



Deep learning with upstream-aware terrain representations and rainfall adapter for generalizable pluvial flood prediction

Thomas Lutonsky^{a,*}, Viktor Birschtzky^a, Stefan Jeschke^a,
Andreas Buttinger-Kreuzhuber^a, Nils Wandel^c, Daniel Cornel^a, Viola Haase^d,
Reinhard Klein^c, Jürgen Waser^a, Michael Wimmer^b

^a VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH, Donau-City-Strasse 11, 1220 Wien, Austria

^b TU Wien, Institute of Visual Computing and Human-Centered Technology, Wiedner Hauptstrasse 8-1, 1040 Wien, Austria

^c University of Bonn, Department of Computer Science, Friedrich-Hirzebruch-Allee 8, 53115 Bonn, Germany

^d Stadtentwässerungsbetriebe Köln, Osterheimer Straße 555, 51109 Köln, Germany

HIGHLIGHTS

- Catchment generalization of deep learning models for pluvial flood prediction.
- Upstream-aware terrain representations for capturing long-range dependencies.
- Validation on international case studies and mitigation measures.
- Two-stage training strategy with rainfall adapter for reduced training effort.

ARTICLE INFO

This manuscript was handled by Ashok Mishra, Editor-in-Chief, with the assistance of Vahid Nourani, Associate Editor

Keywords:

Pluvial flooding prediction
Deep learning
Data-driven modeling
Two-stage training

ABSTRACT

Deep learning-based surrogate models can substantially accelerate pluvial flood simulations. However, most existing approaches struggle to generalize across heterogeneous terrains or modified topographies, limiting their applicability for real-world flood risk assessment unless retrained for each specific terrain. In this work, we investigate how deep learning flood surrogates can be designed to achieve robust spatial generalization across previously unseen terrain configurations without retraining. We train an attention-enhanced U-Net on diverse terrain and introduce upstream-aware input maps that propagate terrain properties into downstream regions, allowing the model to capture long-range spatial dependencies and supporting the learning of generalizable spatial features. In addition, we add architectural design choices to improve the model's capacity to generalize beyond the training domain. We further propose a two-stage training strategy with an adapter-like rainfall encoder that enables efficient adaptation to previously unseen rainfall intensities while preserving spatial generalizability. The resulting model predicts maximum water depths and is evaluated on unseen catchments in Germany, Austria and Denmark, as well as on artificially modified terrains, such as terrains with added dikes. Our results demonstrate that combining a large training dataset with upstream-aware terrain representations and a generalization-oriented architectural design enables accurate flood predictions across unseen terrains, reaching an RMSE of 0.05 m and $CSI_{0.3}$ of 0.83 over all test domains.

1. Introduction

Pluvial flood events, i.e., floods caused by intense rainfall, cause significant damage to urban areas every year (Nicklin et al., 2019). With the increasing frequency and intensity of such events due to climate change (IPCC, 2022; Winsemius et al., 2016), the need for reliable

flood prediction and risk assessment is growing. Accurate estimation of maximal water depths during heavy rainfall events is therefore essential for identifying areas at risk and for quantifying potential damage. Such predictions can support urban planning and mitigation measures, including the design of retention basins, drainage upgrades, or targeted terrain modifications. Reliable prediction is thus indispensable for effective

* Corresponding author.

Email address: lutonsky@vrvis.at (T. Lutonsky).

risk mitigation, infrastructure design, and long-term climate-change adaptation strategies.

Physics-based hydrodynamic models are widely regarded as the gold standard for simulating pluvial flooding using the 2D shallow water equations (Cea and Costabile, 2022). These models reliably predict flood damage locations and severity and support systematic rainfall scenario analysis under varying topographic conditions, as well as flood mitigation strategies (Waser et al., 2011). However, their high computational cost – driven by the need for high-resolution topography, large spatial domains and small time steps for numerical stability – limits the speed and flexibility of analysis, making real-time forecasting and comprehensive scenario exploration challenging. Even when using high-performance computing and GPU acceleration (Buttinger-Kreuzhuber et al., 2022; Luan et al., 2023), simulations over large domains or multiple rainfall scenarios remain computationally demanding.

This motivates the development of machine-learning-based surrogate models that approximate physics-based simulations at substantially lower computational expense. Trained on simulation outputs, these models can quickly predict flood behavior for various rainfall or terrain scenarios, enabling rapid screening to identify cases that warrant detailed numerical evaluation. Deep learning (DL) (LeCun et al., 2015) has proven particularly effective, as neural networks capture nonlinear relationships and spatial dependencies. In this context, pluvial flood modeling is formulated as an image-to-image translation task. Some approaches predict the temporal evolution of flooding by modeling water depths at successive time steps (Bentivoglio et al., 2023; Kazadi et al., 2024; Chaudhary et al., 2024; Bentivoglio et al., 2025; Xu et al., 2025; Cao et al., 2025; Song et al., 2025b). Other approaches instead directly estimate the maximum water depth from the input maps. These methods either reduce computational cost by performing simulations at low resolution and subsequently applying super-resolution (Song et al., 2025a; El baida et al., 2025), or predict maximum water depth using physics-guided methods (Woo et al., 2025; Taghizadeh et al., 2025) or purely data-driven approaches (Guo et al., 2021; Löwe et al., 2021; Cache et al., 2024; Do Lago et al., 2023). In the case of purely data-driven approaches, models are trained to map rainfall intensities and static input features – such as terrain elevation, land use, and flow accumulation – to maximum water depth. Commonly employed deep learning architectures include convolutional neural networks (CNNs) (Liao et al., 2023; Woo et al., 2025), such as U-Nets (Löwe et al., 2021; Guo et al., 2021, 2022), sometimes enhanced with residual connections (Shao et al., 2024), as well as graph neural networks (GNNs) (Taghizadeh et al., 2025) and generative adversarial networks (GANs) (Hofmann and Schüttrumpf, 2021; Berkhahn and Neuweiler, 2024; Shao et al., 2024; Burrichter et al., 2023; Do Lago et al., 2023).

Several of these studies have investigated the ability to generalize to unseen rainfall events within the same spatial domain (Liao et al., 2023; Guo et al., 2021; Shao et al., 2024; Hofmann and Schüttrumpf, 2021; Berkhahn and Neuweiler, 2024; Pang et al., 2026; Burrichter et al., 2023; Cao et al., 2025). While these studies demonstrate promising performance within the given study area, extending the models to conditions outside the training domain or to modified terrains was generally not their primary focus. Only few studies have explored spatial generalization more explicitly. Löwe et al. (2021) trained a U-Net in Odense, Denmark, reserving separate spatial locations for validation, making it one of the earlier works to explicitly consider spatial generalization. Similarly, Cache et al. (2024) trained a multi-resolution model on data from Zurich and demonstrated that accurate predictions for multiple rainfall events in Lucerne and Singapore can be achieved by fine-tuning the model on the respective terrain using only a single rainfall event. Seleem et al. (2023) used a U-Net and a Random Forest to generalize between three catchments in Berlin, Germany, while Do Lago et al. (2023) employed a conditional GAN trained on multiple catchments in San Antonio, Texas. Guo et al. (2022) trained a U-Net to predict maximal water depths and velocities for a single rainfall event using data from a large number of catchments within the Canton of Zurich, reserving some

catchments to evaluate spatial generalization to terrain not seen during training. This study is among the few to consider such a wide variety of terrains, representing an important step toward reliable generalization.

As a result, current deep learning surrogates for estimating maximal water depths for a given rain event perform well within regions seen during training but struggle on unseen terrains (Cache et al., 2024; Seleem et al., 2023) or altered topographies – such as constructed dikes and retention basins – limiting their practical applicability despite their substantial computational speed-ups. This motivates our central research question: How can deep learning-based flood surrogates be designed to achieve robust spatial generalization across unseen terrains and modified topographies without retraining? We investigate four key challenges:

- What extent of spatially distributed training data are required to enable robust transfer to entirely unseen terrain configurations?
- How can feature engineering and input representations be designed to capture long-range spatial dependencies and improve generalization across catchments?
- Which architectural design choices most effectively enhance the model's ability to generalize beyond training data?
- Can the model be efficiently adapted to previously unseen rainfall events without losing spatial generalizability?

To address these challenges, we make the following contributions: First, we explore the data requirements for achieving robust performance, by exposing the model to a wide range of spatial patterns, allocating distinct regions for training, validation, and testing. Second, we introduce input maps that propagate terrain properties into downstream regions, enabling the model to capture long-range spatial dependencies. To the best of our knowledge, this is the first approach to include additional downstream-propagated terrain properties, rather than relying solely on flow accumulation. Third, we investigate architectural design choices to enhance the model's capability to consistently represent various spatial patterns. Specifically, we analyze the effect of incorporating attention mechanisms into the skip connections and the bottleneck, evaluating their impact on feature aggregation and overall model performance. Fourth, we introduce a two-stage training strategy: (i) training is performed on a single rainfall event to focus on learning robust spatial features that generalize to unseen domains; (ii) an adapter-like rain encoder is used to enable efficient rain fine-tuning, maintaining robust performance on additional rain events, with smaller amounts of reference simulation data. Together, these contributions form the basis of a surrogate model designed to predict maximum water depths and to support strong spatial generalization across unseen terrain configurations.

After training the model in Rhineland-Palatinate (RLP), a federal state of Germany, we evaluate it within unseen regions in RLP as well as in Austria and Denmark to examine its ability to maintain strong performance far outside the training area. Additionally, we evaluate performance in estimating changes in maximal water depth caused by artificial terrain modifications to investigate its potential for practical applications in flood risk assessment and management. After applying the two-stage training procedure, we apply the model to artificial and natural rain events with varying intensities and shapes to assess its capability to adapt to varying rainfall conditions. Finally, we investigate how different factors – including a large training dataset and careful input map design – contribute to reliable generalization and support accurate predictions in unseen terrains.

2. Methodology

We present a methodological framework for developing, training, and evaluating a model that predicts maximal water depth during pluvial flood events. We begin by describing the data and simulation setup, including the input datasets and the generation of ground truth through numerical hydrodynamic simulation. Afterwards our input maps and

preprocessing pipeline are introduced, which encode terrain characteristics to support accurate predictions with particular emphasis on capturing long-range dependencies. We then detail the machine learning model architecture and training procedure, both motivated by the need for effective spatial reasoning and stable training. Building on these components, an initial training stage produces a model designed to generalize to unseen domains under fixed rainfall conditions; a second training stage then extends the model's applicability to varying rainfall scenarios by augmenting it with a dedicated rain adapter module and a lightweight fine-tuning process. Finally, we describe the evaluation strategy, including the metrics and validation protocols used to assess model performance in a rigorous and transparent manner.

2.1. Data & simulation

To examine how training data size enables model generalization to unseen domains, we employ a large dataset covering a wide geographic area of RLP. The scale of the dataset exceeds that used in previous studies (Do Lago et al., 2023; Löwe et al., 2021; Cache et al., 2024; Seleem et al., 2023; Guo et al., 2022) and allows us to investigate how the number of training domains affects the model's ability to learn transferable relationships that can be applied to previously unseen terrains.

Spatial domain. Fig. 1(a) shows the study region in RLP, covering approximately 20,000 km². The region was subdivided into 148 simulation domains derived from catchments. Catchment size was artificially limited to 20 km² by marking pixels with larger contributing area as invalid, thereby excluding pixels belonging to larger rivers and associated fluvial flooding from the simulations. Multiple catchments were then combined into a single simulation domain (see Figure S1 in Supplementary Section A.1 for a more detailed description of the domain construction). The resulting domains were divided into 120 training, 14 validation, and 14 test domains. In addition, two evaluation domains outside RLP were analyzed to assess model transferability across different geographic regions. The first domain is located near Odense, Denmark (~173.86 km²), a relatively flat region previously used by Löwe et al. (2021), while the second is located in the vicinity of Mariazell, Austria (~99 km²) to provide a contrasting mountainous setting. As summarized in Table 1, the elevation values in both domains lie outside the distribution observed during training. To ensure comparable scales, we adjusted the height field by shifting its mean to match the mean elevation of the training data. This adjustment is permissible since the target water depth is independent of absolute elevation. Pixels corresponding to the sea in the area of Odense were declared invalid. The boundaries of both evaluation domains were generated using a

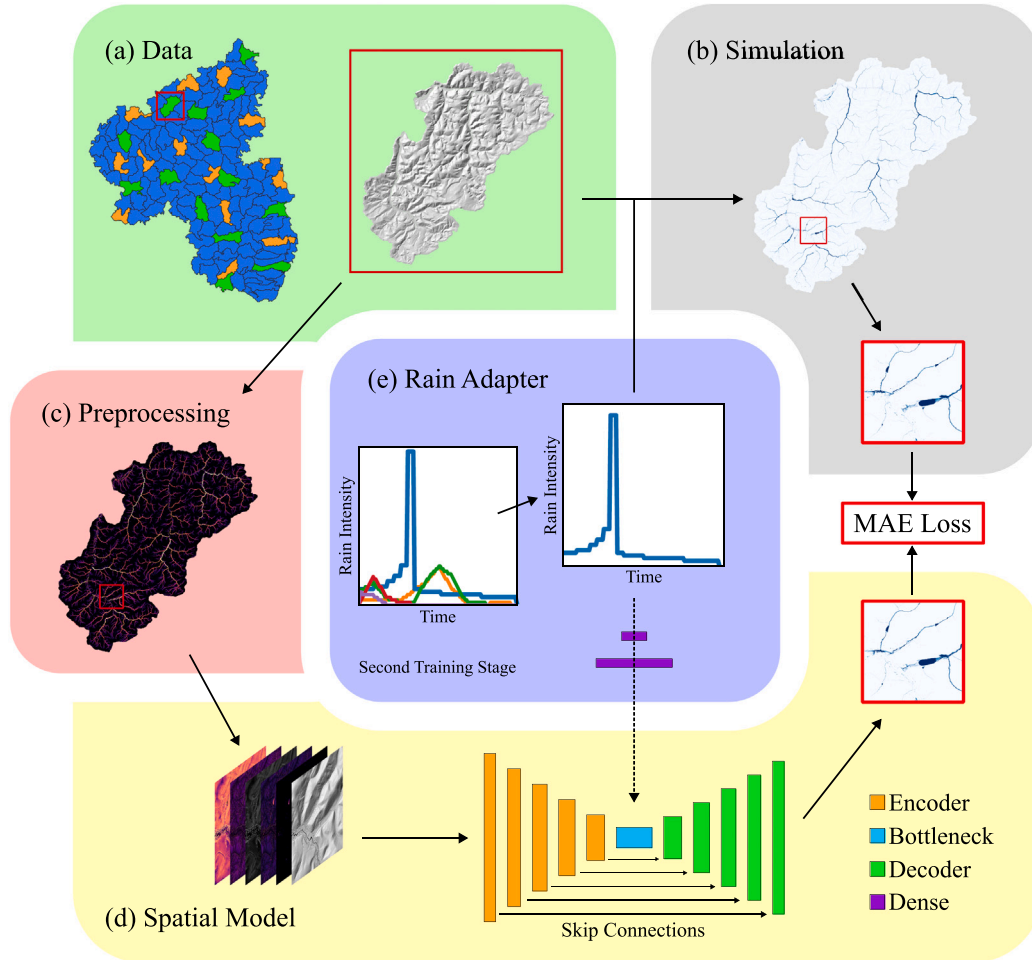


Fig. 1. AI-driven pluvial flooding prediction workflow. The workflow includes the single-event model M_{single} and the multi-event model M_{multi} , which extends M_{single} with a rainfall adapter. (a) Spatial partitioning of catchments in Rhineland-Palatinate (RLP) into training (blue), validation (orange), and test (green) domains (left). Terrain data of a simulation domain (right). (b) Simulated maximum water depth for the simulation domain (top) and extracted patch (bottom), used as ground truth for model training. (c) Example input feature map derived from the domain's terrain. (d) Neural network model M_{single} processing spatially patched input maps: the encoder (orange), bottleneck (blue), and decoder (green) generate a spatial prediction of maximum water depth. (e) Rainfall input workflow: hyetographs of various rainfall events, the selected hyetograph, and the rain adapter architecture (purple), which integrates rainfall data into the spatial model to extend M_{single} to the multi-event model M_{multi} . (For interpretation of the references to colour in this Figure legend, the reader is referred to the web version of this article.)

Table 1

Height statistics for Odense, Mariazell and RLP training domains.

Location	Range [m]	Mean [m]
Odense, Denmark	2–120	27
Mariazell, Austria	750–1200	910
RLP, Germany: Training Data	36–830	324

catchment-derivation algorithm, since, in contrast to RLP, no catchment data were available.

Ground-truths. Fig. 1(b) presents the simulation output—the maximum water depths—including an inset highlighting a subregion of the simulated domain. Hydrodynamic simulations were performed with *scenariFY*, a computational framework that solves the shallow-water equations using finite-volume schemes (Buttinger-Kreuzhuber et al., 2019, 2022). The simulations employed a uniform spatial resolution of 2 m since finer resolutions would lead to prohibitively long computation times, whereas the neural network is designed to provide a rough estimate rather than fine-scale detail. To minimize potential sources of error and focus on generalization to unseen terrains, we use a spatially uniform Manning roughness coefficient ($0.0313 \text{ s/m}^{1/3}$) and neglect infiltration during ground truth generation (see also Do Lago et al. (2023)). While neglecting infiltration leads to less realistic simulations and generally higher maximum water depths, it simplifies the underlying flood dynamics and reduces additional sources of variability. At the same time, the resulting higher water depths can be interpreted as a worst-case scenario, where the soil can no longer absorb any surface water. Cells outside the domain were excluded from simulation and model prediction, with boundary conditions defined to allow water to freely flow out at each time step. This allows for domain outflow while restricting the simulations to pluvial flooding, as fluvial flooding associated with larger river catchments is excluded based on the contributing-area threshold defined above. For the initial conditions, we assume zero initial water depth. Filled lakes and canals are represented as part of the terrain elevation. Similar to Cache et al. (2024); Do Lago et al. (2023); Löwe et al. (2021); Guo et al. (2022, 2021), the simulations were driven by spatially uniform rainfall events. These include a 2 h, 100-year Euler Type II event derived from KOSTRA-DWD-2020 (Jungähnel et al., 2023) – a dataset containing spatially interpolated design rainfall statistics provided by the German Weather Service (Deutscher Wetterdienst) – for RLP, additional 1-hour KOSTRA-DWD-2020 events with varying return periods and amplification factors, as well as natural rainfall patterns from the data provided by Löwe et al. (2021). Each simulation covered a period of 4 h, producing maximum water-depth values across the domain, which were subsequently extracted as ground-truth data. The simulation end time is deliberately chosen to be significantly longer than the rainfall event to capture the runoff that continues after the rainfall has ended. The simulations capture long-range dependencies across the entire domain, with complex flow patterns arising from interactions between local terrain, neighborhood retention effects, and upstream contributing areas.

2.2. Input maps & preprocessing

We designed input maps to systematically capture hydrologically relevant information across multiple spatial scales, from local terrain features to upstream catchment contributions. All input maps were derived from the terrain using the open-source library WhiteboxTools (WBT) (Lindsay, 2017), with Table 2 summarizing each channel, its transformation, and a brief description. Each map is generated for the entire domain, providing a global representation of terrain and hydrological characteristics, as exemplified in Fig. 1(c). The input features can be conceptually organized into three levels: *terrain*, *semilocal*, and *global* descriptors. This hierarchical structure ensures that the model receives

Table 2

Overview of input maps used by our model with the corresponding transformations applied during preprocessing and a short description of the input map.

Channel	Transf.	Notes
Terrain	None	Elevation relative to sea level.
Slope	Square-root	Magnitude of terrain slope.
Sink Depth	Reciprocal	Height distance to outlet in local sink. Zero outside of sinks.
FLACC	$1/(1+x)$	Flow accumulation map.
Average	$\log_{10}$	Mass flux of slope over upstream area; not used in prior studies.
Upstream Slope	$\log_{10}$	Mass flux of sink depth over upstream area; not used in prior studies.
Upstream		
Capacity		

complementary information, ranging from point-wise terrain properties to aggregated upstream context, enabling it to capture both fine-scale and catchment-scale flow dependencies.

Terrain – referring to the digital elevation model (DEM) – characterizes geometric properties at each grid cell independently of its surroundings. This feature encodes elevations, which govern local flow direction and immediate runoff behavior. As such, the terrain provides fundamental point-wise geometric information that is essential for predicting the local distribution of water accumulation. By capturing fine-scale terrain characteristics, local features form the basis upon which more complex, scale-dependent representations can be built.

Semilocal features capture terrain characteristics that depend on the neighborhood surrounding each cell. In our study, this is represented by the *Slope* (calculated using `slope` from WBT) and *Sink Depth* – defined as the height difference from the outlet within a local sink (calculated using `depth_in_sink` from WBT). These features encode retention and storage potential, providing context beyond point-wise terrain information and allowing the model to account for localized water accumulation and flow retention. By representing neighborhood-scale storage effects, *Slope* and *Sink Depth* contribute additional spatial context between point-wise terrain properties and catchment-scale influences.

Catchment-scale features encode aggregated upstream influence and capture long-range spatial dependencies across the entire domain. These features enable the model to account for how upstream terrain, slope, and storage characteristics affect water accumulation and flow at each location. In this study, we considered three such features: *Flow Accumulation (FLACC)*, *Average Upstream Slope*, and *Upstream Capacity* (all calculated using `d8_mass_flux` from WBT after filling the terrain using `fill_depressions` from WBT). The general expression for calculating the upstream-aware maps is given by the *mass flux*:

$$F_i = \left(L_i - \alpha_i + \sum_{j \in \Omega_i} F_j \right) \eta_i, \quad (1)$$

where L_i is the local property to be transported, α_i is local absorption (set to zero for all maps in this study), $\sum_{j \in \Omega_i} F_j$ is the total inflow from all upstream cells Ω_i draining directly into cell i . We set the efficiency η_i to 0 for cells i that are invalid in the simulation domain and 1 for valid cells. This choice mirrors the behavior of the runoff simulation, where water is removed at invalid cells.

The choice of L_i determines the type of catchment-scale feature:

- **FLACC:** $L_i = 1$, measuring the contributing area of each cell. This quantifies the potential amount of runoff accumulating at a given location.
- **Upstream Capacity:** L_i is set to Sink Depth, propagating information about the volume of water that can be captured by upstream depressions.
- **Average Upstream Slope:** L_i is set to the local slope. The map encodes upstream terrain steepness and its influence on flow velocity and runoff dynamics. In this case, the resulting flux is divided by

the FLACC to provide the mean upstream slope, giving the model a measure of the overall steepness of upstream terrain.

By propagating these properties downstream, catchment-scale features provide the model with a multi-scale representation of how upstream terrain and storage characteristics influence water distribution.

Preprocessing and normalization. After feature generation, each channel is transformed and normalized using min-max scaling based on statistics computed from the training set. Following Löwe et al. (2021), the target maximum water depth is square-root transformed to reduce skewness. These preprocessing steps ensure that inputs and targets are represented on comparable numerical scales, which facilitates stable and efficient model training.

2.3. Model architecture

Fig. 1(d) presents a schematic of the network corresponding to the base model M_{single} , while Fig. 1(e) illustrates the rainfall-conditioned extension M_{multi} (see Supplement Section A.2 for more details). The architecture is based on a U-Net with attention in the bottleneck (Petit et al., 2021) and attention gates at the skip connections (Oktay et al., 2018). Key hyperparameters are summarized in Table 3. The architectural design was guided by the following challenges:

Long-range dependencies. Rainfall response in catchments exhibits long-range spatial dependencies, as upstream terrain strongly influences downstream water accumulation. These interactions are partially captured through the input maps. However, the model must additionally ensure effective communication between all locations within an input patch, which is not guaranteed by fully convolutional neural networks – depending on the model parameters and patch size – due to their limited effective receptive field (Luo et al., 2017). We introduce a bottleneck with multi-head self-attention (see also Petit et al. (2021); Cache et al. (2024); Natarajan et al. (2026)), enabling information to propagate across distant spatial positions within the input patch. Relative positional encoding (Shaw et al., 2018) preserves translation invariance while maintaining pairwise positional information. To provide adequate context for attention, we feed the network large input patches, which imposes memory constraints and motivates the use of group normalization instead of batch normalization.

Sparse gradient signals and training efficiency. Flood-relevant regions occupy only a small fraction of the domain, leading to sparse and noisy gradient signals. To facilitate training, residual connections are added to each double convolution block, creating shortcuts that allow the input signal to bypass intermediate layers. These shortcuts preserve original information and improve gradient flow (He et al., 2016). Additionally, attention gates (Oktay et al., 2018) in the skip connections between encoder and decoder suppress irrelevant features and emphasize important regions, enhancing the signal-to-noise ratio of gradient updates and improving overall training efficiency.

Rainfall conditioning. The architecture described above defines the base model M_{single} , which is trained on data from a single rainfall event. The pretrained model is then extended with a lightweight conditioning module, the *Rainfall Adapter* (see Fig. 1e), to obtain M_{multi} . Rainfall characteristics are encoded by a small multilayer perceptron and injected via Feature-wise Linear Modulation (FiLM) (Perez et al., 2018), which applies learned affine transformations to the feature channels, enabling rainfall-specific modulation while preserving the pretrained backbone.

2.4. Training procedure

The proposed model is trained in two stages. First, the base model M_{single} is trained using simulations generated from a single rainfall event, allowing the network to learn robust spatial representations that

Table 3

Key hyperparameters of the attention-enhanced U-Net model.

Hyperparameter	Value
Depth	5
Number of initial features	64
Number of attention heads	8
Dropout	0.01

generalize across catchments. In a second stage, the rainfall-conditioned model M_{multi} is obtained by fine-tuning the network using multiple rainfall scenarios. The following subsections describe both stages.

2.4.1. Spatial representation learning with M_{single}

We train M_{single} using the input maps from the 120 domains selected for training and the corresponding ground truths by simulating a 2 h, 100-year event from KOSTRA-DWD-2000 (Junghänel et al., 2023). The following paragraphs describe the details of the training procedure:

Patch sampling. During training, patches of size 1020×1020 pixels ($2040 \text{ m} \times 2040 \text{ m}$) are sampled from random spatial locations across all training domains. Patches containing more than 10% invalid pixels are discarded, as they provide limited useful information. Invalid cells in the input maps are set to zero, while the corresponding target cells are excluded from loss computation and evaluation metrics. For validation, each domain is subdivided using a structured grid to ensure consistent evaluation across epochs and training runs.

Stability. Due to the large patch size, GPU memory constrains the batch size to 4. Together with the skewed distribution of target values, this small batch size necessitates additional mechanisms to stabilize training. We therefore adopt the mean absolute error as the loss function:

$$\mathcal{L}(y, \hat{y}) = \frac{1}{N} \sum |y - \hat{y}|,$$

which is robust to outliers and avoids large, destabilizing gradients. Optimization is performed using the Adam algorithm (Kingma and Ba, 2014), providing stable and efficient parameter updates during training. The learning rate is gradually reduced from 10^{-4} to 10^{-5} via cosine annealing (Loshchilov and Hutter, 2016), promoting smooth and stable convergence.

Generalization. We apply two additional strategies to promote generalization: (i) We apply *data augmentations* to increase patch diversity and discourage orientation-specific features (see also Do Lago et al. (2023); Cache et al. (2024)) using the 8 symmetry transformations generated by 90° rotations and vertical flips. These transformations are label-preserving for raster inputs, as applying the same transformation to both input and target yields a valid training pair. (ii) To obtain more stable evaluation behavior, we maintain an exponential moving average (EMA) of the model parameters with a decay rate of 0.999 (Morales-Brotons et al., 2024). We maintain two sets of weights: the original network parameters for training and an EMA version, which is updated throughout training and used for validation and testing. This approach can be viewed as a temporal ensemble of recent model states, which tends to favor flatter regions of the loss landscape, thereby improving the stability and generalization of the evaluated model.

After applying this training procedure for 400 epochs with 10,000 patches per epoch, we obtain the model M_{single} , trained to predict maximal water depths for the rain event used during training.

2.4.2. Rainfall adaptation with M_{multi}

The second training stage enables efficient adaptation to additional rainfall scenarios by decoupling spatial and rainfall variability.

While M_{single} learns transferable spatial representations in the first stage, the rainfall adapter extends the model to additional rainfall conditions without retraining the full network, reducing computational and data requirements. Besides the 2 h 100-year Euler Type-II event used in the first training stage, the dataset for training M_{multi} includes 1 h Euler Type-II events with varying return periods from KOSTRA, a 100-year event with an amplification factor of 2, and natural rainfall patterns with varying temporal characteristics derived from Löwe et al. (2021).

Training setup. In contrast to the first training stage, we use only 8 out of the 120 available spatial domains (see Figure S5 in Supplement Section A.5) to reduce the number of simulations required to generate training data. For each sample, we randomly select an input patch and a hyetograph as well as the corresponding target patch. The model is fine-tuned for 15 epochs using 10,000 samples per epoch. During this stage, encoder weights are frozen to preserve the spatial feature representations learned during the first stage.

2.5. Evaluation

After training, we evaluate the model across a range of scenarios to ensure robust performance in predicting maximum water depths. The domains are divided into patches with a 50% output overlap to mitigate boundary artifacts. Predictions from overlapping patches are combined using the arithmetic mean at each spatial location to reconstruct the full domain, and evaluation metrics are computed on the reconstructed outputs. Only in Section 4, do we use patches without output overlaps to analyze the model's performance within individual patches.

To provide a comprehensive assessment, we employ several complementary metrics, each designed to capture different aspects of model performance (see Table 4).

The root mean-squared error (RMSE_{thr}) is used as a global error metric due to its sensitivity to outliers, providing insight into the distribution of prediction errors and highlighting extreme deviations.

The Critical Success Index (CSI_{thr}) evaluates the model's ability to detect pixels exceeding a given threshold.

At low thresholds, the CSI indicates the model's ability to discriminate between wet and dry pixels. At higher thresholds, it reflects the capability to detect extreme water depths.

2.6. Technical implementation

The model was implemented in Python version 3.12.0 using PyTorch version 2.8.0. Training, evaluation, and timings for the physical simulation in Section 3.6 were performed on an NVIDIA GeForce RTX 5090. In order to utilize the GPU's performance on certain data types and reduce memory consumption, training was performed on data in float16.

Table 4

Evaluation metrics. The symbols y_i and $\hat{y}_i$ denote the ground-truth and predicted maximum water depth at pixel i , respectively. N denotes the total number of evaluated pixels, and thr denotes the given threshold. We denote H_{thr} as the number of pixels i where $y_i > \text{thr}$ and $\hat{y}_i > \text{thr}$ (hits), M_{thr} as the number of pixels where $y_i > \text{thr}$ and $\hat{y}_i \leq \text{thr}$ (misses), and FA_{thr} as the number of pixels where $y_i \leq \text{thr}$ and $\hat{y}_i > \text{thr}$ (false alarms).

Metric	Range	Best
$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$	$[0, \infty)$	0
$\text{RMSE}_{\text{thr}} = \sqrt{\frac{1}{ H_{\text{thr}} } \sum_{i: y_i > \text{thr}} (y_i - \hat{y}_i)^2}$	$[0, \infty)$	0
$\text{CSI}_{\text{thr}} = \frac{H_{\text{thr}}}{H_{\text{thr}} + M_{\text{thr}} + FA_{\text{thr}}}$	$[0, 1]$	1

3. Results & discussion

3.1. Performance evaluation

Table 5 summarizes the performance of the single-event model M_{single} in predicting maximum water depth on the training, validation, and test datasets, as well as the case studies discussed in the following sections. Performance decreases from the training to the validation and test datasets, consistent with expected behavior. CSI values of approximately 0.8 and RMSE values near 0.05 m indicate robust predictive performance on unseen domains, while the close agreement between validation and test metrics suggests that the validation set provides a reliable estimate of generalization performance.

The results across different maximum water depth thresholds provide further insights into model behavior. The CSI values for the training dataset remain relatively stable, whereas the CSI values for the validation and test datasets decrease as the threshold increases, which indicates the challenge of accurately predicting high maximum water depths, which is also reflected by the gap between RMSE and $\text{RMSE}_{0.1}$. This is expected because the dataset is dominated by dry cells, resulting in a highly skewed and imbalanced distribution.

Meaningful direct comparison with previous studies is challenging because of substantial differences in simulation configurations, training strategies, and evaluation protocols. Spatial resolutions range from 1 m (Cache et al., 2024) and 2 m (Guo et al., 2022) to 3 m (Do Lago et al., 2023) and 5 m (Löwe et al., 2021), whereas our simulations use a resolution of 2 m. The treatment of infiltration also varies: Löwe et al. (2021) use spatially distributed infiltration to simulate, Cache et al. (2024) and Guo et al. (2022) assume uniform infiltration, whereas Do Lago et al. (2023) and the present study neglect infiltration. Training data composition differs substantially as well. Cache et al. (2024) trained on a single catchment ($\sim 12.7 \text{ km}^2$), while Löwe et al. (2021) used a single, larger catchment ($\sim 135 \text{ km}^2$) from which separate training and validation patches were sampled, with both studies considering many rainfall events. Do Lago et al. (2023) trained on multiple spatially distributed catchments ($\sim 28.13 \text{ km}^2$) using four rainfall events, whereas Guo et al. (2022) used numerous catchments ($\sim 4000 \text{ km}^2$, estimated from the information provided in the manuscript) but only a single rainfall event. In contrast, our framework is first trained on 120 domains ($\sim 17,600 \text{ km}^2$) using a single rainfall event to obtain M_{single} , before being fine-tuned on 10 additional rainfall events across 8 training domains (see Section 3.4).

Consequently, existing studies are best regarded as contextual references rather than direct benchmarks. Models trained on a single domain generally report CSI values below 0.6 when evaluated on unseen terrains Löwe et al. (2021); Cache et al. (2024), whereas the multi-catchment approach of Do Lago et al. (2023) achieves CSI values of up to approximately 0.8. Although these results are not directly comparable, they indicate that training across diverse domains substantially improves spatial generalization. Our results are consistent with this trend and further demonstrate effective generalization to previously unseen terrains.

Some domains, such as the mountainous Test Domain T_1 and Mariazell, achieve lower error metrics and higher CSI values compared

Table 5

Performance of M_{single} for maximal water depth prediction. Results are shown for the overall training, validation, and test sets in RLP. Lower RMSE and higher CSI indicate improved performance.

Dataset	RMSE [m] ↓	$\text{RMSE}_{0.1}$ [m] ↓	$\text{CSI}_{0.05}$ ↑	$\text{CSI}_{0.1}$ ↑	$\text{CSI}_{0.3}$ ↑	$\text{CSI}_{1.0}$ ↑
RLP Training	0.025	0.084	0.883	0.900	0.908	0.894
RLP Validation	0.044	0.149	0.867	0.866	0.834	0.808
RLP Test	0.050	0.159	0.863	0.853	0.828	0.786
RLP Test Domain T_1	0.024	0.110	0.890	0.908	0.913	0.867
RLP Test Domain T_2	0.052	0.125	0.873	0.857	0.780	0.759
Odense	0.043	0.102	0.868	0.863	0.802	0.647
Mariazell	0.028	0.136	0.874	0.897	0.881	0.826

to the aggregated performance in RLP Test, indicating that their characteristics are well captured by the model. In contrast, the flatter test domain T_2 and Odense exhibit mixed behavior, with comparable or improved RMSE and RMSE_{0.1} values but reduced CSI values at higher thresholds compared to the aggregated performance in RLP Test. This reduced performance in flatter terrain may partly be attributed to greater influence of micro-topography, which can cause upstream-aware terrain properties to be less informative, or even potentially misleading. This suggests that prediction quality, particularly for higher maximum water depths, depends on the individual characteristics of the terrain. This dependency is further examined in the following section.

3.2. Case studies

3.2.1. Evaluation in RLP

We complement the global evaluation with a more detailed analysis for two test domains in RLP. As in the previous chapter, we use M_{single} and the corresponding single rain event to reduce computational costs while focusing on the model's spatial generalization capability.

Test domain T_1 – shown in the inset of Fig. 2(a) – is a mountainous region. The scatter plot shows a high density of points along the identity line, indicating strong agreement between predictions and simulations up to a maximum water depth of approximately 8 m, with no significant bias toward over- or underestimation. The corresponding boxplots reveal that the error remains close to zero for pixels with simulated water depths between 0.1 m and 0.5 m. For higher depths, the error distribution widens; however, even for pixels where the ground truth exceeds 3.5 m, 50% of predictions still exhibit an error below 0.25 m.

Test domain T_2 – shown in the inset of Fig. 2(b) – is comparatively flatter and exhibits less pronounced terrain structure. In this case, the scatter plot is more dispersed and indicates a slight tendency toward underestimation for water depths above 3 m. Nevertheless, a high density of points remains along the identity line, indicating overall good agreement between simulation and prediction. As in the first domain, water depths below 0.5 m are predicted accurately, as depicted in the boxplot in Fig. 2(b). For higher water depths, the task becomes more challenging, leading to larger errors. Still, for depths up to 2 m, the majority of pixels exhibit errors below 0.5 m, with noticeable underprediction occurring primarily at higher values.

A closer look at the highlighted region (see Fig. 2c) provides a specific example of the model's performance. The largest errors occur at a location where multiple flows meet and are surrounded by complex building structures. Such structures create narrow passages and retentive behavior, making these cases more challenging for prediction. Despite this challenging scenario, the model successfully captures the locations of deep water, highlighting its reliability even in complex cases.

3.2.2. International case studies

Fig. 3(a) provides an overview map of Central Europe, highlighting the training region and the test domains in Mariazell (Austria) and Odense (Denmark). The geographical separation between these areas illustrates their suitability for evaluating the model's applicability to domains that differ substantially from the training data.

Odense, Denmark. The region shown in Fig. 3(b), located in Odense, is relatively flat and comprises both urban and rural areas. The scatter plot shows a relatively strong clustering around the identity line. Despite the moderate spread reflecting prediction variability, we see that the model captures the overall relationship between the predicted and simulated values. When focusing on a smaller rural subregion, the model closely reproduces the reference solution, as reflected by consistently small prediction errors across the area. The spatial distribution of maximal water depths, including numerous small and deep puddles, is captured well by the model.

Mariazell, Austria. The scatter plot for the mountainous region in Fig. 3(c) shows a strong correspondence between predicted and

simulated water depths except for a single offset diagonal structure (also see Supplement Section A.4.2). Compared to Odense, the spread of errors is reduced, supporting the observation that flat terrain poses a greater challenge for the model. The spatial comparison further confirms that the model captures the distribution and structure of water depths consistently across varying elevations, highlighting its robustness when applied to previously unseen terrain.

3.3. Measures-awareness

Terrain modifications, such as the addition of basins and dikes, can produce disproportionately large changes in maximum water depth relative to their spatial extent. Accurately predicting the maximum water depths for scenarios with and without these measures, and consequently capturing the resulting changes in maximum water depth, poses a significant challenge for the model. As in previous evaluations, we use M_{single} and the corresponding single rainfall event to reduce computational costs while isolating the model's response to localized terrain modifications.

In Fig. 4(a), we see that M_{single} demonstrates reasonable predictions for the maximal water depth with ($\hat{h}_{\text{on}}$) and without the dike ($\hat{h}_{\text{off}}$): Water accumulates upstream, while the dike holds back the first major flood, leading to reduced maximum water depths downstream. A direct comparison between the simulation results and the predictions for both cases is provided in Figure S8 in Supplement Section A.7.

In Fig. 4(b), we visualize the model's ability to estimate the *response* to the dike, defined as the difference between the simulation results with and without the measure. Within the patch where the terrain was modified, the predicted response ($\Delta\hat{h} := \hat{h}_{\text{on}} - \hat{h}_{\text{off}}$) closely matches the reference response ($\Delta h := h_{\text{on}} - h_{\text{off}}$) accurately capturing both the water accumulation upstream of the dike and the reduction in maximum water depths downstream. In order to isolate the effect of the upstream capacity, we run the model without updating this input map (while including the modifications in all other inputs). The predicted maximal water depth $\tilde{h}_{\text{on}}$ and corresponding response ($\Delta\tilde{h} := \tilde{h}_{\text{on}} - \hat{h}_{\text{off}}$) show that the model still predicts water accumulation upstream of the dike, but the downstream impact is completely misrepresented, and beyond the patch borders, no response is visible. This highlights the critical importance of the upstream capacity map for modeling the effects of retained floodwater, as commonly encountered in mitigation measures.

To avoid drawing conclusions from a single example, we introduced eight terrain modifications in the RLP test areas in the form of retention basins, breaches and dikes to further assess the model's performance (see Figures S9 and S10 in Supplementary Section A.7 for visualizations). Table 6 summarizes the quantitative performance of the model across all artificial terrain modifications in the RLP test domains. The response error ($RMSE_{\Delta} := RMSE(\Delta h, \Delta\hat{h})$) is highest for dikes, indicating that their pronounced influence on flood dynamics is the most difficult for the model to capture. The RMSE values are higher than those reported for the case studies because the evaluation is restricted to pixels affected by the terrain modifications in either the simulation or the prediction. Consequently, the analysis focuses exclusively on the most challenging regions. Comparing the RMSE and CSI values for predictions with and without the respective terrain modifications reveals improved performance for scenarios with lower maximum water depths, namely those including dikes or retention basins and those excluding breaches.

The scatter plot in Fig. 4(c) shows strong agreement between predicted and reference values for most pixels, indicating reliable performance in the scenarios with modified terrain. The individual scatter plots for each scenario can be seen in Figures S9 and S10 in Supplement Section A.7.

3.4. Rain adapter fine-tuning

Fig. 5(a) shows the rain events used for training and evaluating M_{multi} , covering a wide range of rainfall intensities and temporal patterns.

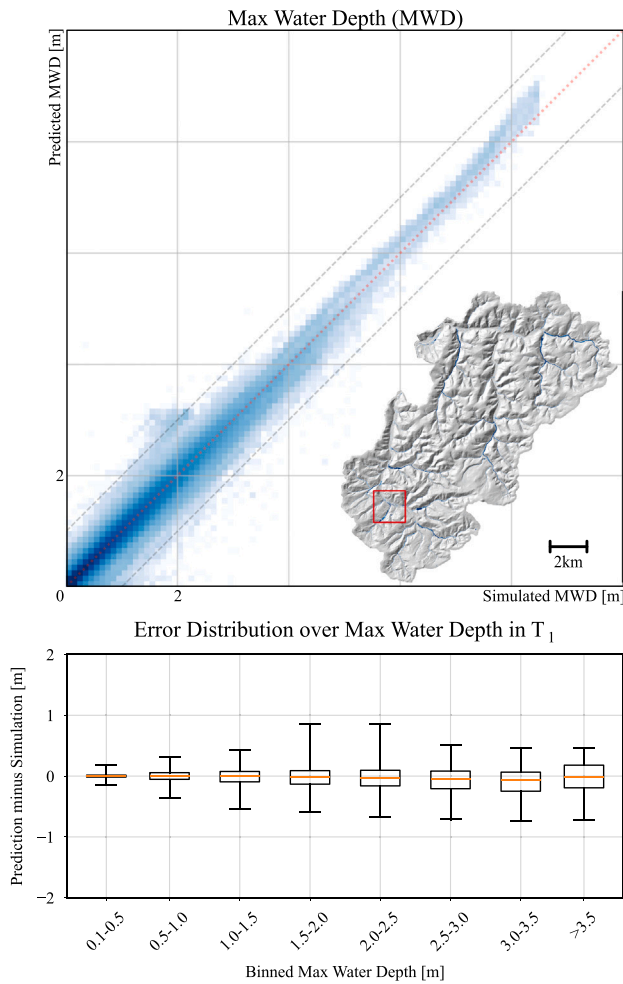
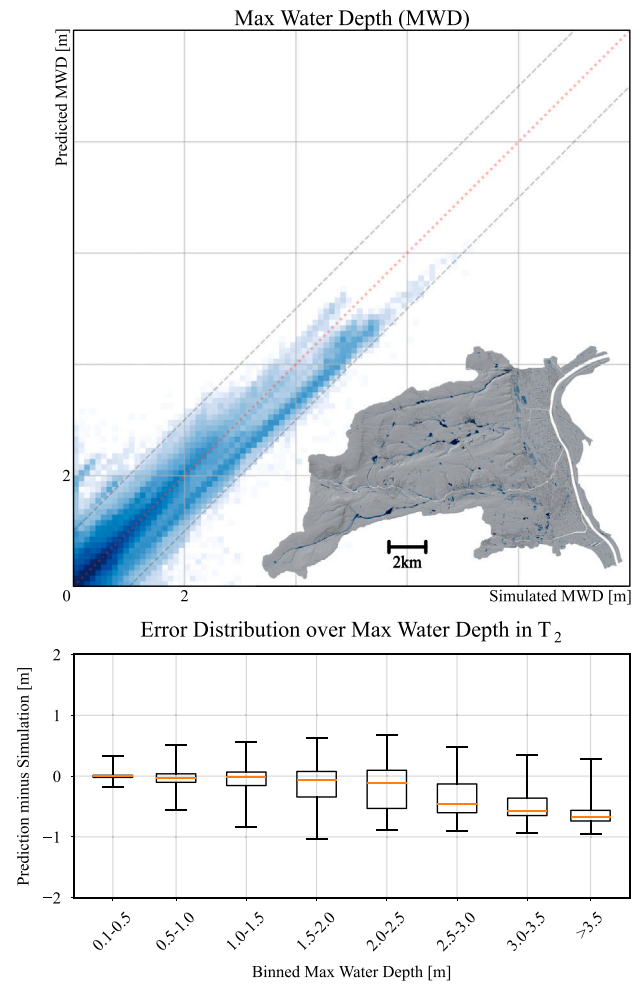
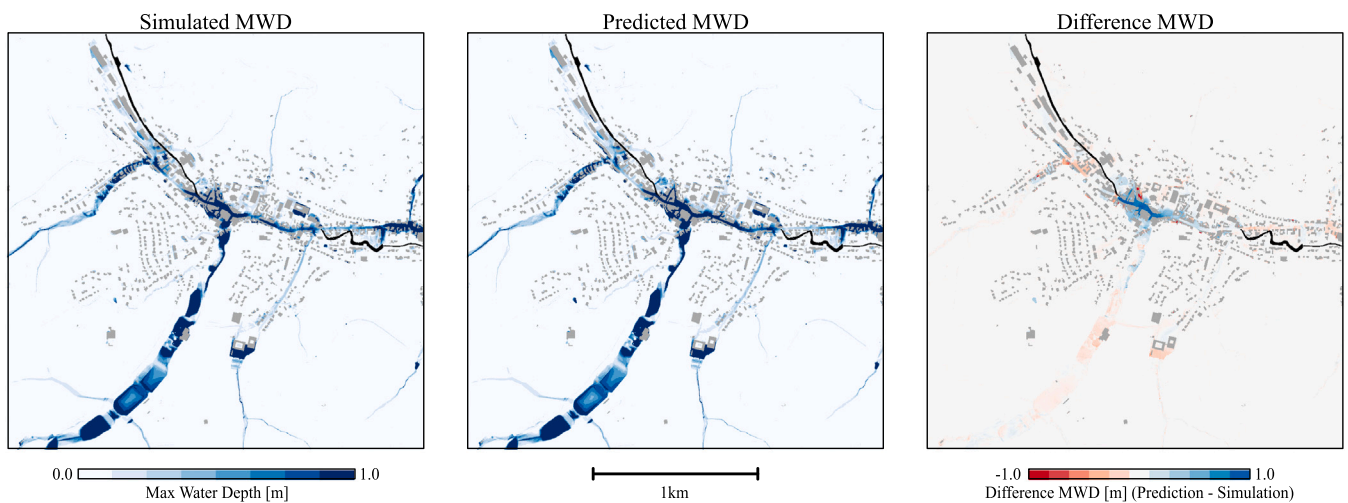
(a) Domain T_1 - RLP(b) Domain T_2 - RLP(c) Simulated vs. Predicted Max Water Depth (MWD) for a Patch from Domain T_1 

Fig. 2. Results for predictions using the model M_{single} for the corresponding single rain event in the test domains of RLP. (a) Scatter plot (density 1 to 10^5 pixels) of test domain T_1 with terrain overlayed with maximum water depth as an inset (top), and corresponding error (prediction minus simulation) box plots binned by maximum water depth with a bin width of 0.5 m (bottom). Dry pixels (water depth below 0.1 m) are omitted. Box plots are restricted to the 1st–99th percentiles to exclude outliers and improve visualization clarity. (b) Scatter plot (density 1 to 10^5 pixels) of test domain T_2 with terrain overlayed with maximum water depth as an inset (top), and corresponding error box plots (prediction minus simulation) binned by maximum water depth with a bin width of 0.5 m (bottom). Dry pixels (water depth below 0.1 m) are omitted. Box plots are restricted to the 1st–99th percentiles to exclude outliers and improve visualization clarity. (c) Detailed plots of the target, prediction, and their difference in the highlighted region of test domain T_1 .

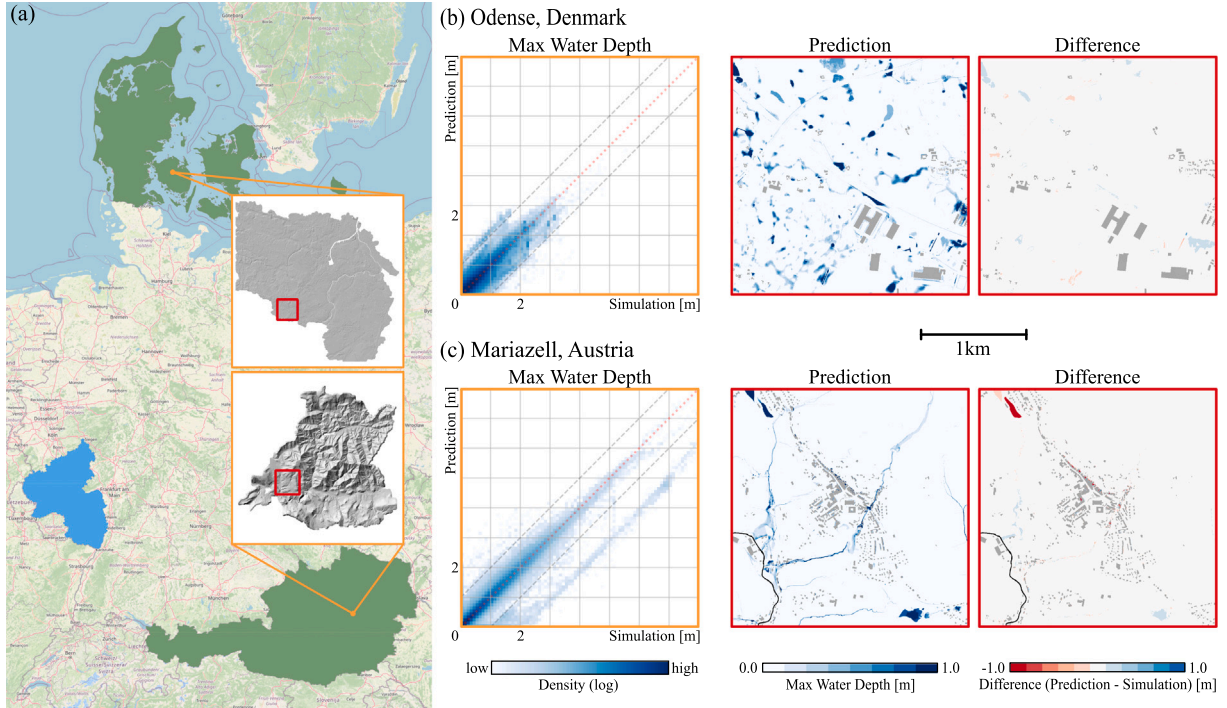


Fig. 3. Evaluation of the model M_{single} on unseen test domains using the corresponding single rain event. (a) Map of Europe showing countries used for training (blue) and test regions (green) in Denmark (top) and Austria (bottom). The insets show the terrains of the simulation domains, and the highlighted regions are analyzed in more detail in the following panels. (b) Scatter plots (density 1 to 10^5 pixels) of simulated vs. predicted maximum water depth for all pixels in the simulation domain of Odense, Denmark (left), and for the highlighted patch (right). (c) Scatter plots (density 1 to 10^5 pixels) of simulated vs. predicted maximum water depth for all pixels in the simulation domain of Mariazell, Austria (left), and for the highlighted patch (right). (For interpretation of the references to colour in this Figure legend, the reader is referred to the web version of this article.)

The CSI values in Fig. 5(b) illustrate the model's performance across different rainfall events. A comparison of M_{single} and M_{multi} for the initial 2-hour 100-year event shows only a minor reduction in CSI, indicating that the fine-tuning stage successfully extends the model to additional rainfall events while largely preserving its original performance (see also Table S3 and Figure S6 in Supplement Section A.5). The lowest CSI values are observed for the 1-year and 2-year KOSTRA events, as well as for the natural events Natural 4 and Natural 5, which differ substantially from the initial event due to their considerably lower rainfall intensities. In contrast, the remaining KOSTRA events and the natural rainfall events Natural 1, Natural 2, and Natural 3 achieve performance levels comparable to those of the initial training event, demonstrating the model's ability to adapt to a range of additional rainfall patterns.

Fig. 5(c) demonstrates the model's ability to accurately detect flooded areas (defined as regions with maximal water depth greater than 0.1 m) across different rainfall intensities. Higher rainfall intensities lead to higher maximal water depths and larger flooded areas. As rainfall intensity increases, the model consistently captures the expanding spatial extent of inundation.

3.5. Ablation

We now examine a series of ablation experiments to investigate the contribution of different factors to the model's capability for spatial generalization in greater detail. These experiments are designed to systematically quantify the impact of our architectural and training choices on overall model performance. By isolating individual components, we aim to better understand which design decisions are most critical for achieving robust generalization to unseen spatial domains.

3.5.1. Experiment setup

To focus on spatial generalizability, we use M_{single} as the baseline along with the corresponding rain event. To reduce computational costs,

all models were trained for 100 epochs, compared to the 400 epochs used in the previous experiments. The results presented in Fig. 6 were obtained from the 14 test domains in RLP Test.

We perform the following experiments:

- **Number of Training Domains.** Providing a sufficient quantity of training data is a central factor in achieving robust spatial generalization. To assess the importance of training data volume, we trained multiple models using varying numbers of domains. The domains were randomly selected such that the sets for 1 ($\sim 200 \text{ km}^2$), 8 ($\sim 1,400 \text{ km}^2$), 16 ($\sim 2,800 \text{ km}^2$), 32 ($\sim 5,100 \text{ km}^2$), 64 ($\sim 10,300 \text{ km}^2$), and 120 ($\sim 17,600 \text{ km}^2$) training domains formed nested subsets of one another. This ensures a controlled and fair comparison, allowing us to isolate the effect of increasing data quantity without introducing bias from different domain selections. The number of samples per epoch was kept constant, independent of the number of domains, to ensure comparability by maintaining equal computational effort across experiments.
- **Input Maps.** The input maps were divided into four categories based on the level of spatial context they provide. The terrain-only category includes solely elevation data. The semilocal category extends this by incorporating terrain, sink depth, and slope. The FLACC category additionally includes the flow accumulation map, which has been used in previous studies (Löwe et al., 2021; Seleen et al., 2023). The global category further augments the input with the upstream capacity and average upstream slope.
- **Model Size.** To evaluate the effect of model capacity, we conducted multiple training runs with varying numbers of parameters. Model size was controlled by adjusting the number of filters in the first convolutional layer. As described in Section 2.3, the baseline configuration uses 64 filters in the first layer. We additionally evaluated configurations with 16 and 32 initial feature maps to analyze how model capacity influences performance as well as a fully

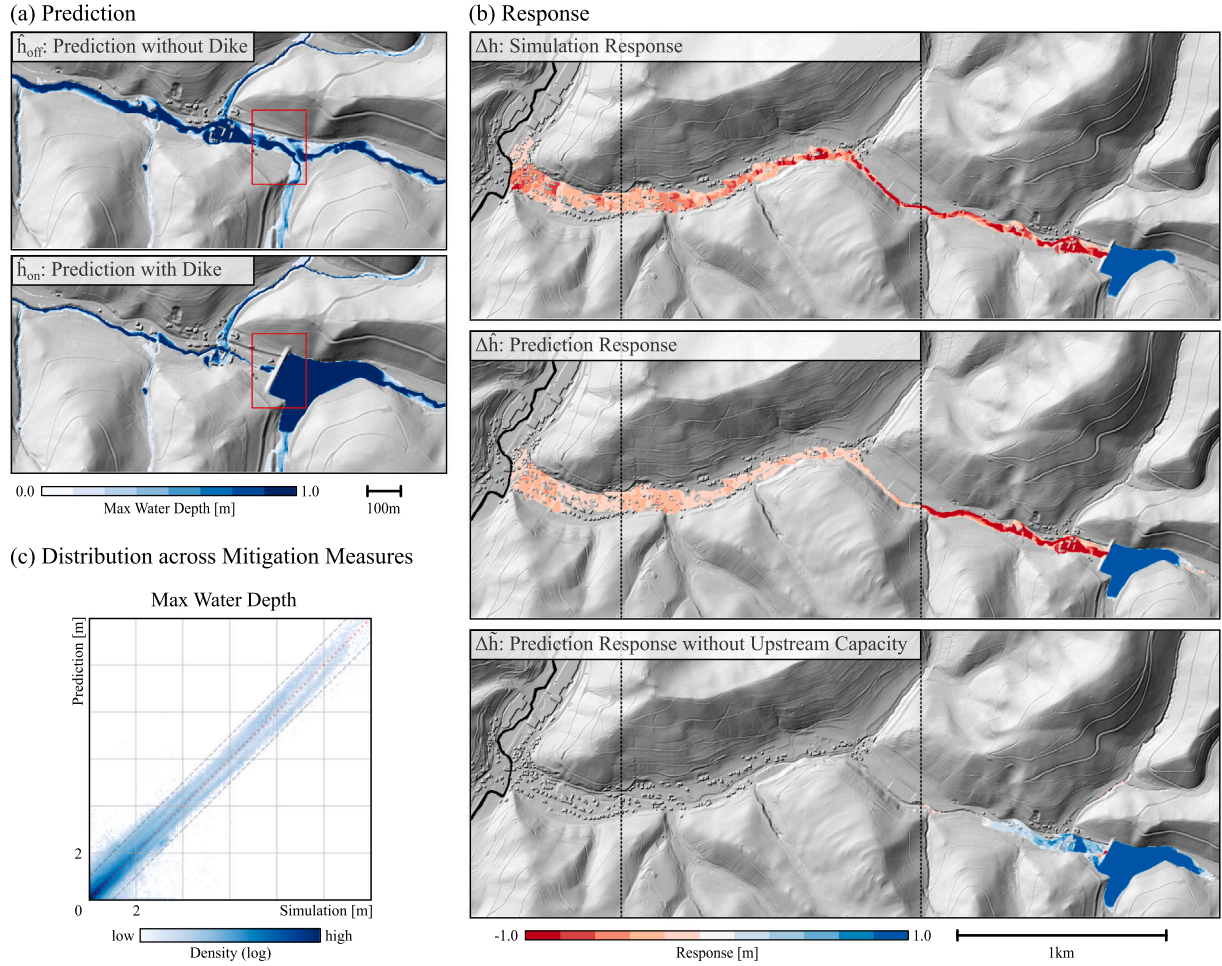


Fig. 4. Evaluation metrics for predictions using M_{single} under the corresponding single rain event in test domains with terrain modifications implemented in RLP. (a) Terrain with maximal water depth without a dike ($\hat{h}_{\text{off}}$) and with a dike ($\hat{h}_{\text{on}}$). The region of the modified terrain is highlighted (in red). (b) Simulation response ($\Delta h = h_{\text{on}} - h_{\text{off}}$), prediction response ($\Delta \hat{h} = \hat{h}_{\text{on}} - \hat{h}_{\text{off}}$), and the prediction response without considering the dike in the upstream capacity ($\Delta \hat{h} = \hat{h}_{\text{on}} - \hat{h}_{\text{off}}$, where $\hat{h}_{\text{on}}$ denotes the prediction obtained using the upstream capacity without the dike). Dotted lines indicate the boundaries between inference output patches. (c) Scatter plot (density 1 to 10^5 pixels) of simulation (x-axis) versus prediction (y-axis) for the scenarios including the mitigation measures. All experiments were performed in the unseen test domains in RLP. (For interpretation of the references to colour in this Figure legend, the reader is referred to the web version of this article.)

Table 6

Results for predictions using M_{single} and the corresponding single rain event in the RLP test domains containing artificial terrain modifications in the form of retention basins, dikes, and breaches. The table reports the error of the predicted response RMSE_{Δ} , defined as the difference between the maximum water depth with (h_{on}) and without (h_{off}) the measure ($\Delta h = h_{\text{on}} - h_{\text{off}}$) as well as the error metrics for simulations and predictions with the measures present and for the corresponding reference state without the measures. All mean-based metrics were computed only over grid cells where either the simulated or predicted response exceeded 0.05 m, thereby excluding irrelevant pixels with negligible response from the evaluation.

Structure	$\text{RMSE}_{\Delta} \downarrow$	Scenario	$\text{RMSE} \downarrow$	$\text{RMSE}_{0.1} \downarrow$	$\text{CSI}_{0.05} \uparrow$	$\text{CSI}_{0.1} \uparrow$	$\text{CSI}_{0.3} \uparrow$	$\text{CSI}_{1.0} \uparrow$
Basins	0.38	Measures On	0.31	0.35	0.88	0.85	0.78	0.82
		Measures Off	0.20	0.21	0.93	0.89	0.83	0.78
Breaches	0.33	Measures On	0.22	0.23	0.92	0.89	0.80	0.65
		Measures Off	0.47	0.52	0.89	0.86	0.79	0.67
Dikes	0.63	Measures On	0.39	0.45	0.88	0.86	0.85	0.81
		Measures Off	0.26	0.29	0.94	0.91	0.88	0.86
Total	0.39	Measures On	0.30	0.33	0.90	0.87	0.81	0.77
		Measures Off	0.34	0.37	0.92	0.88	0.82	0.76

convolutional U-Net with 64 initial filters, no attention mechanism in the bottleneck, and no attention gates.

3.5.2. Results

As shown in Fig. 6(a) (more detailed results and loss curves are provided in Figure S7 in Supplement Section A.6), increasing the number

of training domains has the most significant impact on performance. For a threshold of 0.1 m, the CSI improves from 0.6 to over 0.8 as the amount of training data increases. For thresholds of 0.3 m, 0.5 m, and 1.0 m, training on a single domain yields CSI values below 0.5, indicating unreliable performance, whereas expanding the training dataset substantially improves performance, increasing CSI values to 0.75 or

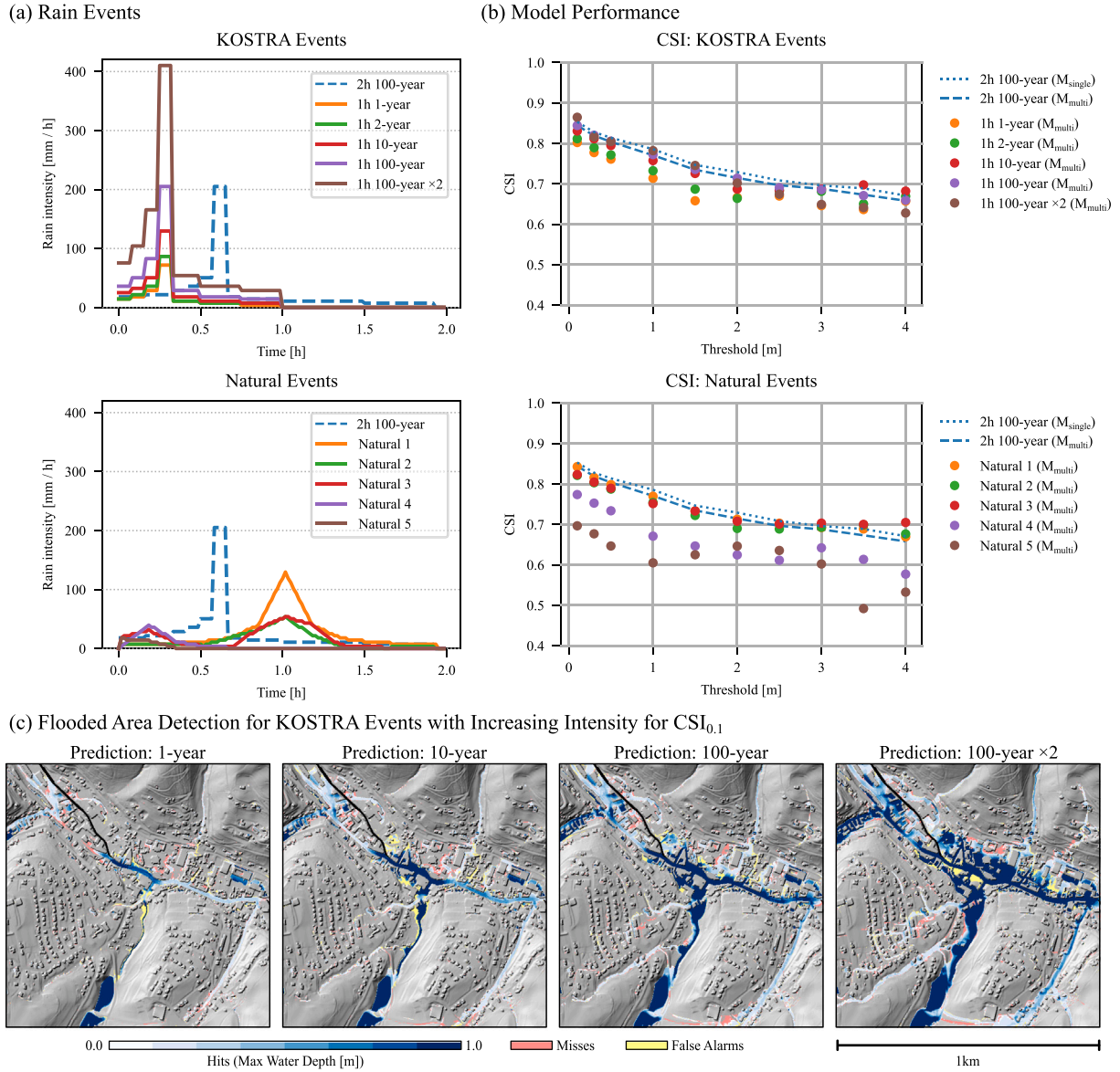


Fig. 5. Evaluation of M_{multi} for various rain events. (a) Hyetographs of the rainfall events used during retraining. KOSTRA events (top) and natural events derived from Löwe et al. (2021) (bottom). (b) CSI values of inference results for KOSTRA events (top) and natural events from Löwe et al. (2021) (bottom). (c) Pixel-wise hits, misses and false alarms for threshold 0.1 m for KOSTRA events with return periods of 1, 10, and 100 years and 100 years with an amplification factor $\times 2$ (left to right). All experiments were performed in the unseen test domains T_1 and T_2 in RLP.

higher. Performance improves rapidly up to 16 domains ($\sim 2800 \text{ km}^2$), after which gains become more gradual. This suggests that capturing the diversity of spatial patterns across different terrains requires a sufficiently large and varied dataset. This effect is particularly pronounced for higher maximal water depths ($> 1 \text{ m}$), where training on fewer than 16 domains results in CSI values below 0.6, indicating that extreme water levels are not accurately captured. These findings further imply that prior work (Löwe et al., 2021; Cache et al., 2024; Guo et al., 2021), which relied on only one or two domains, likely faced substantial limitations in spatial generalization.

Fig. 6(b) illustrates the influence of different input map configurations on model performance across CSI thresholds. While, in principle, sink depth should correlate with maximum water depths, we find no significant improvement in predictive performance from including this information and, in some cases, even observe a decrease in performance (compare Terrain and Semilocal). This may be because the terrain already describes the location and shape of depressions, while the sink

depth alone does not indicate their fill state, making the information potentially ambiguous. Incorporating the FLACC consistently improves performance by providing information about the size of the contributing area. The upstream-aware maps, originally introduced to account for the effects of mitigation measures, yield an additional performance gain. For both FLACC and the upstream-aware maps, the largest performance gains occur at higher CSI thresholds, consistent with the expectation that higher maximum water depths are driven by runoff from larger upstream areas.

Fig. 6(c) shows the effect of changing the model complexity by increasing the number of filters in the first layer or by incorporating attention mechanisms into the U-Net architecture. The results indicate that increasing the number of initial filters improves model performance, suggesting that the additional feature capacity allows the network to better capture relevant spatial patterns. However, the results indicate diminishing returns as model capacity increases. Compared with a fully convolutional U-Net (i.e., without attention in the bottleneck or attention gates)

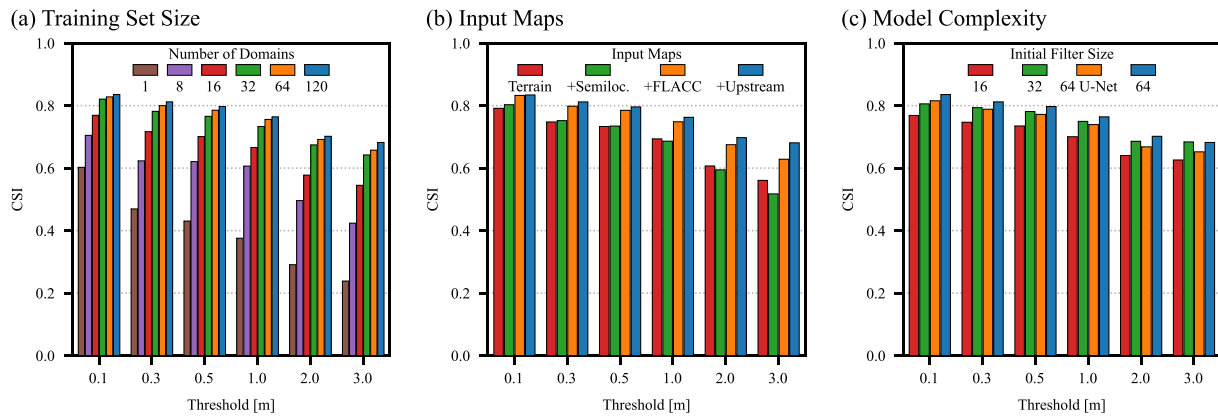


Fig. 6. Ablation studies. (a) CSI values for different training dataset sizes. (b) CSI values for different input maps: terrain only, with semi-local maps (terrain, slope, sink depth), with FLACC (terrain, slope, sink depth, FLACC) and with upstream-aware maps (terrain, slope, sink depth, FLACC, average upstream slope, upstream capacity). (c) CSI values with increasing model complexity. All experiments were performed in the unseen test domains on RLP.

using the same number of initial filters, the 64-filter attention-enhanced model consistently achieves higher CSI values. In addition, the 32-filter model achieves competitive performance despite having approximately one-third of the parameters of the standard model (46 million compared to 133 million; see Table S1 in Supplement Section A.2), and even surpasses the U-Net baseline for higher maximum water depths.

3.6. Timings

The computational performance of the proposed approach is assessed by comparing simulation, preprocessing, and inference times. The reported timings were taken from the two test domains T_1 and T_2 in RLP, each having approximately 40 million valid cells (corresponding to 160 km²).

The numerical simulation requires approximately 10–15 min, depending on the size of the domain and the number of grid cells. In contrast, model inference is significantly faster, taking around 3–20 s depending on oversampling and model size (see Table S2 in Supplement Section A.2). This includes dividing the input maps into patches, performing predictions, and stitching the outputs back together to obtain maximal water depth estimates for the entire domain.

Preprocessing – which involves the computation of terrain-derived features such as sink depth, slope, FLACC, upstream capacity, and average upstream slope as well as transformations for each input map – overall takes approximately 40 s when performed using WhiteboxTools. This step can be further accelerated through optimization strategies, for example by propagating all global maps simultaneously rather than computing them sequentially. Moreover, preprocessing only needs to be performed when the terrain changes; multiple rainfall events over the same domain can reuse the precomputed input features without additional cost.

3.7. Limitations & outlook

Despite the successes in the domain of spatial generalizability, several limitations remain.

Terrain classification. We showed that even though we achieve accurate results for all tested terrains, performance varies depending on the terrain characteristics. The results from RLP, Denmark, and Austria suggest that the model is more accurate in mountainous regions. Further investigation of the characteristics associated with this variation could inform the selection of more suitable input maps. For example, upstream-aware maps derived from coarser-resolution DEMs may reduce the influence of micro-topography in relatively flat domains.

Similarly, DEM-derived features used for flood delineation (Samela et al., 2017) may help identify candidate input maps that capture informative patterns across different slope scales. Beyond the design of more informative input features, classifying terrain characteristics could support a data-selection pipeline that maximizes terrain diversity, potentially improving generalization while reducing the amount of training data required.

Representation of hydrological and hydraulic processes at catchment scale. Another limitation concerns the incorporation of additional properties commonly used in physical flood simulations, such as spatially heterogeneous infiltration and surface roughness. These properties were omitted to reduce experimental complexity and to focus on the spatial generalization of terrain features. After achieving sufficient generalizability with respect to terrain, these properties could be reintroduced. Introducing upstream-aware representations of infiltration and roughness may help capture their long-range effects and improve accuracy when incorporating these properties into the model. Additionally, the catchment size was artificially restricted to 20 km² in the simulations. Consequently, the resulting model may not generalize well to larger catchments, as doing so would require predicting for FLACC values outside the range encountered during training, thereby relying on extrapolation. Enabling realistic estimations for larger catchments would further require accounting for spatially distributed rainfall. Combining our methods for spatial generalization with previous work on runoff prediction under spatially distributed rainfall (Cao et al., 2025; Hofmann and Schüttrumpf, 2021) could further extend the model's applicability to larger and more heterogeneous domains.

Mitigation measures. The application of machine learning models to predict the effects of flood mitigation measures has received some attention in prior work (Guo and Leitão, 2025), but remains relatively underexplored. Our experiments demonstrate the potential of using this model to obtain a rapid, rough first estimate of the effects of potential mitigation measures. Such estimates can support the initial screening and comparison of mitigation options, while the detailed design and assessment of selected measures should subsequently be carried out using physics-based simulations. Future research could build upon this approach to enable rapid machine learning-based assessments as an initial step in flood mitigation planning.

Preprocessing overhead. The preprocessing required to compute Sink Depth, FLACC, and other input maps must be repeated whenever the terrain is modified. Because these maps are globally defined, the maps have to be recalculated for the entire simulation domain, even when

the terrain changes are confined to a small, localized region, leading to significant overhead. Further optimizations, such as vectorization of the upstream propagation maps, can lead to significant reduction of this overhead.

Velocity prediction. Since velocity fields are relevant for assessing impacts such as structural damage, extending our model to predict maximal flow velocity is a natural next step. This approach has been explored in previous work (e.g., Guo et al. (2021)) and would allow a more comprehensive assessment of flood hazards.

Spatio-temporal setting. Previous studies have investigated data-driven spatio-temporal flood prediction (e.g., Bentivoglio et al. (2023); Kazadi et al. (2024); Chaudhary et al. (2024); Bentivoglio et al. (2025); Xu et al. (2025); Cao et al. (2025); Song et al. (2025b)), demonstrating the potential of these approaches for modeling flood evolution. For larger domains divided into patches, high velocities and large time steps can cause information to propagate across patch boundaries. Our proposed modifications for capturing long-range dependencies, may improve performance while enabling larger time steps and reducing computational cost.

Explainable AI. Given the high stakes associated with flood prediction, it is crucial to better understand the capabilities and limitations of data-driven models. We found that despite the model's impressive general performance, out-of-distribution scenarios still occur, leading to larger local errors (see Supplement Section A.4 for examples). Future research should therefore further explore explainable AI (Barredo Arrieta et al., 2020) techniques and uncertainty quantification (as exemplified by Chaudhary et al. (2022)) to provide more transparent and reliable predictions. Such approaches could help stakeholders better assess the confidence of model outputs and support safer, more informed decision-making.

4. Conclusion

In this work, we demonstrated that a deep learning model for flood prediction can achieve strong spatial generalization when trained with informative input maps and a sufficiently large dataset. Our model produces accurate predictions across previously unseen domains and diverse topographies. It achieves an RMSE of around 0.05 m and CSI scores between 0.78 and 0.91 at a threshold of 0.3 m. The model successfully reproduces both flood extents and water depths in lowland and mountainous regions. This indicates that the learned representations capture general hydrological patterns rather than overfitting to specific terrains. We further showed that the model is capable of estimating the response to localized terrain modifications, such as adding dikes, and that its predictions propagate downstream when relevant flow-related inputs are provided, illustrating the potential of spatial generalization to support scenario analysis for structural mitigation measures. By retraining the model on additional rainfall events, we demonstrated rapid adaptation to diverse precipitation scenarios while maintaining performance on previously seen events, further highlighting the robustness of the learned spatial representations. Ablation studies confirmed that a diverse set of training domains, carefully designed input maps and adequate model capacity are crucial for capturing complex spatial patterns.

Overall, these results highlight the effectiveness of combining spatial generalization with a flexible architecture and targeted retraining for rain events, enabling accurate and robust flood predictions across varied terrains, rainfall conditions, and structural modifications.

5. Data sources

The terrain data for RLP was provided by LVerGeoRP (2022). A terrain for Austria with 1 m resolution was provided by BEV. A terrain of Denmark with a spatial resolution of 0.5 meters was taken from open public data GeoDanmark and Danish Climate Data Agency (2023).

The Non-KOSTRA rain events were derived from the supplementary data provided by Löwe et al. (2021).

CRedit authorship contribution statement

Thomas Lutonsky: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Viktor Birschitzky:** Writing – review & editing, Supervision, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Stefan Jeschke:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Andreas Buttinger-Kreuzhuber:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition. **Nils Wandel:** Writing – review & editing, Visualization, Methodology. **Daniel Cornel:** Writing – review & editing, Funding acquisition. **Viola Haase:** Writing – review & editing. **Reinhard Klein:** Writing – review & editing. **Jürgen Waser:** Writing – review & editing, Project administration, Funding acquisition. **Michael Wimmer:** Writing – review & editing, Supervision.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT and Claude for reformulating the text. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The VRVis GmbH is funded by BMIMI, BMWET, Tyrol, Vorarlberg and Vienna Business Agency within the scope of COMET – Competence Centers for Excellent Technologies (911654) and COMET Module (904918), which are managed by FFG. The authors thank the Landesamt für Umwelt Rheinland-Pfalz for providing us with data necessary for setting up the Rheinland-Pfalz study region.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.jhydrol.2026.136500.

Data availability

The authors are unable to, or have chosen not to specify which data have been used.

References

- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Bentivoglio, R., Isufi, E., Jonkman, S.N., Taormina, R., 2023. Rapid spatio-temporal flood modelling via hydraulics-based graph neural networks. *Hydrol. Earth Syst. Sci.* 27, 4227–4246. <https://doi.org/10.5194/hess-27-4227-2023>
- Bentivoglio, R., Isufi, E., Jonkman, S.N., Taormina, R., 2025. Multi-scale hydraulic graph neural networks for flood modelling. *Nat. Hazards Earth Syst. Sci.* 25, 335–351. <https://doi.org/10.5194/nhess-25-335-2025>
- Berkhahn, S., Neuweiler, I., 2024. Data driven real-time prediction of urban floods with spatial and temporal distribution. *J. Hydrol. X* 22, 100167. <https://doi.org/10.1016/j.hydroa.2023.100167>
- Burrichter, B., Hofmann, J., Koltermann Da Silva, J., Niemann, A., Quirnbach, M., 2023. A spatiotemporal deep learning approach for urban pluvial flood forecasting with multi-source data. *Water* 15, 1760. <https://doi.org/10.3390/w15091760>
- Buttinger-Kreuzhuber, A., Horváth, Z., Noelle, S., Blöschl, G., Waser, J., 2019. A fast second-order shallow water scheme on two-dimensional structured grids over abrupt topography. *Adv. Water Resour.* 127, 89–108. <https://doi.org/10.1016/j.advwatres.2019.03.010>

- Buttinger-Kreuzhuber, A., Konev, A., Horváth, Z., Cornel, D., Schwerdtf, I., Blöschl, G., Waser, J., 2022. An integrated GPU-accelerated modeling framework for high-resolution simulations of rural and urban flash floods. *Environ. Model. Softw.* 156, 105480. <https://doi.org/10.1016/j.envsoft.2022.105480>
- Cache, T., Gomez, M.S., Beucler, T., Blagojevic, J., Leitao, J.P., Peleg, N., 2024. Enhancing generalizability of data-driven urban flood models by incorporating contextual information. *Hydrol. Earth Syst. Sci.* 28, 5443–5458. <https://doi.org/10.5194/hess-28-5443-2024>
- Cao, X., Wang, B., Yao, Y., Zhang, L., Xing, Y., Mao, J., Zhang, R., Fu, G., Borthwick, A.G., Qin, H., 2025. U-RNN high-resolution spatiotemporal nowcasting of urban flooding. *J. Hydrol.* 659, 133117. <https://doi.org/10.1016/j.jhydrol.2025.133117>
- Cea, L., Costabile, P., 2022. Flood risk in urban areas: modelling, management and adaptation to climate change. A review. *Hydrology* 9, 50. <https://doi.org/10.3390/hydrology9030050>
- Chaudhary, P., Leitão, J.P., Donauer, T., D'aronco, S., Perraudin, N., Obozinski, G., Perez-Cruz, F., Schindler, K., Wegner, J.D., Russo, S., 2022. Flood uncertainty estimation using deep ensembles. *Water* 14, 2980. <https://doi.org/10.3390/w14192980>
- Chaudhary, P., Leitão, J.P., Schindler, K., Wegner, J.D., 2024. Flood water depth prediction with convolutional temporal attention networks. *Water* 16, 1286. <https://doi.org/10.3390/w16091286>
- Do Lago, C.A., Giacomoni, M.H., Bentivoglio, R., Taormina, R., Gomes, M.N., Mendiondo, E.M., 2023. Generalizing rapid flood predictions to unseen urban catchments with conditional generative adversarial networks. *J. Hydrol.* 618, 129276. <https://doi.org/10.1016/j.jhydrol.2023.129276>
- El Baidi, M., Boushaba, F., Chourak, M., 2025. Drowning overconfidence with uncertainty: mitigating deep learning overconfidence in flood depth super-resolution through maximum entropy regularization. *Stoch. Environ. Res. Risk Assess.* 39, 1159–1177. <https://doi.org/10.1007/s00477-025-02917-1>
- GeoDanmark and Danish Climate Data Agency, 2023. <https://www.geodanmark.dk> (accessed: 8 January 2023).
- Guo, Z., Leitão, J.P., 2025. Designing pluvial flood-safe urban landscapes using differentiable surrogate flood models. *Blue-Green Syst.* 7, 124–138. <https://doi.org/10.2166/bgs.2025.042>
- Guo, Z., Leitão, J.P., Simões, N.E., Moosavi, V., 2021. Data-driven flood emulation: speeding up urban flood predictions by deep convolutional neural networks. *J. Flood Risk Manag.* 14, e12684. <https://doi.org/10.1111/jfr3.12684>
- Guo, Z., Moosavi, V., Leitão, J.P., 2022. Data-driven rapid flood prediction mapping with catchment generalizability. *J. Hydrol.* 609, 127726. <https://doi.org/10.1016/j.jhydrol.2022.127726>
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- Hofmann, J., Schütttrumpf, H., 2021. floodGAN: using deep adversarial learning to predict pluvial flooding in real time. *Water* 13, 2255. <https://doi.org/10.3390/w13162255>
- IPCC, 2022. Global Warming of 1.5°C: IPCC Special Report on Impacts of Global Warming of 1.5°C Above Pre-Industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty, 1 ed. Cambridge University Press, <https://doi.org/10.1017/9781009157940> ed).
- Junghänel, T., Ostermüller, J., Deutschländer, T., 2023. Grids of return periods of heavy precipitation (design precipitation) over Germany (KOSTRA-DWD) raster der wiederkehrintervalle für starkregen (bemessungsniederschläge) in deutschland (KOSTRA-DWD). <https://doi.org/10.5676/DWD/KOSTRA-DWD-2020>
- Kazadi, A., Doss-Gollin, J., Silva, A.L.D., 2024. Pluvial flood emulation with hydraulics-informed message passing. In: Proceedings of the 41st International Conference on Machine Learning. PMLR, pp. 23367–23390.
- Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. [arXiv:1412.6980](https://arxiv.org/abs/1412.6980)
- LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature* 521, 436–444. <https://doi.org/10.1038/nature14539>
- Liao, Y., Wang, Z., Chen, X., Lai, C., 2023. Fast simulation and prediction of urban pluvial floods using a deep convolutional neural network model. *J. Hydrol.* 624, 129945. <https://doi.org/10.1016/j.jhydrol.2023.129945>
- Lindsay, J., 2017. WhiteboxTools. <https://github.com/opengeos/whitebox-python/tree/v2.3.6> (accessed: 19 August 2025).
- Loshchilov, I., Hutter, F., 2016. SGDR: stochastic gradient descent with warm restarts. [arXiv:1608.03983](https://arxiv.org/abs/1608.03983)
- Löwe, R., Böhm, J., Jensen, D.G., Leandro, J., Rasmussen, S.H., 2021. U-FLOOD – topographic deep learning for predicting urban pluvial flood water depth. *J. Hydrol.* 603, 126898. <https://doi.org/10.1016/j.jhydrol.2021.126898>
- Luan, G., Hou, J., Wang, T., Li, D., Zhou, Q., Liu, L., Duan, C., 2023. A 1D-2D dynamic bidirectional coupling model for high-resolution simulation of urban water environments based on GPU acceleration techniques. *J. Clean. Prod.* 428, 139494. <https://doi.org/10.1016/j.jclepro.2023.139494>
- Luo, W., Li, Y., Urtasun, R., Zemel, R., 2017. Understanding the effective receptive field in deep convolutional neural networks. [arXiv:1701.04128](https://arxiv.org/abs/1701.04128)
- LVermGeoRP, 2022. <https://lvermgeo.rlp.de/geodaten-geoshop/open-data> (accessed: 25 February 2025).
- Morales-Brotons, D., Vogels, T., Hendriks, H., 2024. Exponential moving average of weights in deep learning: dynamics and benefits. *Trans. Mach. Learn. Res.* <https://doi.org/10.48550/ARXIV.2411.18704>
- Natarajan, D.K., Stricker, M., Mukherjee, R., Charfuelan, M., Nuske, M., Dengel, A., 2026. Deep learning emulators for large-scale, high-resolution urban pluvial flood prediction. In: Braun, T., Paaßen, B., Stolzenburg, F. (Eds.), KI 2025: Advances in Artificial Intelligence, vol. 15956. Springer Nature Switzerland, Cham, pp. 228–235. https://doi.org/10.1007/978-3-032-02813-6_17
- Nicklin, H., Leicher, A.M., Dieperink, C., Van Leeuwen, K., 2019. Understanding the costs of inaction—an assessment of pluvial flood damages in two European cities. *Water* 11, 801. <https://doi.org/10.3390/w11040801>
- Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-Net: learning where to look for the pancreas. [arXiv:1804.03999](https://arxiv.org/abs/1804.03999)
- Pang, Y., He, J., Canlas, A., Ju, L., Fei, K., Zhang, L., 2026. A skillful self-evolving deep-learning framework for pluvial flood process forecasting in urban areas. *J. Hydrol.* 664, 134593. <https://doi.org/10.1016/j.jhydrol.2025.134593>
- Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A., 2018. FiLM: visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32. <https://doi.org/10.1007/978-3-0321-11671>
- Petit, O., Thome, N., Rambour, C., Themyr, L., Collins, T., Soler, L., 2021. U-Net transformer: self and cross attention for medical image segmentation. In: Lian, C., Cao, X., Reikik, L., Xu, X., Yan, P. (Eds.), Machine Learning in Medical Imaging, vol. 12966. Springer International Publishing, Cham, pp. 267–276. https://doi.org/10.1007/978-3-030-87589-3_28
- Samela, C., Troy, T.J., Manfreda, S., 2017. Geomorphic classifiers for flood-prone areas delineation for data-scarce environments. *Adv. Water Resour.* 102, 13–28. <https://doi.org/10.1016/j.advwatres.2017.01.007>
- Selem, O., Ayzel, G., Bronstert, A., Heistermann, M., 2023. Transferability of data-driven models to predict urban pluvial flood water depth in Berlin, Germany. *Nat. Hazards Earth Syst. Sci.* 23, 809–822. <https://doi.org/10.5194/nhess-23-809-2023>
- Shao, Y., Chen, J., Zhang, T., Yu, T., Chu, S., 2024. Advancing rapid urban flood prediction: a spatiotemporal deep learning approach with uneven rainfall and attention mechanism. *J. Hydroinformatics* 26, 1409–1424. <https://doi.org/10.2166/hydro.2024.024>
- Shaw, P., Uszkoreit, J., Vaswani, A., 2018. Self-attention with relative position representations. [arXiv:1803.02155](https://arxiv.org/abs/1803.02155)
- Song, W., Guan, M., Guo, K., Yu, D., 2025a. Rapid flood inundation mapping by integrating deep learning-based image super-resolution with coarse-grid hydrodynamic modeling. *Eng. Appl. Comput. Fluid Mech.* 19, 2481115. <https://doi.org/10.1080/19942060.2025.2481115>
- Song, W., Guan, M., Yu, D., 2025b. SwinFlood: a hybrid CNN-swin transformer model for rapid spatiotemporal flood simulation. *J. Hydrol.* 660, 133280. <https://doi.org/10.1016/j.jhydrol.2025.133280>
- Taghizadeh, M., Zandsalimi, Z., Nabian, M.A., Shafee-Jood, M., Alemazkoor, N., 2025. Interpretable physics-informed graph neural networks for flood forecasting. *Comput.-Aided Civ. Infrastruct. Eng.* 40, 2629–2649. <https://doi.org/10.1111/mice.13484>
- Waser, J., Ribicic, H., Fuchs, R., Hirsch, C., Schindler, B., Blöschl, G., Grollier, M.E., 2011. Nodes on ropes: a comprehensive data and control flow for steering ensemble simulations. *IEEE Trans. Vis. Comput. Graph.* 17, 1872–1881. <https://doi.org/10.1109/TVCG.2011.225>
- Winsemius, H.C., Aerts, J.C.J.H., Van Beek, L.P.H., Bierkens, M.F.P., Bouwman, A., Jongman, B., Kwadijk, J.C.J., Ligtoet, W., Lucas, P.L., van Vuuren, D.P., Ward, P.J., 2016. Global drivers of future river flood risk. *Nat. Clim. Change* 6, 381–385. <https://doi.org/10.1038/nclimate2893>
- Woo, H., Choi, H., Kim, M., Noh, S.J., 2025. Physics-guided deep learning for spatiotemporal evolution of urban pluvial flooding. *Water* 17, 1239. <https://doi.org/10.3390/w17081239>
- Xu, Q., De Vos, L.F., Shi, Y., Rüther, N., Bronstert, A., Zhu, X.X., 2025. Urban flood modeling and forecasting with deep neural operator and transfer learning. <https://doi.org/10.2139/ssrn.5197692>