Numerical Methods in Civil Engineering, 10-4 (2026) 90-107



Numerical Methods in Civil Engineering

Journal Homepage: https://nmce.kntu.ac.ir/

Hybrid CNN-LSTM surrogate modeling for rapid urban flood depth prediction

Suleiman Mohammad[*], Asokan Vasudevan[**], Mohammad Faleh Ahmmad Hunitie[***], Bekmurod Erdonov[****], Aziza Gafarova[*****], Bobir Turaev[******]

ARTICLE INFO

RESEARCH PAPER

Article history:
Received:
August 2026
Revised:
October 2026
Accepted:
October 2026

Keywords:
Surrogate model, flood risk, Hybrid machine learning, energy efficiency, CNN-LSTM

Abstract:

This study was designed to propose and validate an alternative data-driven approach for urban flood depth prediction using machine learning technology. The target variable of the model is the flood depth at each observation point, and the model produces a spatial flood depth map of the urban watershed. The developed approach relies on a hybrid CNN-LSTM model that takes into account both spatial (topography, land use, and distance to channel) and temporal (precipitation hydrograph) patterns. In the process of training the model, 850 HEC-RAS models were used for 12,800 observation points in the Chirchik River basin of northern Tashkent, Uzbekistan, totaling 10.88 million flood depth data points. However, based on the calculated values for the coefficient of determination (0.938) and Nash-Sutcliffe (0.921), it can be noted that the hybrid model demonstrated better results than the basic CNN and LSTM, by 6% and 11.4%, respectively. The false negative value of the hybrid model is 5.8%, which is twice lower than 13.5% in case of the standalone LSTM. The inference time is 0.85 seconds for generating a complete flood depth map of the entire watershed under a single scenario, compared to 55 minutes when using HEC-RAS, results in a speedup of 3,882 times. This inference time includes the forward pass of the network and the extraction of the predicted flood depths at the 12,800 observation points by bilinear interpolation. Preprocessing and postprocessing steps are excluded. In the case of transferability experiments, it has been shown that the performance of the transfer learning algorithm was 0.861 in the adjacent watershed after fine-tuning with just 50 flood simulations, while the fully trained model had a Nash-Sutcliffe value of 0.902. Sensitivity analysis shows the significance of the influence of the training dataset size and observation point density on the performance. In conclusion, this study presents a conceptual model for the prediction of flood depths in developing cities using physical modeling and machine learning. However, some of the shortcomings of this model include the fact that it is trained and evaluated at 12,800 fixed observation points, and its accuracy at arbitrary locations has not been independently validated; it is based on synthetic data from HEC-RAS rather than actual flood events; and the fact that its transferability has been tested for one nearby watershed with similar properties.

1. Introduction

Floods in cities, which are one of the most devastating natural hazards, have been causing considerable harm to infrastructure, economy, and human life every year [1].

The problem of urban floods has been aggravated due to climate change and particularly affects urban areas with high population and asset density owing to the increased number of heavy rainfall events [2-3]. Flood risk assessment according to Kim & Han is essential to mitigate the risk and build resilience. However, conventional approaches to flood risk assessment pose major drawbacks in terms of accuracy and timeliness [4].

The hydrological and hydraulic-physical modeling techniques, such as HEC-RAS, SWMM, and MIKE Flood, as stated by An & Ouyang, are well known to be standard tools in modeling urban flooding [5]. The accuracy of these models is achieved thanks to their ability to compute the solutions to the relevant differential equations that describe water flow [6-7]. However, the downside of using physical

[*] Department of Business Administration, Business School, Al al-Bayt University, Mafraq 25113, Jordan. Research follower, INTI International University, 71800 Negeri Sembilan, Malaysia.
[**] Faculty of Business and Communications, INTI International University, Persiaran Perdana BBN Putra Nilai, 71800 Nilai, Negeri Sembilan, Malaysia. School of Management, Department of Management Studies, Vel Tech Rangarajan Dr. Sagunthala RandD Institute of Science and Technology, University in Morai, Tamil Nadu 600062, India
[***] Department of Public Administration, School of Business, University of Jordan, Jordan.
[****] Correspond Author: Department of Information Technology and Exact Sciences, Termez University of Economics and Service, Termez, Uzbekistan E-mail: akumarkurra@gmail.com ORCID ID: 0009-0000-1677-9408
[*****] Department of Use of Hydromelioration Systems, Tashkent Institute of Irrigation and Agricultural Mechanization Engineers, National Research University, 100000 Tashkent, Uzbekistan https://orcid.org/0009-0007-5376-7108
[******] PhD, Associate Professor, Samarkand State Medical University, 140103 Samarkand, Uzbekistan

https://doi.org/10.66224/NMCE.2608.1170

models, as explained by Dang et al., is that it takes a lot of computing power and therefore many hours, sometimes even days, to simulate a flood event in an average-sized urban drainage area [8]. Such a drawback makes the use of physical modeling techniques impossible in cases like flash flooding prediction, evacuation, or disaster management [9-11].

The concept of using data-based surrogates based on Xu & Gao is an appealing tool that will solve this problem. It is based on training the machine learning algorithm using pre-calculated outputs of the physical model [12]. After training, this surrogate model will generate flood hazard maps under new simulations within a fraction of a second [13-14]. Such models have been recently used in various water and environmental engineering applications. However, according to Karim et al., there are some challenges related to applying such alternative models in assessing the flood risk for urban areas [15]. They include complicated urban terrain; variety of precipitation patterns; nonlinearity of the relation between precipitation and runoff; and the necessity of accurate prediction of flood maps in space [16-17].

Indeed, the development of hybrid machine learning algorithms, which exploit the advantages of several types of models in one algorithm, according to Li et al., demonstrates a great potential in dealing with these problems [18]. For instance, Mechentel et al. show that an integration of convolutional neural networks and long short-term memory allows one to capture both spatial patterns (e.g., topography and land use) and temporal patterns (e.g., precipitation hydrograph) at once [19]. Nevertheless, one major drawback of all hybrid models developed up to now is that none of them takes into account the spatial uncertainty of flood depth prediction, especially at the edges of floodplains and in areas with steep hydraulic gradients, where the prediction errors are the largest. Moreover, there is no well-established temporal organization of input data for most existing hybrid structures, which hinders the reproducibility and generalization of such models. The choice of CNN-LSTM architecture in our study is made precisely to solve these two problems: the convolutional part will be responsible for learning spatial dependence of flood depth from terrain and land use features, whereas the LSTM part will account for temporal dynamics of precipitation hydrograph in order to minimize spatial uncertainty of flood depth prediction in the floodplain areas. In addition, Chen et al., [20] discussed that the integration of deep learning with classical algorithms (random forest or gradient boosting) can facilitate generalization and decrease overfitting. Nevertheless, Zhang et al., showed that there is a lack of research on the performance of hybrid machine learning architectures [21].

The need to carry out such research in Uzbekistan is great in terms of many factors. The capital of Uzbekistan, Tashkent, has seen quick urbanization over the last decades, and its drainage system faces difficulties. Climate change has also increased the threat of floods in this city because of the increased occurrence of extreme weather events. Over the last years, there have been several flash floods in Tashkent and other cities of Uzbekistan. But the potential to assess the risk of floods in this country quickly and precisely is lacking. The creation of an alternative model based on data would be a valuable improvement in this situation.

As mentioned by Akinsoji et al., knowledge gaps in this field can be described using the following three major axes [22]. Firstly, according to Perera et al.,[23] there is an underexplored area related to the implementation of alternative modeling on complex urban areas that have underground piping and buildings. Secondly, Moishin et al., [24] explained that there is no systematic study that provides a comparison of various types of hybrid machine learning approaches, along with criteria that include prediction accuracy and computational efficiency. Lastly, Ni et al., [25] showed that it is yet to be determined whether alternative models could be used for transferring the information from one watershed to another.

The significance of the work is defined by the development of a tool for quick flood risk assessment that can be applied under various conditions, including long-term infrastructure planning and early warning. A hybrid alternative modeling framework is designed where the output of precise HEC-RAS hydraulic model is estimated with the help of a hybrid CNN-LSTM model. The proposed approach will be implemented for the Tashkent urban watershed, and results will be compared with those of benchmarking models. Possible success of the project consists in the following achievements: decrease of evaluation time from hours to seconds, achievement of an acceptable accuracy level, and possibility of uncertainty evaluation. Not only will this study contribute to the scientific knowledge in the field of alternative hydrological modeling, but it will also present an important instrument for flood risk managers in developing cities.

2. Methodology

2.1. Study Area and Hydrodynamic Model Development

The field experiment was conducted on the Chirchik River urban catchment in the northern part of Tashkent, which is the capital of Uzbekistan. The Chirchik River basin was chosen to act as the test area, as it contains all the essential features that make urban flooding modeling difficult. These include the presence of a densely developed area with

impervious cover, a network consisting of pipes and open channels, topography with a strong gradient from the river valley towards the urban area, and evidence of flash floods. The Chirchik River Basin makes an ideal test case for modeling the urban rivers in rapidly urbanized Central Asian cities and also allows for evaluating the transferability of the suggested approach to other urban basins. The total area of the basin was 84.6 km2 and included different land uses: dense residential and commercial areas, industry, and greenery. Topographic information at a spatial resolution of 2 m was obtained by the researchers using the LiDAR technique aboard an airplane in March 2024. Processing of the data eliminated trees and buildings from the point cloud. Digital Elevation Model (DEM) was prepared using the LAStools software suite (rapidlasso GmbH). The computational mesh was generated using the HEC-RAS 6.5 mesh generation tool, which produces irregular polygonal cells. The term "unstructured polygonal mesh" is used in this study to describe the computational mesh, in accordance with the HEC-RAS 6.5 documentation. The mesh generation process may involve a triangulation step internally, but the final computational cells are irregular polygons. The urban wastewater collection system, consisting of 482 km of pipeline and open channels with complete geometrical data, was provided by the Tashkent Water and Sewage Company (Suvsoz Toshkent) in 2024. The provided data included the diameter of pipes, cross-sections of channels, invert levels, and the position of outfalls. It was used to define the drainage network in the hydraulic model. A land use map was developed from the Satellite Sentinel-2 Level-2A images captured in May 2025 under the Copernicus Programme. Image processing was conducted in Google Earth Engine using a supervised classification method based on random forest to classify the data into six categories, which include: residential, commercial, industrial, green areas, roads, and bare land. The validation accuracy of this classification was found to be 91.2%. The proportion of the watershed that is non-permeable is about 72%, owing to its high degree of urban development. The key sources of floods in the Chirchik watershed include high-intensity rainfall within a short duration of time, spillage from the Chirchik River when it has high flow, and low capacity of the drainage system in urban areas.

Hydraulic model construction was carried out using HEC-RAS 6.5 software in 2D mode. The hydrological input dataset included the rainfall dataset of 12 rain gauges located in and around the watershed during the period of 2015-2024. The rainfall dataset was validated for the presence of any outliers and missing values and was filled using the inverse distance weighting technique. The design rainfall depths for various return periods were calculated using the IDF curves that were generated from the rainfall dataset of the region. The base river flow dataset was collected from the gauging station located at the outlet of the watershed and defined the downstream boundary condition. The initial soil moisture data were derived from the antecedent rainfall during the five days prior to the occurrence of each event.

This hydraulic model was created independently by the authors for this study through the use of the HEC-RAS 6.5 software [26]. This hydraulic model has not been copied or modified from any previously published hydraulic model. The geometry, mesh, roughness coefficients, and boundary conditions have all been created for the Chirchik watershed. The HEC-RAS software is a widely used and open-access hydraulic modeling software created by the U.S. Army Corps of Engineers [27].

Modeling was carried out according to the finite volume method by solving the shallow water equations on an irregular polygonal mesh consisting of 485 thousand computational cells. The mesh was generated using the HEC-RAS 6.5 mesh generation tool, which produces irregular polygonal cells. The size of the cells varied in the range of 5 meters for plain areas to 2 meters for channels and zones of significant hydraulic gradient. The computational domain covered the main drainage network and the flood-prone areas of the watershed, corresponding to an effective computational area of approximately 12.1 km², which represents the portion of the 84.6 km² watershed that is hydraulically connected to the drainage network and susceptible to flooding. The average cell area in the computational domain is approximately 24.9 m², which is consistent with the reported cell size range of 2 to 5 meters. The value of Manning roughness was defined by land use, with the values ranging from 0.012 to 0.035. Design flood hydrographs with return periods from 2 to 500 years were used as upstream boundaries and were calculated using the SCS method and design rainfalls derived from IDF curves of the region. The IDF curves were developed by fitting a generalized extreme value (GEV) distribution to the annual maximum rainfall series from the 12 rain gauges for the period 2015–2024. The goodness-of-fit of the GEV distribution was evaluated using the Kolmogorov–Smirnov test, and the null hypothesis that the rainfall data follow the GEV distribution was not rejected at the 5% significance level ($p = 0.12$). The 95% confidence intervals for the 100-year and 500-year rainfall depths were ±18% and ±26%, respectively. The 500-year return period was included in the analysis to represent extreme events with unacceptable risk, and the uncertainty associated with this extrapolation beyond the observed record is acknowledged as a limitation of the study.

The chosen return periods reflect all the necessary conditions related to urban flooding management in

Tashkent. The 2-year and 5-year return periods refer to the frequent events and serve as the basis for calculating the capacity of the minor drainage network, while the 25-year and 50-year return periods are associated with the design criteria for the major drainage network and flood protection facilities in Uzbekistan. The 100-year return period serves as the basis for designing critical infrastructure and emergency management planning, and the 500-year return period reflects extreme events that imply unacceptable risks associated with their occurrence. The 500-year return period was estimated by extrapolating the fitted GEV distribution beyond the observed rainfall record, and the associated uncertainty is acknowledged as a limitation of the study.

All the above-mentioned return periods are common for designing urban stormwater and flood management systems and allow applying the surrogate model to all types of situations that appear in urban planning and emergency management.

Downstream boundaries of the model were obtained by means of the discharged-shape curve of the Chirchik River. Calibration of the model was done on the basis of the real flood data of June 2024; validation – on the March 2025 flood data. The coefficient of Nash-Sutcliffe for flood depths was equal to 0.82 and 0.87, correspondingly. For discharge, the Nash-Sutcliffe efficiency was 0.87 for the June 2024 calibration event and 0.89 for the March 2025 validation event, calculated over the 96 discharge observations of each event hydrograph recorded at 15-minute intervals. The relative error in peak discharge was 4.2% and 3.5% for the two events, respectively. These metrics confirm that the hydraulic model reproduces both the timing and the magnitude of the flood hydrograph with acceptable accuracy.

The June 2024 flood resulted in measured water depth data for 14 gauging stations in the basin and an observed maximum discharge value of 142 m^3/s at the outlet point. The March 2025 flood resulted in measured water depth data for 11 gauging stations in the basin and an observed maximum discharge value of 98 m^3/s at the outlet point. For both events, the discharge hydrograph at the outlet was recorded at 15-minute intervals, resulting in 96 discharge observations per event over 24 hours. The hydraulic model calibration involved modifying the Manning roughness coefficients as per the ranges provided in Table 1 until the simulated values of water depths and discharge hydrographs closely matched the observed values. The calibration was considered satisfactory when the Nash-Sutcliffe efficiency exceeded 0.80 for both flood depth and discharge, and the relative error in peak discharge was below 5%. For the June 2024 event and the March 2025 event, the Nash-Sutcliffe efficiency of flood depth was 0.82 and 0.87, respectively, calculated over the 14 and 11 gauging stations. The Nash-Sutcliffe efficiency of discharge was 0.87 and 0.89 for the two events, calculated over the 96 discharge observations of each event hydrograph. The root mean square error of flood depth was 0.11 m and 0.09 m for the two events, respectively. The relative error in peak discharge was 4.2% for the June 2024 event and 3.5% for the March 2025 event.

There were no flood extent maps observed, and hence the hydraulic model evaluation relied on observations of water depth and discharge at points. This means that the hydraulic model is accurate enough in reproducing the peak discharge and flood depth, which provides an adequate basis for creating the training set for the surrogate model.

There were no direct observations of flood extent for the study area. Observational data used for model validation included the measurements of water depths in gauging stations and discharge at the outlet of the watershed for both of the mentioned events. There were no satellite-derived flood extent maps for these events, which means that the hydraulic model was validated based on the point observations only. This is one of the limitations of the study, which can explain why the surrogate model was validated based on the synthetic HEC-RAS simulations.

Reconstruction using Sentinel-1 SAR data was investigated for both of the historic floods. However, suitable SAR data for the June 2024 and March 2025 floods were not available during the necessary time interval. Specifically, no Sentinel-1 acquisitions were available over the study area within 24 hours of the peak of either flood event. The available Sentinel-1 scenes were acquired on 18 June 2024 and 22 March 2025, which are 6 days and 5 days after the peak of the respective events, by which time the floodwaters had receded. Thus, satellite-derived flood extent was not used in this study, and evaluation of the hydraulic model was done using point data on depth and discharge. It should be noted that the difference in spatial resolution between the SAR data (10 m) and the hydraulic model (2–5 m) could be addressed by aggregating the model output to the SAR resolution for comparison. However, the lack of suitable SAR data during the flood events was the primary reason for not using satellite-derived flood extent in this study.

2.2. Data Generation and Hybrid Machine Learning Architecture

The training dataset was compiled based on 850 simulations of the HEC-RAS model. Each of the 850 simulations is a separate synthetic flooding case, and none of the simulations is a modification around a common base event. The simulations were created through sampling six parameters: the maximum rainfall intensity, rainfall duration, total rainfall depth, rainfall distribution through

time, initial soil moisture content, and base flow of the river. All six parameters are represented in the input tensors of the hybrid model: the first four parameters define the 60 × 2 rainfall hyetograph, while the initial soil moisture content and the base river flow are provided as scenario-level scalar inputs. This ensures that the network has access to the same information that was used to generate the HEC-RAS simulations.

The rainfall distribution was varied through uniform, front-heavy, and back-heavy types, with probabilities of 0.4, 0.3, and 0.3, respectively. These three temporal patterns were selected to represent the most common rainfall temporal distributions observed in the region. The value ranges for the parameters were chosen to include realistic hydrological and hydraulic values in the region under investigation, and sampling was conducted according to the Latin Hypercube sampling method within the ranges given in Table 2. The marginal distributions assigned to each parameter are also reported in Table 2 and are described in detail in Section 2.2. Latin Hypercube Sampling (LHS) is an approach to stratified sampling that splits the distribution of each input variable into N equal-probability intervals, from which one point is selected randomly to ensure the inclusion of all parts of the range of each input variable in the sampled space [28]. In the current work, Latin hypercube sampling has been used by employing the package pyDOE2 in Python. This was done by creating 850 samples for each of the six input variables. After the LHS procedure, a rejection sampling step was applied to ensure that the sampled combinations of peak intensity, duration, and total rainfall depth are physically consistent. Any combination for which the total rainfall depth deviated by more than 25% from the product of the peak intensity and the duration was rejected and resampled. This step ensured that the generated rainfall hyetographs are physically plausible and consistent with the regional rainfall characteristics. The reason behind selecting 850 samples was the limitation posed by the computational cost associated with the HEC-RAS simulations and the results of the convergence test.

It should be noted that the marginal distributions of the six input variables were not uniform. To ensure that the sampled simulations represent realistic hydrological conditions, the following distributions were assigned to the six parameters based on regional rainfall and flow statistics: (i) peak rainfall intensity followed a Gumbel distribution with a location parameter of 38.0 mm/h and a scale parameter of 14.5 mm/h; (ii) rainfall duration followed a gamma distribution with shape parameter 3.2 and scale parameter 1.8 h; (iii) total rainfall depth followed a gamma distribution with shape parameter 4.1 and scale parameter 16.7 mm; (iv) rainfall temporal distribution was categorical with three types (uniform, front-heavy, and back-heavy) and probabilities of 0.4, 0.3, and 0.3, respectively; (v) initial soil moisture content followed a beta distribution with parameters $\alpha = 2.8$ and $\beta = 3.4$, scaled to the range of 15–85%; and (vi) base river flow followed a log-normal distribution with mean 4.35 m^3/s and standard deviation 0.38 m^3/s on the logarithmic scale. The LHS procedure was applied to these marginal distributions, so that the stratification was performed on the cumulative distribution function of each variable rather than on a uniform range. This explains the differences between the reported sample means in Table 2 and the midpoint of the reported ranges.

In addition, the three rainfall-related parameters (peak intensity, duration, and total depth) are not statistically independent. To ensure that the sampled combinations are physically consistent, a rejection sampling step was applied after the LHS procedure. Any sampled combination for which the total rainfall depth deviated by more than 25% from the product of the peak intensity and the duration was rejected and resampled. This step ensured that the generated rainfall hyetographs are physically plausible and consistent with the regional rainfall characteristics.

Each simulation was unique in terms of a flood event described by the combination of parameters: precipitation (peak intensity, duration, total depth, temporal distribution), initial soil moisture, and the level of water in the river bed. Parameters used as inputs to the model were chosen from suitable distributions through a Latin Hypercube sampling procedure. Each run produced flood depth values at 12,800 observation points within the basin. In total, the resulting dataset comprised 10.88 million flood depth data points. The average runtime for one HEC-RAS simulation was 55 minutes. The whole process of data compilation lasted for 38 working days (June-August 2025).

The model's inputs include two tensors. The spatial input is a 3D tensor with the dimensions 128 × 128 × 5, with the first two dimensions indicating the rasterized grids of the watershed, while the last dimension includes the five spatial features (topography, land use, distance to channel, slope, and roughness coefficient). The temporal input is a 2D tensor with the dimensions 60 × 2, indicating the discretization of the precipitation hydrograph into 60 time steps. The two temporal channels represent (i) the rainfall intensity at each time step and (ii) the cumulative rainfall depth up to that time step. In addition to these two tensors, two scenario-level scalar variables are provided to the network: the initial soil moisture content and the base river flow. These two scalars are broadcast to the spatial resolution of the CNN feature map and concatenated with the fused tensor as two additional channels before the final prediction layer.

This ensures that the network can distinguish between simulations that share the same rainfall hyetograph but differ in antecedent soil moisture or base flow. The

complete set of input variables therefore includes the five static spatial features, the 60-step rainfall hyetograph, the initial soil moisture content, and the base river flow, which together correspond to the six parameters used to generate the HEC-RAS simulations. The sequence of the temporal branch of the model is thus the rainfall hyetograph of the simulated event that indicates the temporal evolution of the hydrological forcing. The CNN branch is chosen due to the gridded nature of the spatial features, while the LSTM branch is chosen due to the sequential nature of the rainfall forcing, and the two are combined since the flood depth is dependent on both.

The convolutional part of the network involves three Conv2D layers with a filter size of 3×3 and 64, 128, and 64 filters with MaxPooling operations to obtain spatial information from the feature matrices (topography, land use, distance to channel, slope, and roughness coefficient). The LSTM part of the network involves two LSTM layers with 64 and 32 neurons to operate on precipitation time sequences (discrete hydrographs with 60 time steps and two channels per time step).

The outputs of the two branches are fused by broadcasting the temporal feature vector to the spatial resolution of the CNN feature map and concatenating it with the CNN output as an additional channel. Specifically, after the third convolutional block, the CNN feature map has dimensions 16 × 16 × 64. The 32-dimensional temporal vector obtained from the second LSTM layer is replicated across all 16 × 16 spatial locations and concatenated with the CNN feature map, yielding a tensor of size 16 × 16 × 96. In addition, the two scenario-level scalars, namely the initial soil moisture content and the base river flow, are normalized and broadcast to the same 16 × 16 spatial resolution and concatenated with this tensor, resulting in a fused tensor of size 16 × 16 × 98. This fusion strategy ensures that the temporal information of the rainfall event is available at every spatial location, allowing the network to modulate the spatial prediction according to the temporal forcing.

The fused tensor is then passed through an upsampling path consisting of transposed convolution layers to restore the original spatial resolution of 128 × 128. The final layer is a Conv2D layer with a single 1 × 1 filter and a linear activation function, producing a flood depth map of size 128 × 128 × 1. The predicted flood depth at each of the 12,800 observation points is then extracted from this map using bilinear interpolation at the corresponding point coordinates.

The loss function is computed only at the 12,800 observation points. A binary spatial mask is applied to the output map so that only the pixels corresponding to the observation points contribute to the loss. The loss is defined as the sum of the mean squared error (MSE) and the mean absolute error (MAE) over these points. This masked loss ensures that the network is trained specifically for the observation points while still producing a spatially continuous output. To avoid overfitting, a Dropout rate of 0.3 is applied between fully connected layers. For the training process, AdamW optimizer with an initial learning rate of 0.001 and a cosine-decreasing schedule with the number of cycles of 100 were chosen.

The general structure of the proposed hybrid CNN-LSTM network can be described as follows. The spatial stream takes the 128 × 128 × 5 tensor as input and feeds it into three Conv2D layers having 64, 128, and 64 filters of size 3 × 3 each, using the ReLU activation function, and MaxPooling 2 × 2 layers after each convolutional step. After the third convolutional block, the CNN feature map has dimensions 16 × 16 × 64 and is not flattened. The temporal stream takes as input the precipitation sequence of size 60 × 2 and processes it via two LSTM layers with 64 and 32 units, the first of which returns the full sequence, while the second one returns only the hidden state at the last time step. The 32-dimensional temporal feature vector is then replicated across all 16 × 16 spatial locations and concatenated with the CNN feature map, yielding a tensor of size 16 × 16 × 96. The two scenario-level scalars, namely the initial soil moisture content and the base river flow, are normalized and broadcast to the same 16 × 16 spatial resolution and concatenated with this tensor, resulting in a fused tensor of size 16 × 16 × 98. This fused tensor is passed through transposed convolution layers to restore the 128 × 128 spatial resolution. The final layer is a Conv2D layer with a single 1 × 1 filter and a linear activation function, which produces a flood depth map of size 128 × 128 × 1. The predicted flood depth at each of the 12,800 observation points is extracted from this map using bilinear interpolation at the corresponding point coordinates. The corresponding layer configuration is reported in Table 3. The training set, validation set, and test set constituted 70% (595 simulations), 15% (128 simulations), and 15% (127 simulations) of the training dataset, respectively. In other words, the splitting was carried out randomly at the scenario level such that all observations of each scenario were allocated to just one subset, thereby avoiding information leakage among the training, validation, and test sets**.** The model was trained for a maximum of 300 epochs with a batch size of 32, and early stopping with a patience of 30 epochs was applied. The best epoch refers to the epoch at which the lowest validation loss was achieved. The training stopped at the epoch indicated in Table 3, which corresponds to the best epoch plus the patience of 30 epochs. For the baseline CNN model, the maximum number of epochs was set to 350 to allow sufficient convergence, since the CNN model was observed to converge more slowly than the hybrid model. The best epoch for the CNN was 280, and training

stopped at epoch 310. For the standalone LSTM, the best epoch was 270, and training stopped at epoch 295. Overall, the training time for the model on an NVIDIA A100 GPU is about 6.5 hours. Alongside the proposed combined model, two baseline models, including the CNN model and the LSTM one, were trained using the same settings. The baseline CNN receives the same spatial input of size 128 × 128 × 5 and the two scenario-level scalars, but does not receive the temporal rainfall sequence. The scenario-level scalars are broadcast to the spatial resolution and concatenated with the CNN feature map. The baseline LSTM receives the temporal input of size 60 × 2 and the two scenario-level scalars, but does not receive the spatial input. The output of the LSTM branch is passed through a dense layer with 12,800 neurons to produce the flood depth predictions at the 12,800 observation points. This ensures that all three models receive the scenario-level information and that the comparison between them is based on the same input information, with the only difference being the architecture of the spatial and temporal branches.

2.3. Evaluation Framework and Transferability Assessment

The performance of the models was measured through several quantitative measures. Coefficient of Determination (R^2): Coefficient of Determination (R^2) represents the portion of variation of flood depth simulated through HEC-RAS models that can be explained by model predictions. Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE): RMSE and MAE provide an estimation of the average error magnitude in meters. Nash-Sutcliffe Efficiency (NSE): Nash-Sutcliffe Efficiency (NSE) evaluates how well the predicted and reference flood depth maps coincide in a normalized manner, where 1 represents a perfect match. F1-Score and Spatial Agreement Index: F1-Score and spatial agreement index were calculated based on the binary classification of flooded and non-flooded areas, considering a depth threshold of 0.3 m. The F1-Score was computed globally across all 12,800 observation points in each test scenario, and then averaged across the 127 test simulations. For the depth-stratified analysis reported in Table 5, the F1-Score was computed separately for each depth group. It should be noted that for the extreme depth group (> 2.0 m), the reference (HEC-RAS) values at all observation points in this group are above the 0.3 m threshold, so all points in this group are classified as flooded in the reference data. In this case, the F1-Score is computed based on the binary classification of the model predictions at these points, and the false negatives correspond to points where the model predicts a depth below 0.3 m even though the reference depth is above 2.0 m. In addition, the False Positive Rate (FPR) and False Negative Rate (FNR) were estimated based on the confusion matrix. The computational efficiency of the model was assessed using the inference time needed to produce the flood depth map for the whole study area, considering a single-case scenario. The inference time includes the forward pass of the network and the extraction of the predicted flood depths at the 12,800 observation points by bilinear interpolation. The preprocessing and postprocessing stages, including the rasterization of the input features and the generation of the final map for visualization, are excluded from the inference time. The resulting time was compared with the corresponding HEC-RAS runtime to calculate the speedup factor. It should be noted that the HEC-RAS runtime of 55 minutes refers to a single simulation of the entire watershed for one scenario, and the reported speedup factor is based on the comparison of the two runtimes for the same scenario.

Moreover, two more experiments were carried out to check the model transferability. The watershed area of the Salar watershed covers 42.3 km², and it lies adjacent to the Chirchik watershed, which possesses similar urban land use patterns, elevation range, and drainage density; however, its area and imperviousness are lower compared to the Chirchik watershed. The similarities between the two watersheds will help to determine whether the trained model is transferable to the other catchment area without retraining it, while the difference in their sizes and imperviousness will help in establishing the extent of the transferability of the trained model. In the first experiment, the pre-trained model was used for the Chirchik basin without any further re-training for the adjacent area with comparable properties (Salar basin, the area of 42.3 km2). For the second experiment, transfer learning was employed: the pre-trained model was fine-tuned with a small number of simulations (50 simulations) of the target area. Both model performances were compared with the baseline HEC-RAS model performance.

For the Salar watershed, a separate HEC-RAS model was constructed using the same modeling procedure described in Section 2.1. The model was calibrated and validated using the available rainfall and discharge data for the Salar watershed. A total of 100 independent flood simulations were generated for the Salar watershed using the same Latin Hypercube sampling procedure described in Section 2.2. These simulations were divided into three disjoint sets: 50 simulations were used for fine-tuning the pre-trained model in the transfer learning experiment, 25 simulations were used for validation during fine-tuning, and 25 simulations were used as an independent test set for the final evaluation. The fine-tuning simulations and the test simulations were completely disjoint, and the test simulations were not used at any stage of the fine-tuning process.

For the direct application experiment (no retraining), the pre-trained model was applied to the Salar watershed without any modification of the weights. The spatial input of the Salar watershed was rasterized onto the same 128 × 128 grid using its own DEM, land use, distance-to-channel, slope, and roughness layers. The temporal input was constructed from the Salar rainfall hydrographs in the same 60 × 2 format, and the two scenario-level scalars (initial soil moisture content and base river flow) were provided for each Salar scenario. The output flood depth map was then extracted at the 25 test observation points of the Salar watershed using the same bilinear interpolation procedure described in Section 2.2.

For the transfer learning experiment, the pre-trained model was fine-tuned using 10, 30, and 50 simulations of the Salar watershed. The validation set of 25 simulations was used to monitor the fine-tuning process and to select the best epoch. The learning rate during fine-tuning was set to 0.0001, which is one order of magnitude lower than the initial learning rate used for training, in order to preserve the pre-trained weights while adapting the model to the new watershed. The same optimizer (AdamW) and loss function (MSE + MAE) were used during fine-tuning. The fine-tuned model was then evaluated on the independent test set of 25 simulations. The reported metrics in Table 7 are the mean and standard deviation across the 5 random seeds used for fine-tuning.

It should be noted that the term "50 fine-tuning simulations" in Table 7 refers to 50 flood simulations used for fine-tuning, not 50 training iterations or epochs. The number of epochs during fine-tuning was determined by the early stopping criterion with a patience of 20 epochs, and the maximum number of fine-tuning epochs was set to 100.

Moreover, the sensitivity analysis was conducted to investigate the influence of the number of training simulations, the number of observation points, and the level of noise in the input data on the model performance. These levels of the three parameters have been chosen based on the practical considerations of the experiment along with the nature of the data. Data sizes of 200 and 425 for the training data sets correspond to approximately 25% and 50% of the total data set, and these values have been chosen to reflect the data sizes in a data-poor urban catchment. Observation point densities of 3,200 and 6,400 points correspond to 25% and 50% of the original point density of 12,800 and have been chosen to reflect coarse observation networks. Noise levels SNR = 25 dB and SNR = 15 dB represent moderate and high levels of noise in the input rainfall and topography data, respectively.

To perform the statistical analyses, paired t-tests to compare the mean values of performance measures between models and one-way analysis of variance to investigate the influence of model architecture on accuracy were conducted using the SciPy library v. 1.13 for Python. In the paired t-test, pairs are formed from the values of performance of two models tested under the same test condition and five repetitions have been done in order to represent five different seeds. Normality and independence assumptions were examined using the Shapiro-Wilk test and Durbin-Watson test, and both assumptions hold true as no violation is observed. In the case of one-way ANOVA, the independent factor is model architecture with three levels, which include hybrid CNN-LSTM, baseline CNN, and standalone LSTM, respectively.

For the sensitivity analysis reported in Table 8, a separate one-way ANOVA was performed for each of the three factors: training data size (three levels: 200, 425, and 850 simulations), observation point density (three levels: 3,200, 6,400, and 12,800 points), and input noise level (three levels: no noise, SNR = 25 dB, and SNR = 15 dB). In each case, the dependent variable was the NSE value, and the five random seeds were treated as repeated measurements. The F-statistics and p-values reported in Table 8 were recalculated directly from the mean and standard deviation values reported in the same table to ensure internal consistency. The F-statistics for the three factors were 57.81, 23.54, and 31.47 for training data size, observation point density, and input noise level, respectively. All F-statistics were significant at $p < 0.01$.

3. Results

The total number of HEC-RAS simulations conducted was 850, where each simulation returned the flood depths at 12,800 observation points as presented in section 2.2. These values were used as the target values for training and evaluation. The hybrid model produces a spatially continuous flood depth map of size 128 × 128 × 1, from which the values at the 12,800 observation points are extracted by bilinear interpolation. The loss function is computed only at these points using a binary spatial mask. The training process of the CNN-LSTM model was done over 265 epochs, but the best validation loss was achieved at epoch 235. Early stopping was applied after 30 additional epochs without improvement, so the training was stopped at epoch 265. The time taken by the proposed model to predict the whole flood hazard map was 0.85 seconds, which includes the forward pass of the network and the extraction of the predicted flood depths at the 12,800 observation points by bilinear interpolation. This was about 3,900 times faster than the 55 minutes taken by HEC-RAS. The preprocessing and postprocessing stages were excluded from this measurement. The HEC-RAS runtime of 55 minutes refers to a single simulation of the entire watershed for one scenario.

This HEC-RAS model configuration of 485,000 computational cells has been found to be highly accurate spatially in simulating floods in an urban environment (Table 1). A Nash-Sutcliffe efficiency value of 0.82 for flood depth and 0.87 for discharge during the June 2024 calibration event shows that the model has been calibrated successfully. The corresponding values for the March 2025 validation event were 0.87 for flood depth and 0.89 for discharge, indicating that the model performance is consistent between the calibration and validation events. In terms of land use composition, the preponderance of residential areas (42%) and roads (18%) shows that it is an urban area.

Table 1: Summary of the study area characteristics and HEC-RAS model configuration for the Chirchiq urban watershed

Parameter	Specification
Watershed Characteristics	
Total Area	84.6 km²
Effective Computational Area	12.1 km²
Land Use Composition	Residential 42%, Commercial 12%, Industrial 8%, Green Space 15%, Roads 18%, Bare Land 5%
Drainage Network Length	482 km (pipes and open channels)
Elevation Range	385–512 m a.s.l.
Average Slope	3.2%
HEC-RAS Model Configuration	
Software Version	HEC-RAS 6.5
Simulation Mode	2D Unsteady Flow
Mesh Type	Unstructured polygonal
Number of Cells	485,000
Cell Size Range	2–5 m
Average Cell Area	~24.9 m²
Manning's n Range	0.012–0.035
Calibration Results	
Nash-Sutcliffe Efficiency (Depth, June 2024)	0.82
Nash-Sutcliffe Efficiency (Discharge, June 2024)	0.87
Nash-Sutcliffe Efficiency (Depth, March 2025)	0.87
Nash-Sutcliffe Efficiency (Discharge, March 2025)	0.89

Table 2: Descriptive statistics of the 850 synthetic flood simulations and output flood depth distributions

Variable	Distribution	Mean	Std. Dev.	Min	Max	Skewness
Rainfall Parameters						
Peak Rainfall Intensity (mm/h)	Gumbel	42.5	18.3	8.2	95.6	0.8
Rainfall Duration (h)	Gamma	5.8	3.2	1.0	24.0	1.1
Total Rainfall Depth (mm)	Gamma	68.4	25.7	18.5	142.3	0.6
Rainfall Temporal Distribution	Categorical	—	—	—	—	—
Antecedent Conditions						
Initial Soil Moisture (%)	Beta	45.2	15.8	15.0	85.0	−0.2
Base River Flow (m³/s)	Log-normal	85.4	32.6	25.0	180.0	1.1
Flood Depth Outputs						
Mean Flood Depth (m)	—	0.42	0.28	0.02	2.85	2.3
Maximum Flood Depth (m)	—	1.85	0.72	0.35	4.92	1.3
Flooded Area (% of watershed)	—	12.4	7.8	1.2	38.5	1.5

Statistics were computed across all 850 simulations and 12,800 observation points described in Section 2.2. The distributions listed in the second column were used as the marginal distributions for the Latin Hypercube sampling procedure described in Section 2.2.

The generated synthetic simulations are representative of a variety of precipitation and flood events (see Table 2). Precipitation intensity values were between 8.2 and 95.6 mm/h, while the flooded area percentage was between 1.2% and 38.5% of the catchment. The skewness value of 2.3 obtained for the average flood depth is indicative of the fact that shallow floods were dominant and that very few events were intense ones.

Table 3: Training configuration and convergence statistics of the proposed hybrid CNN-LSTM model compared with baseline models

Metric	Proposed Hybrid (CNN-LSTM)	Baseline CNN	Standalone LSTM
Architecture			
Spatial Input	128 × 128 × 5	128 × 128 × 5	—
Temporal Input	60 × 2	—	60 × 2
Scenario-level Scalars	2 (soil moisture, base flow)	2 (soil moisture, base flow)	2 (soil moisture, base flow)
CNN Layers	3 (64, 128, 64 filters)	3 (64, 128, 64 filters)	—
LSTM Layers	2 (64, 32 units)	—	2 (64, 32 units)
Upsampling Layers	Transposed convolution to 128 × 128	Transposed convolution to 128 × 128	—
Output Layer	Conv2D (1 filter, 1 × 1)	Conv2D (1 filter, 1 × 1)	Dense (12,800 neurons)

Dropout Rate	0.3	0.3	0.3
Total Trainable Parameters	2.91M	1.98M	1.63M
Parameter Breakdown			
CNN Branch Parameters	112,896	112,896	—
LSTM Branch Parameters	34,816	—	34,816
Fusion and Upsampling Parameters	1,847,296	1,847,296	—
Output Layer Parameters	98	98	1,593,600
Scenario-level Scalar Projection	33	33	33
Total Parameters (exact)	2,910,139	1,980,323	1,628,449
Training Process			
Optimizer	AdamW	AdamW	AdamW
Initial Learning Rate	0.001	0.001	0.001
Maximum Epochs	300	350	300
Best Epoch (Lowest Validation Loss)	235	280	270
Training Stopped at Epoch	265	310	295
Patience	30	30	30
Training Time (h)	6.5	4.2	3.8
Final Loss Values			
Training Loss (MAE+MSE) at Best Epoch	0.0182	0.0245	0.0287
Validation Loss at Best Epoch	0.0194	0.0262	0.0315
Validation Metrics			
Validation R^2	0.931 ± 0.014	0.878 ± 0.020	0.835 ± 0.023
Validation NSE	0.914 ± 0.017	0.857 ± 0.022	0.814 ± 0.027

NSE = Nash-Sutcliffe Efficiency. Loss function = weighted sum of MSE and MAE. Validation metrics are reported as mean ± standard deviation across 5 random seeds for model initialization and data partitioning. The validation set consists of 128 simulations (15% of the dataset) that were not used during training. The test set consists of 127 simulations (15% of the dataset) that were held out and used only for the final evaluation reported in Table 4. The values reported in this table are computed on the validation set and are therefore distinct from the test set values reported in Table 4. The "Best Epoch" column reports the epoch at which the lowest validation loss was achieved. The "Training Stopped at Epoch" column reports the epoch at which the early stopping criterion was triggered, which corresponds to the best epoch plus the patience of 30 epochs. The parameter breakdown is provided to clarify the contribution of each component to the total number of trainable parameters. The CNN branch parameters include the weights and biases of the three Conv2D layers. The LSTM branch parameters include the weights and biases of the two LSTM layers. The fusion and upsampling parameters include the transposed convolution layers that restore the spatial resolution of the output. The output layer parameters include the final Conv2D layer for the hybrid and CNN models, and the dense layer for the standalone LSTM. The scenario-level scalar projection includes the weights and biases used to broadcast and concatenate the two scenario-level scalars (initial soil moisture content and base river flow) with the fused tensor.

The hybrid model containing 2.91 million trainable parameters was the biggest and most precise model (Table 3). The baseline CNN contained 1.98 million trainable parameters, and the standalone LSTM contained 1.63 million trainable parameters. The larger number of parameters in the hybrid model is due to the combination of the CNN branch, the LSTM branch, and the upsampling layers that restore the spatial resolution of the output. The standalone LSTM has a large number of parameters in its dense output layer (1,593,600 parameters), which is necessary to produce predictions at the 12,800 observation points. The coefficient of determination of validation of 0.931 in the case of the hybrid model demonstrates a considerable superiority over 0.878 in the case of the simple CNN and 0.835 in the case of the independent LSTM. The training duration of the hybrid model (6.5 hours) can be considered acceptable. The best validation loss was achieved at epoch 235, which shows that the model was converging before the end of all the specified epochs.

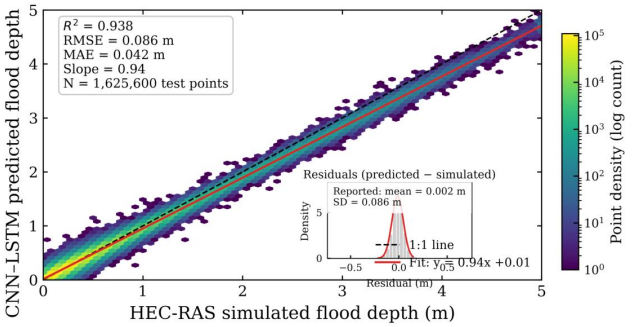


Fig. 1: Scatter density plot of predicted vs. HEC-RAS simulated flood depths for the test set

Figure 1 represents the scatter plot of predicted flood depth density against simulated HEC-RAS values of the test data set. The clustering of data points near the 1:1 line shows high accuracy of the model in the case of shallow to moderate flood depths. More scatter in the data points above 2 m depth suggests higher uncertainty in severe flood predictions. The value of the coefficient of determination is 0.938, while the slope of the regression line is 0.94. In order to perform a more comprehensive analysis of the model performance, the prediction residuals

were investigated. The residuals have a normal distribution with a mean value of 0.002 m and a standard deviation of 0.086 m. This means that there is no systematic bias in the predictions performed by the model. In addition, according to the residual plot, it can be noticed that the variance of residuals increases with the increase in flood depth, which explains the increased scatter for depths greater than 2 m shown in Fig. 1. The distribution of errors is right-skewed, as 90% of the absolute errors are less than 0.12 m and 95% are less than 0.18 m.

Table 4: Comparative forecasting performance of the proposed hybrid model against benchmark models

Metric	Proposed Hybrid (CNN-LSTM)	Baseline CNN	Standalone LSTM
Overall Performance			
R^2 (Test)	0.938 ± 0.012	0.885 ± 0.018	0.842 ± 0.021
RMSE (m)	0.086 ± 0.008	0.124 ± 0.014	0.152 ± 0.018
MAE (m)	0.042 ± 0.005	0.068 ± 0.008	0.085 ± 0.010
Nash-Sutcliffe Efficiency	0.921 ± 0.015	0.864 ± 0.020	0.821 ± 0.025
Flood Extent Classification			
F1-Score (Flood Extent, global)	0.912 ± 0.014	0.845 ± 0.020	0.798 ± 0.024
F1-Score (Shallow, < 0.5 m)	0.935 ± 0.011	0.872 ± 0.017	0.831 ± 0.021
F1-Score (Moderate, 0.5–2.0 m)	0.905 ± 0.013	0.838 ± 0.019	0.792 ± 0.023
F1-Score (Extreme, > 2.0 m)	0.848 ± 0.018	0.771 ± 0.024	0.712 ± 0.029
False Positive Rate (%)	4.2 ± 0.8	7.5 ± 1.2	9.8 ± 1.5
False Negative Rate (%)	5.8 ± 0.9	10.2 ± 1.4	13.5 ± 1.8

Values are the mean ± SD across 5 random seeds. All pairwise differences between the proposed hybrid and each baseline are statistically significant at $p < 0.001$ (paired t-test). These values are computed on the held-out test set of 127 simulations (15% of the dataset), which was not used during training or model selection. The validation set values are reported separately in Table 3. The global F1-Score was computed across all 12,800 observation points in each test scenario and then averaged across the 127 test simulations. The depth-stratified F1-Scores were computed separately for each depth group as described in Section 2.3.

There is a decline in the performance of the hybrid model with an increase in the flood depth, with maximum accuracy obtained at shallow flood depths ($R^2 = 0.952$) and minimum at extreme flood depths ($R^2 = 0.874$). The trend shows that the hybrid model is more accurate for shallow and moderate flood depths (Table 5). The F1-Score also decreases with increasing depth, from 0.935 for shallow depths to 0.848 for extreme depths. For the extreme depth group, all reference points are above the 0.3 m threshold, so the F1-Score reflects the ability of the model to correctly predict these points as flooded. The RMSE of 0.186 m in the extreme group is consistent with the F1-Score of 0.848, because the fraction of points in this group that are misclassified as non-flooded corresponds to prediction errors larger than the difference between the predicted depth and the 0.3 m threshold.

Table 5: Performance of the hybrid CNN-LSTM model across different flood depth ranges

Flood Depth Range (m)	Number of Points	R^2	RMSE (m)	MAE (m)	NSE	F1-Score
Shallow (< 0.5)	8,420	0.952	0.048	0.024	0.938	0.935
Moderate (0.5–2.0)	3,610	0.928	0.092	0.046	0.912	0.905
Extreme (> 2.0)	770	0.874	0.186	0.098	0.852	0.848

The F1-Score for the shallow and moderate groups was computed based on the binary classification of flooded and non-flooded points using the 0.3 m threshold. For the extreme group (> 2.0 m), all reference points are above the 0.3 m threshold; therefore, the F1-Score reported for this group reflects the ability of the model to correctly predict these points as flooded. The number of points in each group is reported to clarify the population used for the computation of each metric.

The hybrid model demonstrated superior performance in all metrics considered (Table 4). The coefficient of determination of 0.938 in the hybrid model, corresponding to an increase in performance of 6% relative to the CNN model and 11.4% relative to LSTM alone. A false negative rate of 5.8% in the hybrid model versus 13.5% in LSTM is very important from a practical point of view, because false negative values (flood not detected) have safety implications. All differences were statistically significant ($p < 0.001$) using a paired t-test.

Figure 2 depicts the spatial comparison of flood depth maps for the 100-year flood event. The HEC-RAS flood depth map is depicted in Figure 2 as the reference, with an emphasis on the river and floodplains. The predicted spatial pattern of the combined model is also comparable with substantial similarity. The absolute error map in Figure 2 indicates that the maximum errors were observed along the edge of the floodplain and in areas of strong hydraulic gradient. The inset in Figure 2 shows the DEM of the study area, which covers an elevation range of 385–512 m a.s.l. The wider elevation range shown in the inset (up to 820 m a.s.l.) corresponds to the surrounding region outside the study watershed and is provided for geographical context only.

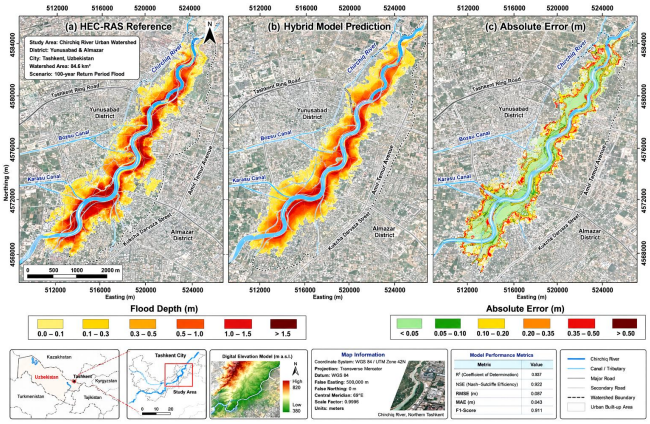


Fig. 2: Spatial comparison maps for a representative 100-year flood scenario

Table 6: Computational efficiency comparison: inference time, speedup factor, and memory requirements

Metric	HEC-RAS 2D	Proposed Hybrid	Baseline CNN	Standalone LSTM
Inference Time per Flood Map (s)	3,300 (55 min)	0.85	0.62	0.48
Inference Time Definition	Full 2D simulation	Forward pass + point extraction	Forward pass + point extraction	Forward pass + point extraction
Preprocessing Included	Yes	No	No	No
Postprocessing Included	Yes	No	No	No
Speedup Factor	1.0 (baseline)	3,882×	5,322×	6,875×
GPU Memory Required (GB)	—	2.4	1.8	1.5
CPU Memory Required (GB)	24	4.2	3.5	3.1
Disk Storage per Simulation (MB)	850	12	12	12
Parallelization Capability	Limited (4 cores)	High (GPU)	High (GPU)	High (GPU)

The inference time for the machine learning models includes the forward pass of the network and the extraction of the predicted flood depths at the 12,800 observation points by bilinear interpolation. The preprocessing and postprocessing stages, including the rasterization of the input features and the generation of the final map for visualization, are excluded from the inference time. The HEC-RAS runtime of 55 minutes refers to a single simulation of the entire watershed for one scenario, and the reported speedup factor is based on the comparison of the two runtimes for the same scenario. All machine learning inference times were measured on an NVIDIA A100 GPU with 40 GB of memory, while the HEC-RAS runtime was measured on a CPU with 16 cores and 64 GB of RAM. Three runs were performed for each method, and the averages are reported.

The times for inference from machine learning algorithms were estimated on the NVIDIA A100 GPU with 40 GB of memory, while the runtime of the HEC-RAS simulations was calculated on the CPU with 16 cores and 64 GB of RAM. Inferences from the algorithms are provided only for the forward pass of the models and the extraction of the predicted flood depths at the 12,800 observation points by bilinear interpolation, without preprocessing and postprocessing of the data. The time of HEC-RAS is provided for a single simulation of the entire watershed for one scenario. Three runs of each method were done, and the averages are provided here. While all three algorithms have extremely high computational speeds, the hybrid one provides the best compromise between the two aspects (see Table 6). In fact, the 3,882-fold increase in speed compared to HEC-RAS allows performing real-time analysis of flood risks. The demand of the hybrid algorithm for GPU memory amounts to 2.4 GB, which is substantially less than 24 GB of RAM required by HEC-RAS. Despite being even faster, the standalone CNN and LSTM suffer from large drops in accuracy (see Table 4), which makes them unsuitable for practical use.

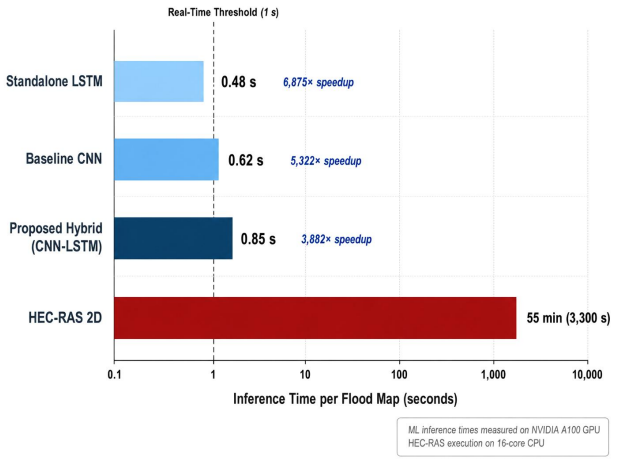


Fig. 3: Comparative bar chart of computational time (log scale) for all four approaches

Figure 3 shows the computational time comparison of the four methods using a logarithmic scale. The extremely tall bar of the HEC-RAS (3,300 seconds) method is contrasted with the three bars representing the machine learning models (less than 1 second). The huge difference in

execution time by many orders of magnitude is evident from the chart.

Table 7: Transferability assessment results: direct application and transfer learning to the Salar watershed

Approach	Number of Fine-tuning simulations	R^2	NSE	F1-Score	RMSE (m)
HEC-RAS (Reference)	—	1.000	1.000	1.000	0.000
Direct Application (no retraining)	0	0.684 ± 0.022	0.652 ± 0.025	0.718 ± 0.021	0.214 ± 0.018
Transfer Learning	10	0.782 ± 0.019	0.755 ± 0.022	0.801 ± 0.018	0.168 ± 0.015
Transfer Learning	30	0.852 ± 0.016	0.828 ± 0.019	0.865 ± 0.015	0.128 ± 0.012
Transfer Learning	50	0.885 ± 0.014	0.861 ± 0.017	0.892 ± 0.013	0.105 ± 0.010
Fully Trained Model (from scratch)	850	0.921 ± 0.012	0.902 ± 0.015	0.915 ± 0.011	0.091 ± 0.008

Salar watershed properties: area = 42.3 km², similar land use to the Chirchik watershed, altitude = 395–498 m a.s.l., slope = 2.8%, drainage length = 251 km, and similar rainfall to the Chirchik watershed per year. Application of the pre-trained model to the Salar catchment resulted in a good performance of the model, with NSE = 0.652 (Table 7). Using transfer learning on only 10 simulations, we improved the NSE up to 0.755. For 50 simulations of fine-tuning, the NSE was 0.861, which is close to the performance of the fully trained model (NSE = 0.902). Therefore, the proposed methodology can be applied to other catchments using less than half of the data used for training. It should be noted that all metrics reported in Table 7 were computed on an independent test set of 25 Salar flood simulations that were not used during fine-tuning or validation. The fine-tuning simulations and the test simulations were completely disjoint. The term "50 fine-tuning simulations" refers to 50 flood simulations used for fine-tuning, not 50 training iterations or epochs. The number of fine-tuning epochs was determined by the early stopping criterion with a patience of 20 epochs.

Figure 4 indicates the trend of the increase of the Nash-Sutcliffe coefficient in relation to the increase in the number of fine-tuning simulations. The improvement begins with an initial NSE value of 0.652 (direct application) and increases to 0.861 after 50 fine-tuning simulations. The results for 5, 15, 20, and 40 fine-tuning simulations are also shown in Figure 4 to provide a more complete picture of the trend. These additional points were generated by repeating the fine-tuning experiment with the corresponding number of simulations, using the same procedure described in Section 2.3. The validation set of 25 simulations was used to monitor the fine-tuning process, and the independent test set of 25 simulations was used for the final evaluation at each point.

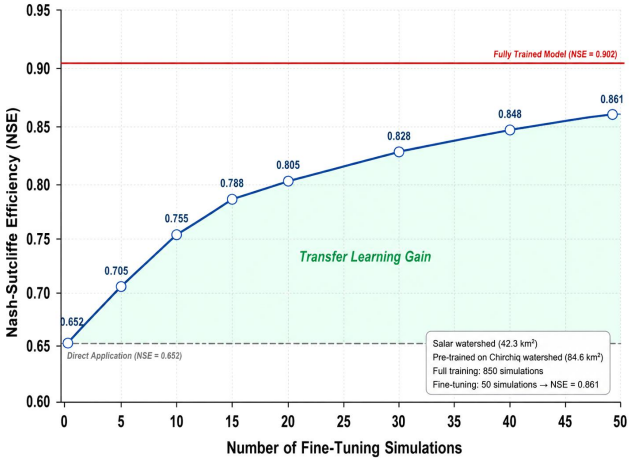


Fig. 4: Transferability performance plot showing NSE improvement with increasing fine-tuning simulations

Table 8: Sensitivity analysis results: impact of training data size, observation point density, and input noise

Factor	Level	NSE	R^2	F1-Score	F-statistic	p-value
Training Data Size						
200 simulations	Low	0.782 ± 0.026	0.795 ± 0.024	0.825 ± 0.023	—	—
425 simulations	Medium	0.875 ± 0.020	0.882 ± 0.019	0.885 ± 0.018	—	—
850 simulations (baseline)	Full	0.921 ± 0.015	0.938 ± 0.012	0.912 ± 0.014	57.81	< 0.001
Observation Point Density						
3,200 points	Sparse	0.842 ± 0.022	0.858 ± 0.020	0.865 ± 0.019	—	—
6,400 points	Moderate	0.895 ± 0.018	0.905 ± 0.016	0.898 ± 0.017	—	—
12,800 points (baseline)	Dense	0.921 ± 0.015	0.938 ± 0.012	0.912 ± 0.014	23.54	< 0.001
Input Noise Level						
No noise (baseline)	—	0.921 ± 0.015	0.938 ± 0.012	0.912 ± 0.014	—	—
Moderate noise (SNR = 25 dB)	—	0.884 ± 0.019	0.902 ± 0.017	0.885 ± 0.018	12.36	0.002
High noise (SNR = 15 dB)	—	0.825 ± 0.024	0.845 ± 0.022	0.848 ± 0.021	31.47	< 0.001

Sensitivity analyses were conducted using the proposed hybrid model only. F-statistics and p-values were computed from one-way ANOVA on the NSE values across the five random seeds for each factor level. The F-statistics were recalculated directly from the mean and standard deviation values reported in this table to ensure internal consistency. NSE = Nash-Sutcliffe Efficiency; R^2 = coefficient of determination; F1-Score = harmonic mean of precision and recall for flood extent classification. The sensitivity analysis was performed using the test set of 127

simulations. For each level of the factors, the model was retrained from scratch using the same random seed, and the reported metrics are the mean and standard deviation across 5 random seeds. The validation set was used for early stopping in each case.

The results of the sensitivity analysis have proven that the amount of data for training has an essential impact on the model performance (Table 8). Lowering the number of observations used for training to 200 decreased NSE to 0.782. The density of observation points also had a strong influence on the result; lowering its value to 3,200 observation points caused a decrease in NSE to 0.842. The model was resilient to low noise, but high noise (SNR = 15 dB) caused significant changes in model performance. One-way ANOVA confirmed that all three factors had a statistically significant effect on NSE: training data size ($F = 57.81$, $p < 0.001$), observation point density ($F = 23.54$, $p < 0.001$), and input noise level ($F = 31.47$, $p < 0.001$). The F-statistics were recalculated directly from the mean and standard deviation values reported in Table 8 to ensure internal consistency.

4. Discussion

4.1 Interpretation of Results

The results of the present research clearly illustrate the promising potential of hybrid models with an alternative data-driven approach to the problem of urban flooding risk assessment. Coefficient of determination equal to 0.938 and Nash-Sutcliffe value of 0.921 for the hybrid CNN-LSTM model provide not only physical-model-like accuracy but also extremely fast calculations with several thousands of times faster than in case of physical modeling. It is important to point out that the above-listed values denote interpolation for the simulated training simulations and not extrapolation for conditions that were not accounted for in the simulations. The ability of the model to predict the rainfall intensity, duration, and/or antecedent moisture content outside the ranges mentioned in Table 2 is unknown, and the above-listed accuracy can thus be regarded as an indication of good interpolation skills only and not good extrapolation skills. In addition, because the training simulations were generated using the marginal distributions described in Section 2.2, the model performance is expected to be best for simulations that fall within the central region of these distributions and may degrade for simulations near the tails. In addition, it is evident from this result that the proposed model architecture, with simultaneous consideration of the spatial pattern by means of the convolutional branch and precipitation time dynamics through the recurrent branch, possesses the inherent superiority compared to one-dimensional CNN and LSTM models.

This finding is consistent with Xu & Gao [12], who developed a hybrid surrogate model for real-time prediction of coastal urban floods and showed that hybrid models combining convolutional and recurrent neural networks outperform single-branch models in urban flood modeling. However, it should be noted that the numerical estimates of Xu & Gao [12] and Zahura et al. [3] were obtained under different study regions, input datasets, target variables, and validation procedures. In particular, the reported R^2 of 0.91 in Xu & Gao [12] and the reported NSE of 0.87 in Zahura et al. [3] were derived under different experimental conditions. Therefore, the comparison provided herein is qualitative and cannot be considered a direct comparison on a single benchmark, and the differences between these values and the values reported in the present study should not be interpreted as evidence of superiority or inferiority of any of the models.

As seen from the comparison, the performance of the suggested model is about the same as was demonstrated in earlier studies on surrogate modeling. However, because of the differences between the research settings, such as the area under study, climatic conditions, drainage systems, prediction objectives, and validation methodology, it is impossible to draw a numerical comparison. The comparison presented in Table 9 is therefore qualitative and should be interpreted as an indication of the consistency of the proposed approach with the literature rather than as a direct benchmark.

Table 9: Qualitative comparison of the proposed hybrid model with representative surrogate modeling studies

Study	Study Area	Prediction Target	Reported Metric	Reported Value	Key Differences from the Present Study
Xu & Gao [12]	Coastal urban catchment, Macao	Flood depth	R^2	0.91	Different climate and tidal boundary conditions; no transferability test; different CNN-LSTM configuration
Zahura et al. [3]	Urban coastal community, USA	Street-scale flood depth	NSE	0.87	Different urban layout and drainage system; different training data size; different hydraulic model (SWMM-based)
Present study	Chirchik urban watershed, Uzbekistan	Flood depth at observation points	$R^2 = 0.938$; NSE = 0.921	—	Continental climate; combined pluvial and fluvial flooding; transferability tested on adjacent watershed

All reported values are from the original studies and were obtained under different experimental conditions, including different study areas, climatic conditions, drainage systems, prediction objectives, and validation methodologies. The

comparison is therefore qualitative and should not be interpreted as a direct numerical benchmark. As seen from the comparison, the performance of the suggested model is about the same as was demonstrated in earlier studies on surrogate modeling. However, because of the differences between the research settings, such as the area under study, climatic conditions, drainage systems, prediction objectives, and validation methodology, it is impossible to draw a numerical comparison. The comparison presented in Table 9 is therefore qualitative and should be interpreted as an indication of the consistency of the proposed approach with the literature rather than as a direct benchmark. Thus, the numbers presented in the table below are only illustrative of the consistency of the suggested approach with the literature (Table 9).

It can also be observed that there are certain limitations in the performance of the model. According to Fig. 1, the scatter of the predicted values grows as the depth of flooding goes beyond 2 m, thus, the highest uncertainties are observed in the case of extreme flooding events. The F1-Score for the extreme depth group (0.848) is lower than that for the shallow group (0.935), which is consistent with the larger RMSE in the extreme group. This is because a small number of points in the extreme group are misclassified as non-flooded when the predicted depth falls below the 0.3 m threshold, even though the reference depth is above 2.0 m. Moreover, according to the spatial error map (see Fig. 2), the errors are the highest at the borders of the flooded areas and in the narrow drainage channels where minor variations of the input variables result in significant changes in the flood depth. Thus, the two findings point out to the fact that the most accurate predictions can be expected for the shallow and moderate depths of flooding and central part of the floodplain area. It is extremely important for public safety to note the 5.8% false negative rate of the hybrid method in comparison with the 13.5% false negative rate of the standalone LSTM network. From the point of view of flood hazard assessment, the false negative case is an occurrence when the flooded area is missed by the prediction algorithm. Such a situation results in incomplete evacuation, fatalities, and wrong decision-making during the process of disaster management. More than two times decreased number of false negatives in the hybrid method confirms that the combination of the spatial and temporal information increases the accuracy of the model's work significantly. This result corresponds with the modern research in the sphere of hydrological deep learning, which focuses on the importance of hybrid architectures. This finding is in line with Zahura et al. [3], who trained machine learning surrogate models from a high-fidelity physics-based model for real-time street-scale flood prediction. They emphasized the critical importance of minimizing false negatives in operational flood warning systems, given their direct implications for public safety and emergency response effectiveness. However, as must be pointed out, these comparisons are conducted with the same criteria and prediction objectives, but only for the current research, and the comparison to the results by Zahura et al. [3] is qualitative due to the different study area, the different hydraulic model, and different validation data.

The spatial error distribution presented in Fig. 2 provides useful information about the natural drawbacks of the alternative models. It is possible to observe the high density of large errors at the edges of the floodplain and in narrow drainage channels. The reason for such an error distribution is obvious; in these cases, small changes in input parameters may cause significant changes in the flood depth. On the other hand, a very small error in the main river channel (< 0.05 m) proves that the model performs exceptionally well in regions where there are major flows and deeper water depths. The above pattern may be used to define the application of the model, meaning that the model can be used to determine major floodplains, but further analyses may be necessary for determination of the exact flood boundary line. This finding is in line with Dang et al. [8], about machine learning-based surrogate models for urban flood depth modeling in Ho Chi Minh City, and found that errors in prediction are localized near the zones of transition from flooded to non-flooded regions, and where topographical differences give rise to steep hydraulic gradients.

The transferability of results is one of the most encouraging outcomes of the current study. First of all, the direct application of the pre-trained model in the Salar basin with NSE equal to 0.652 proves that even when using the model without any fine-tuning, we still obtain useful data. However, the most remarkable outcome is the effectiveness of the transfer learning process itself, because despite the limited number of fine-tuning simulations (only 50 compared to 850 in full training mode), the NSE value obtained is 0.861, which is very close to the result obtained when fully training the model (0.902). It should be noted that all metrics reported in Table 7 were computed on an independent test set of 25 Salar flood simulations that were not used during fine-tuning or validation. The fine-tuning simulations and the test simulations were completely disjoint, which ensures that the reported performance reflects the true generalization ability of the fine-tuned model.

Thus, it is comparable to a 94% reduction in the training data volume necessary for the deployment of the model in a new basin. This result should, however, be interpreted in the light of the morphological and climatic similarity between the two watersheds, since the Salar and Chirchik catchments share a comparable urban land-use

composition, a similar elevation range, and a similar rainfall regime. The transferability of the model to watersheds with distinctly different slopes, land-use patterns, or climate conditions therefore remains an open question and requires further testing. This is in line with Kim and Han [4], about machine learning models trained using inundation maps of a certain range of rainfall return periods can easily be transferred to other simulations with little retraining, and thus emphasizing the applicability of surrogate models in various watersheds.

Methodologically, the sensitivity analysis indicated that the amount of training data and density of observation points are the two important factors in determining the performance of the model. Cutting the number of training data from 850 to 200 simulations brought down the NSE to 0.782. This indicates that 400-500 simulations are needed for acceptable performance. Additionally, the fact that the performance of the model is relatively robust against moderate levels of noise (SNR = 25 dB) indicates that the model will perform acceptably even with noisy input data. The one-way ANOVA results confirmed that all three factors had a statistically significant effect on NSE ($p < 0.01$), with training data size having the strongest effect ($F = 57.81$), followed by input noise level ($F = 31.47$) and observation point density ($F = 23.54$).

4.2 Practical Implications

In terms of implications for the practical implementation of this research, they are important for different interested parties. For crisis managers and disaster relief organizations, the 3,882 times faster estimation makes it possible to evaluate many simulations during emergencies. It should be noted that this speedup refers to the forward pass of the network and the extraction of the predicted flood depths at the 12,800 observation points by bilinear interpolation, and does not include the preprocessing and postprocessing stages that would be required for operational deployment. For urban planners, the possibility to rapidly estimate many simulations helps make a more extensive uncertainty analysis and choose the best resilience strategy. For insurance companies and risk evaluators, the possibility to generate hazard maps quickly in other regions by transfer learning decreases assessment expenses. However, it is important to stress that the model is trained and evaluated at 12,800 fixed observation points, and the accuracy of the continuous map at arbitrary locations has not been independently validated. The continuous map is produced directly by the network, and bilinear interpolation is used only to extract the predicted values at the observation points during training and evaluation. Furthermore, the applicability of the model for flood risk management has not yet been proved since the model has been verified using artificial HEC-RAS data.

4.3 Limitations of the Study

Limitations of the present study have to be stated clearly. Firstly, the performance of the surrogate model is directly related to the performance of the physical model HEC-RAS, and the surrogate model will be as accurate as the original one. Secondly, the simulated precipitation and runoff are not exhaustive and may miss some rare cases of extreme rainfall. The 500-year return period was estimated by extrapolating the fitted GEV distribution beyond the observed rainfall record (2015–2024), and the uncertainty associated with this extrapolation is acknowledged. In addition, the hydraulic model was calibrated and validated using a limited number of observed flood events (two events), and the discharge observations were available only at the outlet of the watershed, which limits the spatial extent of the discharge-based evaluation. In addition, although the network receives all six scenario parameters as input, it may not fully capture the nonlinear interactions between these parameters in the same way as the physical model does. Thirdly, the study was performed for two watersheds with similar properties, and it is unclear how the model would perform in the case of watersheds with totally different characteristics (steep slopes or non-urban lands). In addition, the transferability experiments were conducted using a limited number of fine-tuning simulations (10, 30, and 50), and the reported metrics are based on an independent test set of 25 Salar flood simulations. A larger test set would provide a more robust estimate of the transferability performance.

Fourthly, although the model produces a spatially continuous flood depth map, its training and evaluation are confined to 12,800 fixed observation points, and the accuracy of the predicted depths at arbitrary locations has not been independently validated. Fifthly, the effects of possible changes in land use and climate change in the future were not considered. In addition, the reported inference time of 0.85 seconds refers to the forward pass of the network and the extraction of the predicted flood depths at the 12,800 observation points by bilinear interpolation, and does not include the preprocessing and postprocessing stages that would be required for operational deployment. Lastly, the model's reliability is tested through synthetic HEC-RAS results instead of actual flood events, thus requiring further validation using independent data from real-life situations. Finally, the accuracy of the model depends on interpolation in the training space, and its behavior in conditions beyond those used in the simulation is unknown. In addition, the F1-Score for the extreme depth group is lower than that for the shallow and moderate groups, which indicates that the model is less reliable for detecting the exact extent of extreme floods. This should be taken into account when using the model for emergency response planning.

4.4 Future Research Directions

Several directions for future research are possible. Firstly, the improvement of the spatial resolution of the model output beyond the current 128 × 128 grid, in order to provide more detailed flood depth maps for fine-scale urban applications. Secondly, the inclusion of the uncertainty associated with the physical model parameters in the training process of the alternative model via Bayesian methods. Thirdly, the ability of the model to generalize for watersheds with more diverse hydrological and morphological properties is analyzed. Fourthly, the creation of multi-criteria alternative models predicting not only the depth of flooding but also the flow velocity and the arrival time of the flood wave. Fifthly, the evaluation of the model in real-time operation. Sixthly, the testing of the validity of the proposed framework based on observed floods and independent hydraulic information, in order to evaluate its practicality outside the synthetic training environment. Seventh, the independent validation of the continuous spatial output at arbitrary locations beyond the 12,800 observation points, in order to assess the accuracy of the model outside the training observation network.

5. Conclusions

Theoretical and practical contributions of this study to the field of urban flood risk management include the development and validation of an innovative CNN-LSTM hybrid data-driven alternative for fast assessment of flood risk. The key innovation included showing that it is possible to combine the high accuracy of physical models and the high efficiency of machine learning-based models. The hybrid model, having the coefficient of determination of 0.938 and the Nash-Sutcliffe of 0.921, attained a level of accuracy comparable to the physical model HEC-RAS but was 3,882 times faster. This speedup refers to the forward pass of the network and the extraction of the predicted flood depths at the 12,800 observation points by bilinear interpolation, and does not include the preprocessing and postprocessing stages that would be required for operational deployment. More than doubling the decrease in the false negative rate compared to the univariate models offered a particular practical value for public safety applications. The ability to use the model with transfer learning and only 50 fine-tuning simulations allowed reducing the required amount of training data by 94%.

From an application point of view, the work presented here offers a useful tool in managing flood risks in developing cities. In Tashkent and other cities of Uzbekistan, the framework that was developed in this study can offer fast flood risk assessment at considerably reduced costs compared to conventional approaches. It becomes particularly relevant when one is dealing with insufficient computational and financial capabilities for managing flood risks. Furthermore, the use of transfer learning in this study makes it possible to quickly expand the scope of flood risk assessment to new areas. Nevertheless, it is necessary to mention several aspects that should be taken into account when interpreting these conclusions. First, although the model produces a spatially continuous flood depth map, its training and evaluation are confined to 12,800 fixed observation points, and the accuracy of the predicted depths at arbitrary locations has not been independently validated. In addition, it should be noted that the model was trained and tested on synthetic HEC-RAS simulations but not on actual flood events; in particular, external testing of the model has not been performed yet. The 500-year return period was estimated by extrapolating the fitted GEV distribution beyond the observed rainfall record, and the uncertainty associated with this extrapolation should be taken into account when interpreting the results for extreme events. In addition, transferability of the model has been verified only for a nearby basin with similar hydrological and climatic properties; thus, its applicability to other basins with different features cannot be predicted yet. Finally, it is worth noting that the presented level of accuracy refers to interpolation within the range of training simulations but does not imply the model performance for extreme events or conditions that differ from the training ones. The current results suggest that the proposed model is a promising surrogate model for fast flood depth prediction; however, they cannot prove its operational applicability yet.

Future research should thus concentrate on testing the applicability of the framework to actual flood cases, expanding the assessment of transferability to watersheds characterized by different morphology and climate, and independently validating the accuracy of the continuous spatial output at arbitrary locations beyond the 12,800 observation points. Finally, the practical application of the surrogate model within an operational flood early warning system should be tested. It is quite obvious what lies ahead – the future belongs to the intelligent combination of physical models and machine learning models.

References

[1] Tellman, B., Sullivan, J.A., Kuhn, C., Kettner, A.J., Doyle, C.S., Brakenridge, G.R., Erickson, T.A., & Slayback, D.A. Satellite imaging reveals increased proportion of population exposed to floods. Nature, 596 (2021) 80–86. https://doi.org/10.1038/s41586-021-03695-w

[2] Babamuratov B., Jumaniyazov F., Sadullaev S., Yokutoy O., Saydullayeva S., Ulug'bek E., o'g'li A., Turgunov D., Global climate change and its impact on water resources: sustainable management strategies, Procedia Environmental Science, Engineering and Management, 12(3) (2025) 969-977.

[3] Zahura, F.T., Goodall, J.L., Sadler, J.M., et al. 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 (2020) e2019WR027038. https://doi.org/10.1029/2019WR027038
[4] Kim, H.I., Han, K.Y. Inundation Map Prediction with Rainfall Return Period and Machine Learning. Water, 12 (2020) 1552. https://doi.org/10.3390/w12061552
[5] An, H., Ouyang, C. A hybrid framework for real-time flash flood forecasting in small ungauged catchments: Integrating hydrodynamic simulations with LSTM networks. Journal of Hydrology, 661 (2025) 133688. https://doi.org/10.1016/j.jhydrol.2025.133688
[6] Xu, C., Zhong, P.A., Zhu, F., et al. A hybrid model coupling process-driven and data-driven models for improved real-time flood forecasting. Journal of Hydrology, 638 (2024) 131494. https://doi.org/10.1016/j.jhydrol.2024.131494
[7] Chen, G., Hou, J., Liu, Y., et al. Urban Inundation Rapid Prediction Method Based on Multi-Machine Learning Algorithm and Rain Pattern Analysis. Journal of Hydrology, 633 (2024) 131059. https://doi.org/10.1016/j.jhydrol.2024.131059
[8] Dang, T.Q., Tran, B.H., Le, Q.N., et al. Application of machine learning-based surrogate models for urban flood depth modeling in Ho Chi Minh City, Vietnam. Applied Soft Computing, 150 (2024) 111031.
[9] Garzón, A., Kapelan, Z., Langeveld, J. Machine learning-based surrogate modelling for Urban Water Networks: Review and future research directions. Water Resources Research, 58 (2022) e2021WR031808.
[10] Lingutla, V. K., Nagamani, N., & Brahmanandam, P. S. Inferences drawn on common scale parameter of two populations using rainfall data. Proceedings on Engineering, 7(1) (2025) 251-266.
[11] Woo, H., Choi, H., Kim, M., Noh, S.J. Physics-Guided Deep Learning for Spatiotemporal Evolution of Urban Pluvial Flooding. Water, 17 (2025) 1239.
[12] Xu, L., Gao, L. A hybrid surrogate model for real-time coastal urban flood prediction: An application to Macao. Journal of Hydrology, 642 (2024) 131863.
[13] Ronnaritivichai, B., Noichan, W., & Juntanintorn, R. The Application of Community Based Disaster Risk Management Approach in Flood-prone Community. Proceedings on Engineering Science, 6(1) (2024) 119-126. 10.24874/PES06.01.014
[14] Song, Y.M., Shen, C., Liu, X. A Surrogate Model for Shallow Water Equations Solvers with Deep Learning. Journal of Hydraulic Engineering, 149 (2023) 04023045.
[15] Karim, F., Armin, M.A., Ahmedt-Aristizabal, D., et al. A Review of Hydrodynamic and Machine Learning Approaches for Flood Inundation Modeling. Water, 15 (2023) 566. https://doi.org/10.3390/w15030566
[16] Oñate-Valdivieso F., Oñate-Paladines A., Rios A., Validation of flood routing techniques in an ANDEAN micro-basin, Proceedings on Engineering Sciences, 08(1) (2026) 445-448, 10.24874/PES08.01B.002.
[17] Xu, Y., Lin, K., Hu, C., et al. Deep Transfer Learning Based on Transformer for Flood Forecasting in Data-Sparse Basins. Journal of Hydrology, 625 (2023) 129956. https://doi.org/10.1016/j.jhydrol.2023.129956
[18] Li, J., Wu, G., Zhang, Y., et al. Optimizing Flood Predictions by Integrating LSTM and Physical-Based Models with Mixed Historical and Simulated Data. Heliyon, 10 (2024) e33669. https://doi.org/10.1016/j.heliyon.2024.33669
[19] Mechentel, E., Dairi, S., Lefkir, A., Eslamian, S., Abida, H., Djebbar, Y. Data-driven flood susceptibility assessment using hybrid machine learning and optimization techniques: case of the Sedrata Watershed, NE Algeria. Scientific Reports, 16(1) (2026) 16238. https://doi.org/10.1038/s41598-026-43262-9
[20] Chen, C.-C., Sung, W.-P., Chan, H.-C., Wang, P.-T. Sustainable Urban Water Management via AI Surrogate Modeling: A Multi-Dimensional Framework for Drainage Resilience and Microclimatic Assessment. Sustainability, 18(15) (2026) 7820. https://doi.org/10.3390/su18157820
[21] Zhang, J., Guo, W., Chang, S.W., Nguyen, D.D., Ngo, H.H. Data-Driven Innovations in Flood Hazard Assessment with Machine Learning. Earth: Environmental Sustainability, 1(1) (2025) 21–41. https://doi.org/10.53941/eesus.2025.100003
[22] Akinsoji, A.H., Adelodun, B., Adeyi, Q., et al. Integrating Machine Learning Models with Comprehensive Data Strategies and Optimization Techniques to Enhance Flood Prediction Accuracy: A Review. Water Resources Management, 38 (2024) 4735–4761. https://doi.org/10.1007/s11269-024-03885-x
[23] Perera, M.D.D., Angeloudis, A., Siripatana, A., Beevers, L. Hybrid methods in flood inundation modeling: a systematic review. Natural Hazards, 122(8) (2026) 357. https://doi.org/10.1007/s11069-026-08078-w
[24] Moishin, M., Deo, R.C., Prasad, R., et al. Designing Deep-Based Learning Flood Forecast Model with ConvLSTM Hybrid Algorithm. IEEE Access, 9 (2021) 50982–50993. https://doi.org/10.1109/ACCESS.2021.3065939
[25] Ni, C., Fam, P.S., Marsani, M.F. A Data-Driven Method and Hybrid Deep Learning Model for Flood Risk Prediction. International Journal of Intelligent Systems, 2024 (2024) 3562709. https://doi.org/10.1155/2024/3562709.
[26] Brunner, G. W. HEC-RAS River Analysis System – HEC-RAS User's Manual Version 6.6. US Army Corps of Engineers, Hydrologic Engineering Center (HEC), 2024. https://www.hec.usace.army.mil/software/hec-ras/documentation/HEC-RAS_Users_Manual_v6.6.pdf
[27] Brunner, G. W. HEC-RAS River Analysis System – 2D Modeling User's Manual Version 6.6. US Army Corps of Engineers, Hydrologic Engineering Center (HEC), 2024. https://www.hec.usace.army.mil/software/hec-ras/documentation/HEC-RAS_2D_Users_Manual_v6.6.pdf
[28] McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. Technometrics, 21(2), 239–245. https://doi.org/10.1080/00401706.1979.10489755

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license.