseline.

Conclusion: "beyond approximately 14,400 sequences" no, it is 14,400 time steps.

Response: The revised manuscript now states that the threshold is 14,400 *samples* (where each sample is an extracted 45-minute time-series window).

We hope these revisions address your concerns. The updated manuscript reflects all changes discussed above. Thank you for your time and careful review.