https://doi.org/10.5194/hess-2024-176
Preprint. Discussion started: 5 August 2024

Hydrology and Earth System Sciences
Open Access
Discussions

A deep learning model for real-time forecasting of 2-D river flood inundation maps

Matteo Pianforini[1], Susanna Dazzi[1], Andrea Pilzer[2], and Renato Vacondio[1]

[1]Department of Engineering and Architecture, University of Parma, Parco Area delle Scienze 181/A, 43124 Parma, Italy.
[2]NVIDIA AI Technology Center, Italy.

Correspondence: Matteo Pianforini (matteo.pianforini@unipr.it), Susanna Dazzi (susanna.dazzi@unipr.it), Andrea Pilzer (apilzer@nvidia.com), and Renato Vacondio (renato.vacondio@unipr.it)

Abstract. Floods are among the most hazardous natural disasters worldwide. Accurate and rapid flood predictions are critical for effective early warning systems and flood management strategies. The high computational cost of hydrodynamic models often limits their application in real-time flood simulations. Conversely, data-driven models are gaining attention due to their high computational efficiency. In this study, we aim at assessing the effectiveness of transformer-based models for forecasting the spatiotemporal evolution of fluvial floods in real-time. To this end, the transformer-based data-driven model FloodSformer (FS) has been adapted to predict river flood inundations with negligible computational time. The FS model leverages an autoencoder framework to analyze and reduce the dimensionality of spatial information in input water depth maps, while a transformer architecture captures spatiotemporal correlations between inundation maps and inflow discharges using a cross-attention mechanism. The trained model can predict long-lasting events using an autoregressive procedure. The model's performance was evaluated in two case studies: an urban flash flood scenario at the laboratory scale and a river flood scenario along a segment of the Po River (Italy). Datasets were numerically generated using a two-dimensional hydrodynamic model. Special attention was given to analyzing how the accuracy of predictions is influenced by the type and severity of flood events used to create the training dataset. The results show that prediction errors generally align with the uncertainty observed in physically based models, and that larger and more diverse training datasets help improving the model's accuracy. Additionally, the computational time of the real-time forecasting procedure is negligible compared to the physical time of the simulated event. The performance of the FS model was also benchmarked against a state-of-the-art convolutional neural network architecture and showed better accuracy. These findings highlight the potential of transformer-based models in enhancing flood prediction accuracy and responsiveness, contributing to improve flood management and resilience.

1 Introduction

Floods are the most hazardous natural disasters worldwide (Wallemacq and House, 2018). The catastrophic repercussions of flooding events include loss of human lives, economic damage, environmental degradation, and profound social disruptions. In 2023, over 300 disasters related to floods and storms occurred globally, accounting for approximately 75% of all natural disasters (CRED, 2024). These catastrophes affected tens of millions of people, resulting in the loss of tens of thousands of lives and extensive economic damage. Consequently, understanding and accurately simulating floods have become imperative for safeguarding communities and enhancing inundation resilience. In addition to structural flood mitigation measures, the implementation of efficient emergency action plans, based on Early Warning Systems (EWS), can significantly mitigate the impact of extreme inundations (Pappenberger et al., 2015), reducing flood damage by up to 35% (Rogers and Tsirkunov, 2011).

The development of effective EWS requires rapid and accurate predictions of flood dynamics. Traditionally, this task has been approached using physically based models, which rely on the discretization of partial differential equations to describe the physical processes. For simulating river floods, hydrodynamic models that solve the two-dimensional (2D) Shallow Water Equations (SWE) numerically using Finite Difference, Finite Element or Finite Volume schemes are frequently employed. These models provide accurate results for simulating flood propagation in regions with complex topographies and flood dynamics. However, their application for real-time forecasting is often hindered by the typically high computational times (Bomers and Hulscher, 2023), especially when high spatial resolution is required. To address this issue, research has focused on reducing the computational cost of physically based models by leveraging the efficiency of modern Graphics Processing Units (GPUs) (e.g., Morales-Hernández et al., 2021; Vacondio et al., 2014; Xia et al., 2019). Despite these efforts, the computational cost of 2D hydrodynamic models remains significant, and access to High-Performance Computing (HPC) clusters is necessary to accelerate computations (Turchetto et al., 2020). Consequently, the use of 2D-SWE solvers for early warning is limited and typically relies on databases composed of pre-simulated scenarios for various severities and inundation characteristics (e.g., Dazzi et al., 2022).

In the last decade, “black-box” models, also known as “data-driven” or “surrogate” models, have gained significant attention in predicting hydrological variables. These algorithms learn the complex relationships between input and outputs variables from observed or simulated data, thereby neglecting the physics of the process involved. The high computational efficiency of surrogate models facilitates their use in forecasting flood scenarios, which is the focus of this study. A wide variety of data-driven models for river flood forecasting exist in the literature (Bentivoglio et al., 2022; Mosavi et al., 2018). Early models focused on forecasting the temporal variation of discharges and/or water stages in specific river sections based on hydrological variables observed in the upstream catchment (e.g., Campolo et al., 1999; Kratzert et al., 2018) or water levels observed at previous instants in upstream river sections (e.g., Dazzi et al., 2021b; Tayfur et al., 2018). Various machine-learning (ML) and deep-learning (DL) models have been employed for these prediction tasks, such as Nonlinear Autoregressive Exogenous (NARX) networks (e.g., Bomers, 2021), Long-Short-Term Memory (LSTM) networks (e.g., Kratzert et al., 2018; Nevo et al., 2022), Convolutional Neural Networks (CNNs; e.g., Wang et al., 2019), and transformer architectures (e.g., Yin et al., 2022). However, forecasting the temporal variation of hydraulic variables in a specific river section only partially describes the severity of floods. For effective EWS, understanding the spatial and temporal distribution of hydraulic variables (e.g., water depths and velocities) is crucial. Consequently, recent studies have increasingly focused on forecasting the temporal variation of inundation maps. For pluvial floods, some works (e.g., Burrichter et al., 2023; Hop et al., 2024; Kao et al., 2021) have focused on predicting inundation maps based on rainfall observations and, in some cases, terrain elevation. Differently, for river floods, researchers have developed data-driven models that use upstream hydrograph inflows to predict the spatiotemporal propagation of inundation maps (e.g., Kabir et al., 2020; Wei et al., 2024; Zhou et al., 2022). A first application of a CNN to predict inundation maps for a fluvial flood was proposed by Kabir et al. (2020). This surrogate model predicts a water depth map at a specific instant using a time series of inflow discharge values at previous time steps as input data. The primary drawback of this type of surrogate models is their inability to account for spatiotemporal correlations between consecutive inundation maps, relying solely on the values of the upstream boundary conditions for forecasting.

Innovative methods such as Graph Neural Networks (GNN; Bentivoglio et al., 2023), Gaussian Process (GP) models combined with dimensionality reduction schemes (Donnelly et al., 2022; Fraehr et al., 2023), and transformer architecture (Pianforini et al., 2024a) have also been explored for flood prediction. Bentivoglio et al. (2023) applied GNN to emulate SWE solvers for predicting flood dynamics on randomly generated unseen topographies in synthetic case studies. Despite achieving promising results, the relatively high computational time may restrict the application of GNN in real-time flood forecasting. Furthermore, their applicability to real-word cases remains to be investigated. Differently, Donnelly et al. (2022) combined a GP framework with a principal component analysis to emulate numerical model results and quantify prediction uncertainty. Similarly, Fraehr et al. (2023) employed a low-resolution hydrodynamic model to provide a preliminary estimation of flood inundation, which is then enhanced to high resolution using empirical orthogonal function analysis and GP models, aiming to replicate the results of a high-resolution hydrodynamic model. However, when combining a DL model with a numerical scheme, the prediction's efficiency may be hindered by the computational time and stability of the hydrodynamic model (Fraehr et al., 2023). Therefore, the adoption of a rapid and accurate numerical model is fundamental to speed up the prediction process.

Transformer architectures, initially proposed by Vaswani et al. (2017) for natural language processing tasks, have been applied to various hydrological applications due to their ability to analyze long-range dependencies and attend to different spatiotemporal information in input sequences. These models have been used in tasks such as rainfall-runoff modeling (Li et al., 2024; Xu et al., 2023; Yin et al., 2022, 2023), dam-break scenarios (Pianforini et al., 2024a), pluvial floods (Burrichter et al., 2024; Chaudhary et al., 2024; Jin et al., 2024), and streamflow or water level prediction in rivers (Castangia et al., 2023; Liu et al., 2022). Some studies have shown that transformer-based models generally outperform other DL models in prediction tasks (e.g., Li et al., 2024; Yin et al., 2022). However, this type of model has not been applied to fluvial flood maps yet. Furthermore, existing frameworks typically handle homogeneous data in terms of dimensionality (either all matrices or all vectors), whereas simulating river floods involves the challenge of correlating information from heterogeneous data sources (i.e., time series of upstream inflows and sequences of inundation maps).

In this work, we aim at assessing the effectiveness of transformer-based models for forecasting the spatiotemporal evolution of fluvial floods in real-time. To this end, we started from the FloodSformer (FS) model (Pianforini et al., 2024a), which was successfully developed and applied to predict inundation maps for dam-break scenarios. A limitation associated with this initial version of the FS model is its inability to incorporate upstream boundary conditions as input data, as such information was not used in dam-break studies. In contrast, the FS model has been substantially modified in this work to address fluvial floods, for which the inclusion of time-varying boundary conditions (i.e., inflow hydrographs) is fundamental. The enhanced architecture combines an autoencoder and a transformer-based framework to analyze spatiotemporal information from input inundation maps and the temporal correlation of upstream boundary data, predicting long sequences of future water depth maps using an autoregressive procedure. Unlike the original model, the adapted architecture employs the cross-attention (CA) mechanism (Vaswani et al., 2017) to capture dependencies between different input sequences, namely inundation maps and discharge values, allowing the model to effectively consider and correlate heterogeneous input data. This innovation enables the prediction of long-lasting river flood events with complex flow dynamics and topographies. We analyzed two case studies: an urban flash flood scenario at the laboratory scale and a river flood scenario along a stretch of the Po River (Italy). For each case study, we considered different training datasets (i.e., varying the type, number, and intensity of flood scenarios) to analyze the influence of these configurations on the surrogate model's capability to generalize to unseen flood maps. Furthermore, the performance of the FS model was benchmarked against the CNN architecture proposed by Kabir et al. (2020). The inundation maps used to train and evaluate the FS model were generated with a 2D SWE solver (i.e., PARFLOOD code; Vacondio et al., 2014). This approach allows for generating large datasets, including both real events and synthetic scenarios, thus addressing the issue of scarce observed inundation maps for real flood events.

The paper is structured as follows: Section 2 describes the FS model and the case studies. Section 3 and Section 4 present and discuss the results, respectively. Finally, conclusions are drawn in Section 5.

2 Methods and materials

2.1 FloodSformer model description

The FloodSformer framework introduced by Pianforini et al. (2024a) represents a data-driven model designed to forecast the temporal evolution of inundation maps, emulating the results of physically based schemes. This surrogate model operates under the assumption that inundation maps can be conceptualized as images with dimensions $H \times W$, wherein each pixel corresponds to a computational cell of a Cartesian grid. Consequently, the model draws inspiration from transformer-based architectures commonly employed for video frame prediction tasks (Ye and Bilodeau, 2023). The FS model is composed of three consecutive blocks (Figure 1a): a 2D CNN encoder, a video prediction transformer (VPTR) framework (Ye and Bilodeau, 2023), and a 2D CNN decoder. The first and last blocks constitute the ResNet-based autoencoder (AE) of the Pix2Pix model (Isola et al., 2017), and are used to analyse spatial information in the input maps and reduce their dimensionality, decreasing the memory and time consumption required for the training process. In the original implementation, the second block is the fully autoregressive VPTR model proposed by Ye and Bilodeau (2023), employed to learn the spatiotemporal relationships between consecutive maps thanks to the self-attention (SA) mechanism (Vaswani et al., 2017).

The FS model was initially developed to forecast inundations resulting from dam-break scenarios (Pianforini et al., 2024a). In that context, the description of the flood propagation exclusively relies on the initial conditions within the upstream reservoir. In contrast, when addressing river floods, the incorporation of time-varying open boundary conditions in the data-driven model is fundamental for ensuring the surrogate model's ability to learn the flood propagation. Consequently, in the current study, we modified the original FS model to deal with open boundary conditions. Specifically, the SA mechanism of the VPTR framework has been replaced with a cross-attention (CA) mechanism (see Section 2.1.1). This enhancement enables the handling of different data types (i.e., sequence of maps and time series of discharge values) to predict future maps.

Figure 1 illustrates the general workflow of the modified version of the FS model. Considering a sequence $I+1$ frames (water depth maps at consecutive instants), the 2D CNN encoder takes the first I ground-truth maps ($t = 1, \ldots, I$) and, for each of them, extracts the spatial information creating the latent features. These features are then passed as value ($\boldsymbol{V}$) and key ($\boldsymbol{K}$) matrices to the VPTR framework, together with the inflow discharges at instants $t = 2, \ldots, I+1$, passed as query ($\boldsymbol{Q}$) matrix. The VPTR combines this information using the cross-attention mechanism and predicts the latent features at the next instants ($t = 2, \ldots, I+1$). Finally, the predicted water depth maps are reconstructed by the 2D CNN decoder.

The structure of the FS model can be summarized with the following equations:

$$\boldsymbol{z}_t = Enc(\boldsymbol{x}_t), \qquad t \in [1, \ldots, I] \tag{1}$$

$$\hat{\boldsymbol{z}}_t = \mathcal{T}(\boldsymbol{K}, \boldsymbol{V} : [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{t-1}]; \boldsymbol{Q} : [q_2, \ldots, q_t]), \qquad t \in [2, \ldots, I+1] \tag{2}$$

$$\hat{\boldsymbol{x}}_t = Dec(\hat{\boldsymbol{x}}_t), \qquad t \in [2, \ldots, I+1] \tag{3}$$

where $\boldsymbol{x}_t \in \mathbb{R}^{H \times W}$ and $\hat{\boldsymbol{x}}_t \in \mathbb{R}^{H \times W}$ are the ground-truth and predicted maps at time t, respectively. $\boldsymbol{z}_t \in \mathbb{R}^{h \times w \times d_{\text{model}}}$ and $\hat{\boldsymbol{z}}_t \in \mathbb{R}^{h \times w \times d_{\text{model}}}$ are the latent features at time t in input and output from the VPTR block, respectively. $q_t \in \mathbb{R}^1$ is the inflow discharge at time t. $Enc(\ldots)$, $Dec(\ldots)$ and $\mathcal{T}(\ldots)$ are the encoder, decoder and VPTR blocks, respectively. $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ are the query, key, and value matrices of the cross-attention computation, respectively. H and W are the height and width (in pixels) of the water depth maps along the south-north and west-east directions, while $h = H/2^k$ and $w = W/2^k$ are the height and width of the latent features, respectively. k is the number of convolutional layers of the AE, and d_{model} is the number of channels of the latent features. The hyperparameter I represents the number of input frames, i.e., the maximum length of the sequence of input maps for the FS model. Consequently, a higher value for I allows the model to extract spatiotemporal information from a longer sequence of maps, enhancing its ability to predict the map at time $I+1$. However, increasing I results in longer computational times and greater memory requirements for the training process, thus an optimal value should be defined (see Section 2.1.4).

The training process aims to minimize the differences between the maps predicted by the surrogate model and the ground-truth maps obtained from a hydrodynamic model. More details about the training strategy are available in Section 2.1.2.

Once the surrogate model is trained, real-time forecasting of future frames is achieved through an autoregressive (AR) procedure, described in detail in Section 2.1.3. This technique utilizes a recursive method wherein input frames are substituted with predicted ones. As a result, beginning from P past frames, the AR procedure iteratively predicts F future frames. The total lead time of this prediction is constrained by the potential loss of accuracy stemming from error accumulation and the length of the inflow discharge forecast. Notably, the AR procedure relies on the availability of the entire time series of upstream discharges, which is provided as input data to the surrogate model. This is in line with the approaches commonly employed in EWS based on physically-based hydrodynamic models, in which the inflow time series derives from meteorological/hydrological models.

2.1.1 Cross-attention mechanism and VPTR module

Pianforini et al. (2024a) employed the self-attention (SA) mechanism (Vaswani et al., 2017) within the VPTR framework to capture the dependencies in a single embedding sequence representing latent features derived from the encoder block. As

https://doi.org/10.5194/hess-2024-176
Preprint. Discussion started: 5 August 2024

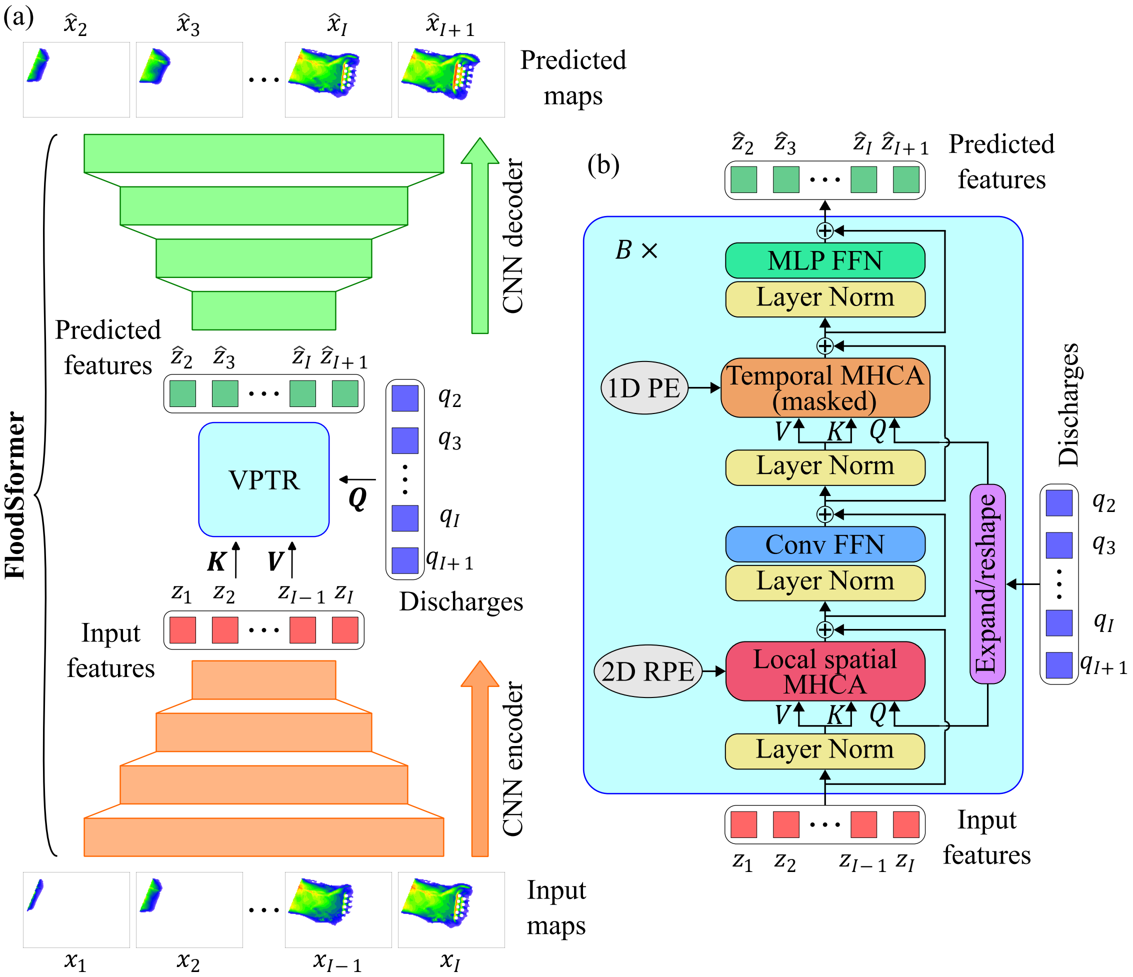

Figure 1. Overview of the FloodSformer model architecture. **(a)** Schematic depiction of the model's overall workflow. Input maps are transformed into key (***K***) and value (***V***) matrices for the VPTR block, while discharge values serve as the query (***Q***) matrix. **(b)** Detailed illustration of the layers within a VidHRFormer block. The video prediction Transformer (VPTR) module comprises a sequence of B consecutive VidHRFormer blocks.

already mentioned, accurate prediction of river floods necessitates integrating inflow discharge values into the data-driven model. In this study, we address this challenge by replacing the SA mechanism with the cross-attention (CA) one (Vaswani et al., 2017). Unlike SA, CA facilitates attention across distinct input sequences, namely latent features and discharge values.

Considering two different input matrices $\boldsymbol{X}_Q$ and $\boldsymbol{X}_{KV}$, the cross-attention process can be formulated as:

165
$$\boldsymbol{Q} = \boldsymbol{X}_Q \boldsymbol{W}_Q; \qquad \boldsymbol{K} = \boldsymbol{X}_{KV} \boldsymbol{W}_K; \qquad \boldsymbol{V} = \boldsymbol{X}_{KV} \boldsymbol{W}_V;$$
$$CA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{4}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ are query, key, and value matrices, obtained by a linear transformation of the input matrices through the trainable weight matrices $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, $\boldsymbol{W}_V$. d_k is the embedding dimension of the key matrix. To attend information from different representation subspaces simultaneously, the transformer adopts the multi-head cross-attention (MHCA) mechanism, where several CA computations, called "heads", are performed in parallel:

$$MHCA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(CA_1,\ldots,CA_p)\boldsymbol{W}_{MHCA} \tag{5}$$

where $\boldsymbol{W}_{MHCA}$ is a projection matrix to integrate the outputs of all attention processes, while p is the number of heads.

In the FS model, the input matrix $\boldsymbol{X}_{KV}$ represents the sequence of latent features $\boldsymbol{Z} = [\boldsymbol{z}_1,\ldots,\boldsymbol{z}_I] \in \mathbb{R}^{I\times h\times w\times d_{model}}$ provided by the CNN encoder layer, while the $\boldsymbol{X}_Q$ is obtained expanding and reshaping the sequence of discharge values $[q_2,\ldots,q_{I+1}] \in \mathbb{R}^I$ in order to obtain the same dimensions of $\boldsymbol{X}_{KV}$. Consequently, the VPTR block takes in input the embedded water depth maps $[\boldsymbol{z}_1,\ldots,\boldsymbol{z}_I]$, and the corresponding inflow discharge values $[q_2,\ldots,q_{I+1}]$, and predicts the embedded maps $[\hat{\boldsymbol{z}}_2,\ldots,\hat{\boldsymbol{z}}_{I+1}]$ (see Figure 1b). We emphasize that the temporal translation of one frame between maps and discharge values used as input data (e.g., the MHCA correlates the feature at time $t = 1$ with the discharge at instant $t = 2$) is essential to predict the map at the subsequent instant (e.g., $t = 2$).

The VPTR module is composed of B consecutive VidHRFormer blocks (Ye and Bilodeau, 2023), represented in Figure 1b. The key layers of the VidHRFormer block are the local spatial MHCA and the temporal MHCA, which apply the attention computation in space and time, respectively. With the aim of reducing the overall complexity compared to a standard joint space-time attention scheme, the use of two different layers for the spatial and temporal CA analysis has been adopted (Ye and Bilodeau, 2023). Masking is employed within the attention mechanism of the temporal MHSA layer to prevent the prediction at a specific time from being influenced by subsequent instants. A convolutional feed-forward neural network (Conv FFN), a multilayer perceptron (MLP) and normalization layers complete the VidHRFormer block. Furthermore, a 2D relative positional encoding (RPE) and a fixed absolute 1D positional encoding (PE) are used in the spatial and temporal MHSA, respectively.

2.1.2 Training methodology

Both the computational time and GPU memory demands necessary for model training have been minimized by dividing the training process in two distinct stages.

The first stage, denoted as "AE training", focuses on training the encoder and decoder blocks, concatenated to form a conventional AE. During this phase, the primary objective is to analyze spatial information within the input maps and undertake feature extraction neglecting temporal information and boundary conditions. The batch consists of randomly selected N individual frames $\boldsymbol{X} \in \mathbb{R}^{N\times H\times W}$. During the training process, the encoder extracts the latent feature $\boldsymbol{z}$ from a map $\boldsymbol{x}$ in the dataset. Then, this feature is used by the decoder to reconstruct the input map $\hat{\boldsymbol{x}}$. The AE training procedure aims at minimizing the following loss function (Ye and Bilodeau, 2023):

$$L_{AE} = L_{MSE} + \lambda_{GDL}L_{GDL} + \lambda_{GAN}\arg\min_G\max_D L_{GAN}(D,G) \tag{6}$$

where:

$$L_{MSE} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}^n - \hat{\boldsymbol{x}}^n)^2 \tag{7}$$

$$L_{GDL} = \frac{1}{N}\sum_{n=1}^{N}\left[\sum_{i=1}^{W}\sum_{j=1}^{H}\left(\left||\boldsymbol{x}_{i,j}^n - \boldsymbol{x}_{i-1,j}^n| - |\hat{\boldsymbol{x}}_{i,j}^n - \hat{\boldsymbol{x}}_{i-1,j}^n|\right|^{\alpha} + \left||\boldsymbol{x}_{i,j-1}^n - \boldsymbol{x}_{i,j}^n| - |\hat{\boldsymbol{x}}_{i,j-1}^n - \hat{\boldsymbol{x}}_{i,j}^n|\right|^{\alpha}\right)\right] \tag{8}$$

$$L_{GAN}(D,G) = \mathbb{E}_{\boldsymbol{X}}[\log D(\boldsymbol{X})] + \mathbb{E}_{\hat{\boldsymbol{X}}}[\log(1 - D(G(\boldsymbol{X})))] \tag{9}$$

L_{MSE}, L_{GDL} and L_{GAN} are the mean square error (MSE), the image gradient difference loss (GDL), and the generative adversarial network (GAN) loss, respectively. The GDL is designed to minimize differences between the gradients of water depths in the original ($\boldsymbol{x}$) and reconstructed ($\hat{\boldsymbol{x}}$) maps. The GAN loss encompasses both a generator G, which represents the AE, and the PatchGAN discriminator D, as proposed by Isola et al. (2017). $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$ and $\hat{\boldsymbol{X}} = [\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_N]$ are the original and reconstructed maps, respectively. λ_{GDL}, λ_{GAN} and α are hyperparameters (see Section 2.1.4).

During the second training phase ("VPTR training"), only the VPTR block is trained, and the parameters of the encoder and decoder blocks remain fixed and equal to the optimized weights obtained from the AE training. This phase prioritizes the analysis of temporal information between subsequent water depth maps and the upstream boundary condition. Consequently, each batch in the VPTR training comprises a sequence of $I+1$ consecutive maps $\boldsymbol{X} \in \mathbb{R}^{N\times(I+1)\times H\times W}$, and the corresponding inflow discharges shifted of one time step. The VPTR block takes the latent features in output from the trained encoder $[z_1, \ldots, z_I]$ and the inflow discharge values $[q_2, \ldots, q_{I+1}]$ to predict the latent features $[\hat{z}_2, \ldots, \hat{z}_{I+1}]$ (see Figure 2a). These latent features are then used to reconstruct the predicted maps $[\hat{\boldsymbol{x}}_2, \ldots, \hat{\boldsymbol{x}}_{I+1}]$ through the previously trained decoder. The VPTR training procedure aims at minimizing the following loss function:

$$L_{VPTR} = \frac{1}{I}\sum_{t=2}^{I+1}(L_{MSE}(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_t)) + \lambda_{GDL}\frac{1}{I}\sum_{t=2}^{I+1}(L_{GDL}(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_t)) \tag{10}$$

For practical applications, the predicted frames within the range of $2 \le t \le I$ are typically not of primary interest during the forecasting procedure, as the corresponding target maps are already known, and they are provided within the input sequence. However, including these forecasted frames in the loss computation improves the prediction performance and, during the autoregressive procedure, allows to consider a lower number of frames used as initial conditions (i.e., the hyperparameter P) than the total sequence length I of the training process (see Section 2.1.3).

Once the FS model is trained (i.e., both training stages are completed), the accuracy in predicting the map at time $I+1$ is assessed using the unseen samples in the testing dataset. This procedure is referred to as "FS test".

2.1.3 Autoregressive prediction

Once trained and tested, the FS model can be applied for real-time inundation forecasting using an autoregressive (AR) procedure for producing water depth maps in F future frames. This involves substituting observed frames with predicted frames as input maps to recursively forecast subsequent maps (Figure 2b). Starting with P past frames (i.e., a few maps used as initial

https://doi.org/10.5194/hess-2024-176
Preprint. Discussion started: 5 August 2024

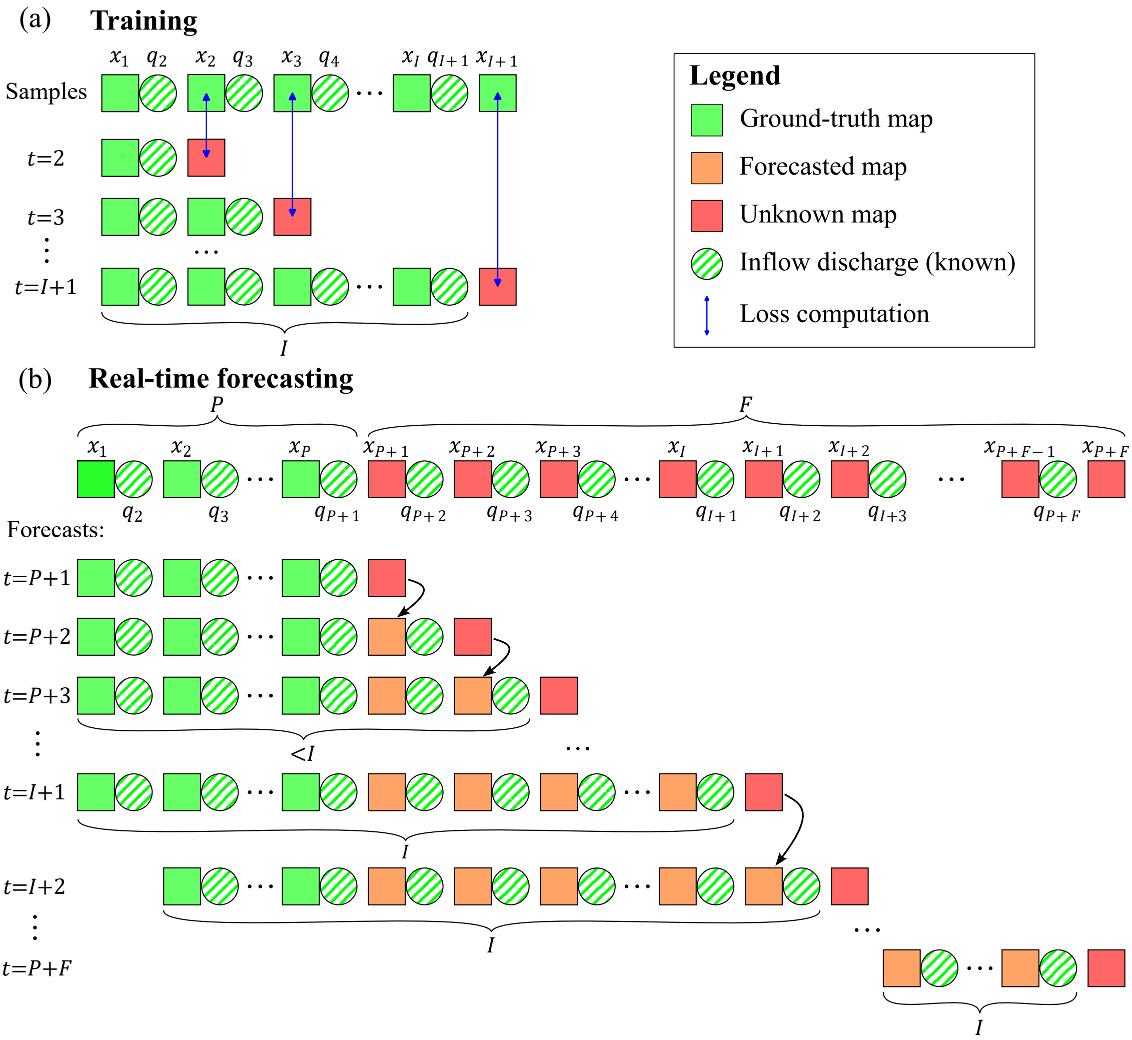
(a) Training
Samples
x_1 q_2 x_2 q_3 x_3 q_4 x_I q_{I+1} x_{I+1}
t=2
t=3
t=I+1
I
Legend
Ground-truth map
Forecasted map
Unknown map
Inflow discharge (known)
Loss computation
(b) Real-time forecasting
P
F
x_1 x_2 x_P x_{P+1} x_{P+2} x_{P+3} x_I x_{I+1} x_{I+2} x_{P+F-1} x_{P+F}
q_2 q_3 q_{P+1} q_{P+2} q_{P+3} q_{P+4} q_{I+1} q_{I+2} q_{I+3} q_{P+F}
Forecasts:
t=P+1
t=P+2
t=P+3
<I
t=I+1
I
t=I+2
I
t=P+F
I

Figure 2. Sketches of the VPTR training and real-time forecasting procedures. The squares represent the water depth maps, while the circles represent the discharge values. For simplicity, only the input and output maps of the prediction are illustrated, neglecting the latent feature computations. **(a)** Illustration of the VPTR training procedure on a sequence of $I+1$ frames. For example, the forecast of the frame at time $t = I+1$ (red square) is achieved by considering the ground-truth maps at instants $t \in [1, I]$ (green squares) and the discharge values at time $t \in [2, I+1]$ (green circles) as inputs. **(b)** Real-time forecasting of F future maps conducted through the autoregressive procedure. The inflow hydrograph spanning the entire duration of the prediction must be known (green circles). Each forecasted future frame (orange squares) is concatenated with the inflow discharge value of the following instant and used to predict the next future map (red squares). Starting from the prediction of the frame at time $t = I+2$, a sliding window is introduced to constrain the length of the input sequence to I. In this illustration we assumed $P < I < F$.

condition, with $P \leq I$) and the sequence of upstream discharge values for the entire flood event (i.e., $P+F$ values), the procedure recursively generates up to F forecasted future frames. The number of future frames that can be predicted is constrained by the availability of a sufficiently extended forecasting period for the inflow hydrograph and by the loss of accuracy due to the error accumulation during the AR procedure. We emphasize that the entire sequence of upstream inflow must be provided as input data to the surrogate model (green circles in Figure 2b); this is not a limitation of the surrogate model herein proposed but a necessity coming from the physics of the phenomenon we are trying to reproduce.

Figure 2b presents a sketch of the AR prediction process using the FS model. In the initial step of the recursive procedure (at $t = P+1$), the FS model takes the ground-truth maps of the past frames $[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_P]$ (green squares) and the corresponding inflow discharges $[q_2,\ldots,q_{P+1}]$ (green circles) to predict the first unknown future map $\hat{\boldsymbol{x}}_{P+1}$ (red square at $t = P+1$). Subsequently, the forecasted map $\hat{\boldsymbol{x}}_{P+1}$ (orange square at $t = P+2$) is associated to the discharge q_{P+2}, and concatenated at the end of the sequence of past frames. This new sequence of $P+1$ frames and discharges is then fed into the FS model to predict the map $\hat{\boldsymbol{x}}_{P+2}$ (red square at $t = P+2$). This process continues until all the F future frames are predicted. It is worth noting that, starting from the prediction of the frame at instant $t = I+2$ (i.e., when the sum of past frames P and concatenated ones exceeds I), the oldest maps are discarded to constrain the length of the input sequence to I, which is the maximum value allowed from the training phase.

Generally, the AR procedure can be summarized with the following formulation:

$$\hat{\boldsymbol{x}}_j = \begin{cases} Dec\left(\mathcal{T}(\boldsymbol{K},\boldsymbol{V} : Enc([\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{j-1}]);\ \boldsymbol{Q} : [q_2,\ldots,q_j])\right) & \text{if } P+1 < j \leq I+1 \\ Dec\left(\mathcal{T}(\boldsymbol{K},\boldsymbol{V} : Enc([\boldsymbol{x}_{j-P},\ldots,\boldsymbol{x}_P,\hat{\boldsymbol{x}}_{P+1},\ldots,\hat{\boldsymbol{x}}_{j-1}]);\ \boldsymbol{Q} : [q_{j-P+1},\ldots,q_j])\right) & \text{if } I+1 < j \leq I+P \\ Dec\left(\mathcal{T}(\boldsymbol{K},\boldsymbol{V} : Enc([\hat{\boldsymbol{x}}_{j-I},\ldots,\hat{\boldsymbol{x}}_{j-1}]);\ \boldsymbol{Q} : [q_{j-I+1},\ldots,q_j])\right) & \text{if } j > I+P \end{cases} \tag{11}$$

We emphasize that the AR procedure works even when $P < I$, i.e. when only a shorter sequence of initial condition maps is available. This ability stems from the methodology used for the loss computation during the VPTR training procedure. Indeed, as already mentioned in Section 2.1.2, the training loss is computed considering all the predicted frames in range $2 \leq t \leq I+1$.

2.1.4 Surrogate model implementation details

One of the primary objectives of the AE framework is to reduce the size of the input maps, thereby limiting the time and memory consumption during training. Hence, the number of convolutional layers (k) in the AE and the number of channels of the latent features (d_{model}) must be proportional to the dimensions of the input maps. Specifically, a higher value of k yields a lower dimension of the height and width of the latent features and a higher value of d_{model}. Therefore, the values of these hyperparameters depend on the size of the case study analyzed (refer to Table 1).

Following the original implementation of the video prediction transformer framework (Ye and Bilodeau, 2023), the VPTR module, whose structure has been previously described in Section 2.1.1, comprises 12 consecutive VidHRFormer blocks. The number of parallel attention heads (p) for the MHCA computation is set to 8. Additionally, the local spatial MHCA uses a local patch size of 4. A Sigmoid function serves as the output layer of the surrogate model. This activation function returns values within the range of $0-1$. Consequently, given a normalization of the samples in the dataset in the same range of values, the creation of non-physical negative water depths is automatically prevented.

Developing a large transformer-based model from scratch presents considerable challenges, often resulting in compromised predictive performance, especially when data resources are constrained, as noted by Bertasius et al. (2021). As the datasets utilized in our study contain significantly fewer samples compared to those typically employed for tasks such as video classification and video prediction, addressing the potential for overfitting (i.e., the model's reduced capability to generalize to unseen data) and minimizing training costs are pivotal considerations. Therefore, the AE and VPTR models were initialized with weights pretrained on a large dataset composed of thousands of sequences of video frames provided by Ye and Bilodeau (2023).

The number of input frames (I) was set to 8, based on a sensitivity analysis conducted by Pianforini et al. (2024a), which identified this value as optimal for balancing result accuracy with the memory and time consumption required for the training process. Similarly, we selected a batch size of 4 for analogous reasons.

As previously mentioned, the training procedure is divided into two phases. For the initial phase (AE training), we employed the Adam optimizer with beta values set to (0.5, 0.999) and a learning rate of 2e-4. Differently, for the training of the VPTR framework, we employed the AdamW optimizer with beta values set to (0.9, 0.999) and learning rate value of 1e-3. This training configuration is consistent with that employed in a previous study which applied the FS model to predict dam-break scenarios (Pianforini et al., 2024a). For the VPTR training, we implemented an early stopping technique with automatic training restarts, in order to stop the training process after a specified number of epochs without metric improvement and to subsequently restart it for a predetermined number of iterations. This approach aims to mitigate overfitting and enhance the accuracy of the autoregressive procedure of the surrogate model (Goodfellow et al., 2016).

For the loss computation (Eq. 6 and Eq. 10), we set $\lambda_{GDL} = 0.01$ and $\alpha = 1.0$ for both training phases. The value of λ_{GDL} was calibrated to obtain the same order of magnitude between L_{MSE} (Eq. 7) and L_{GDL} (Eq. 8) losses. Differently, the value of λ_{GAN} varies during the AE training procedure. Specifically, until the L_{GAN} loss (Eq. 9) converges, we set λ_{GAN} to 0.1. Then, we set $\lambda_{GAN} = 0$ to exclude its influence in the total loss computation (Eq. 6). The number of epochs required for the L_{GAN} loss convergence depends on the case study and the dimension of the AE model, typically ranging between 5 to 50 epochs. The values of all hyperparameters related to the loss computation were determined through a trial-and-error process.

2.2 Hydrodynamic Model

The water depth maps used as samples for the training and testing phases of the surrogate model were obtained through the PARFLOOD code (Vacondio et al., 2014, 2017), a hydrodynamic model that solves the fully dynamic SWE using the Finite Volume methodology. The efficient implementation of this model using the Computer Unified Device Architecture (CUDA) language implies a significant reduction in simulation time compared to conventional serial codes, owing to the exploitation of GPUs (Vacondio et al., 2014). The accuracy and efficiency of the PARFLOOD code have undergone rigorous validation through various challenging case studies, including river flood scenarios (e.g., Dazzi et al., 2021a, 2022; Ferrari et al., 2020, 2023). For an in-depth description on the model's details, the reader is referred to Vacondio et al. (2014).

Table 1. Case studies summary and surrogate model hyperparameters.

	Toce River flood	Po River flood	
		Res. 20 m	Res. 10 m
Spatial resolution $[\mathrm{m}]$	0.05	20	10
Number of cells	16,384	1,306,624	5,226,496
Temporal resolution	0.5 s	3 h	3 h
Maximum water depth in test dataset $[\mathrm{m}]$	0.14	25.6	25.6
Depth normalization value $[\mathrm{m}]$	0.18	28.0	28.0
Discharge normalization value $[\mathrm{m}^3/\mathrm{s}]$	0.25	13,000	13,000
ϵ_{wet} (for metric computation) $[\mathrm{m}]$	0.001	0.2	0.2
CNN layers (k)	3	5	6
Latent feature size ($h \times w \times d_{\mathrm{model}}$)	16 × 16 × 256	44 × 29 × 1024	44 × 29 × 1280
AE parameters [million]	11	182	300
VPTR parameters [million]	33	585	826

2.3 1D CNN model

In the present study, the 1D CNN model proposed by Kabir et al. (2020) is used to benchmark the accuracy of the FS model. The 1D CNN model predicts a water depth map at time t by using the series of inflow discharges between instants t and $t-r$ as input data. Input data are processed using a sequence of two convolutional layers and three fully connected layers to produce the corresponding inundation map. Consequently, the forecast relies solely on the temporal information of the boundary conditions, without considering neither the previous maps nor the spatiotemporal correlation between consecutive maps. This architecture was used as benchmark model in different studies concerning the simulation of flood events on study areas with various extensions, up to approximately 1,500 km^2, with a number of cells ranging from 100k to 3.7M (e.g., Donnelly et al., 2022; Fraehr et al., 2024).

2.4 Case studies

The FS model was trained and tested considering two very different case studies. The first case study involves impulsive flood events in the urbanized valley of the Toce River (Italy), reproduced at the laboratory scale (see Section 2.4.1). The Toce River case has been widely used in the literature for the validation of numerical models (e.g., Costabile et al., 2017; Ferrari et al., 2019; Xia et al., 2017). Consequently, it was adopted in this work to evaluate the proposed surrogate model. The second case study focuses on predicting river flood events along a stretch of the Po River in Italy (see Section 2.4.2). Unlike the Toce River case, Po River floods are characterized by slow flow dynamics and long propagation times, with flood events lasting from days to weeks. Additionally, the presence of defended floodplains in the study area significantly increases the prediction difficulty.

https://doi.org/10.5194/hess-2024-176
Preprint. Discussion started: 5 August 2024

The overall goal is to assess the surrogate model's capability to predict flood events with varying flow dynamics and different topographies.

For each case study, the water depth maps comprising the datasets were generated by running the hydrodynamic model multiple times with different upstream boundary condition (i.e., considering different flood events) and producing water depth maps at predefined time intervals. The samples for the training dataset were obtained by extracting sequences of I consecutive water depth maps from the numerical results and associating them with the corresponding series of inflow discharge values to create the input data, while the output data only includes the water depth map at time $I+1$. The sliding window of length I is moved in time to extract several samples for each simulation, and the procedure is repeated for all simulations. Overall, each flood event of duration $T+I$ provides T samples. To ensure a robust training procedure, several real and synthetic flood events were included in the dataset formation (see Table 2). The corresponding samples were randomly divided to form the training and validation datasets, with a ratio of 95% for the training and 5% for validation.

To assess the FS model's capability to generalize beyond the training data, we evaluated the performance of the surrogate model using the testing dataset, which comprises additional flood events unseen during the training process (see Table 2). The testing samples used for the FS test (i.e., evaluation of the model's performance in predicting the frame at time $I+1$; see Appendix A) were created following the same procedure described for the training dataset generation. Please notice that, to evaluate the surrogate model's ability to predict floods with different levels of severity, testing events with varying flood intensities and dynamics were considered.

The testing events were also used for the application to forecasting entire flood events using the autoregressive procedure. In this case, no samples are necessary. The recursive procedure only requires one water depth map representing the initial condition and the full time series of inflow discharges (i.e., for the entire event duration). The accuracy of the AR procedure is evaluated by computing the errors of recursively predicted sequence of inundation maps for the entire event duration against the ground-truth numerical results.

The water depth maps in the datasets are non-dimensionalized dividing them by a value slightly higher than the maximum simulated water depth for the specific case study (see Table 1). Similarly, for the inflow discharges, a value exceeding the largest peak discharge was chosen (see Table 1). This normalization procedure ensures that the samples are scaled within the range $[0,1]$, thereby enhancing the effectiveness of the training process and ensuring consistency with the activation function used in the output layer (i.e., the Sigmoid function; see Section 2.1.4).

We filtered out insignificant water depths in the numerical model's output maps by zeroing values lower than a predefined threshold. Specifically, we set thresholds equal to 1e-5 m and 0.05 m for the Toce and Po case studies, respectively.

Table 1 provides a summary of the different configurations and hyperparameters adopted. It is important to note that the two case studies have different spatial and temporal scales. Therefore, parameters such as spatial and temporal resolutions, normalization values, and the wet-dry threshold (ϵ_{wet}) need to be appropriately scaled to suit the specific characteristics of each case study.

Table 2. Training configurations for the two case studies. The training and validation datasets were obtained by dividing the total number of samples with a ratio of 95% for the training and 5% for the validation. The flood events included in the testing datasets differ from those used to generate the training datasets.

Training case	Spatial resolution [m]	Training and validation		Testing		Autoregressive prediction
		Flood events	# of samples	Flood events	# of samples	
Toce River						
Toce1	0.05	15 *Rapid*	1056	3 *Rapid* (*Low*, *Medium*, *High*) + 1 *Gradual*	396	Testing events (i.e., 4)
Toce2	0.05	15 *Rapid* + 7 *Gradual*	1782			
Po River						
Po1	20	19 real	2320	3 real (Nov 2011, Nov 2014, and Jun 2020)	423	Testing events (i.e., 3)
Po2	20	19 real + 6 synthetic	3144			
Po3	10	19 real + 6 synthetic	3144			

2.4.1 Toce River flood

The first case study involves an urban flash flood at the laboratory scale. The geometry is derived from a physical model developed by Testa et al. (2007), representing a 1:100 scale replica of a stretch of the Toce River valley in Italy. In the model, 18 concrete cubic blocks arranged in a staggered configuration were positioned in the centre of the domain to reproduce an urban district (Figure 3a). During their experiments, Testa et al. (2007) examined three inflow hydrographs with different values of the peak discharge (named as *Low*, *Medium*, and *High* in Figure 4c). Water depths were consistently monitored at 9 gauge points (labelled P2 to P10 in Figure 3a) throughout the entire duration of the experiments.

To generate the "ground-truth" water depth maps used for training and testing the surrogate model, the PARFLOOD model was setup using a Digital Terrain Model (DTM) with a spatial resolution of 0.05 m. A calibration process was conducted by comparing experimental and simulated water depths at control points P2$-$P10, to ensure the reliability of the ground-truth maps for the surrogate model's training. The results of the calibration confirmed the adoption of a uniform Manning roughness coefficient equal to 0.0162 $\mathrm{sm}^{-1/3}$, as suggested by Testa et al. (2007). In all numerical simulations, a far-field boundary condition was imposed downstream, and the domain was considered initially dry. The water depth maps were sampled at intervals of 0.5 s. This temporal resolution is needed to adequately describe the rapid flood propagation within the study domain.

To build a comprehensive dataset of water depth maps sequences, we considered different inflow hydrographs (illustrated in Figure 4), derived from recorded discharges in a gauge station on the Toce River, which were then scaled in time and magnitude to reproduce a type of flood compatible with the scale of the physical model. The three inflow hydrographs recorded by Testa et al. (2007) during their experiments and 23 synthetic events were used as upstream boundary conditions to run the numerical simulations for the dataset creation.

https://doi.org/10.5194/hess-2024-176
Preprint. Discussion started: 5 August 2024

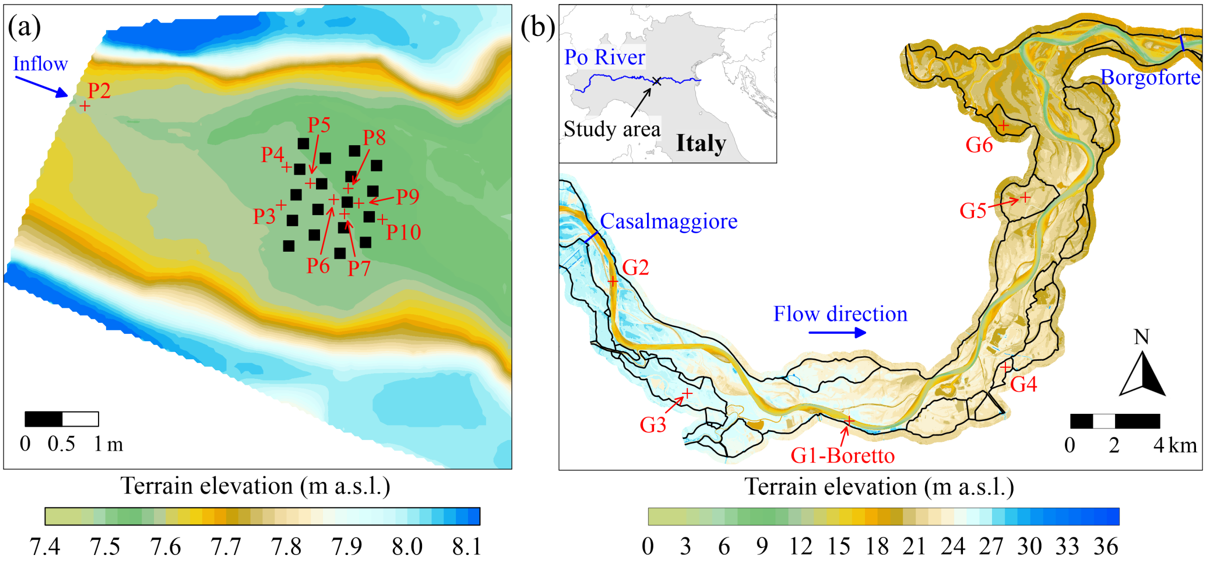

Figure 3. Study area for the two case studies. **(a)** Toce River case study. The black squares represent the urban district. **(b)** Po River case study. The black lines represent the Po River levees.

The present case study is used to: (a) verify the performance of the FS model in predicting impulsive flood events with complex flow dynamics, and (b) assess how the type of flood events in the training dataset influences the accuracy of the FS model in forecasting unseen floods with varying characteristics. Consequently, the surrogate model underwent two distinct training processes based on two different training datasets (see Table 2). The first training case, denoted as *Toce1*, is characterised by a dataset encompassing 15 synthetic events (Figure 4a), referred to as *Rapid* hydrographs, which feature a steep rising limb and a peak occurring within the initial 6 s, similar to the trend of the experimental tests. Differently, the second training case, named *Toce2*, considers a training dataset comprising the previous 15 *Rapid* flood events along with additional 7 synthetic hydrographs, named as *Gradual* (Figure 4b), with slower rising limb compared to the *Rapid* hydrographs.

The testing dataset remains consistent across both training processes (Table 2) and includes samples from 3 simulations of experimental tests and 1 synthetic hydrograph exhibiting a *Gradual* trend (Figure 4c), which were unseen during the two training processes.

2.4.2 Po River flood

The second case study focuses on fluvial flood prediction in a 48 km-long stretch of the Po River in Italy, situated between the stations of Casalmaggiore and Borgoforte (Figure 3b). In the study area the river varies in width from 300 m to 3 km, whereas the main channel itself maintains a width of approximately 300 m. Additionally, the study area includes defended floodplains, confined by minor levees designed to withstand floods with a return period higher than 50 years. Moreover, main embankments exceeding 8 m in height protect the extensive lowland surrounding the Po River from a 1 in 200 years' flood.

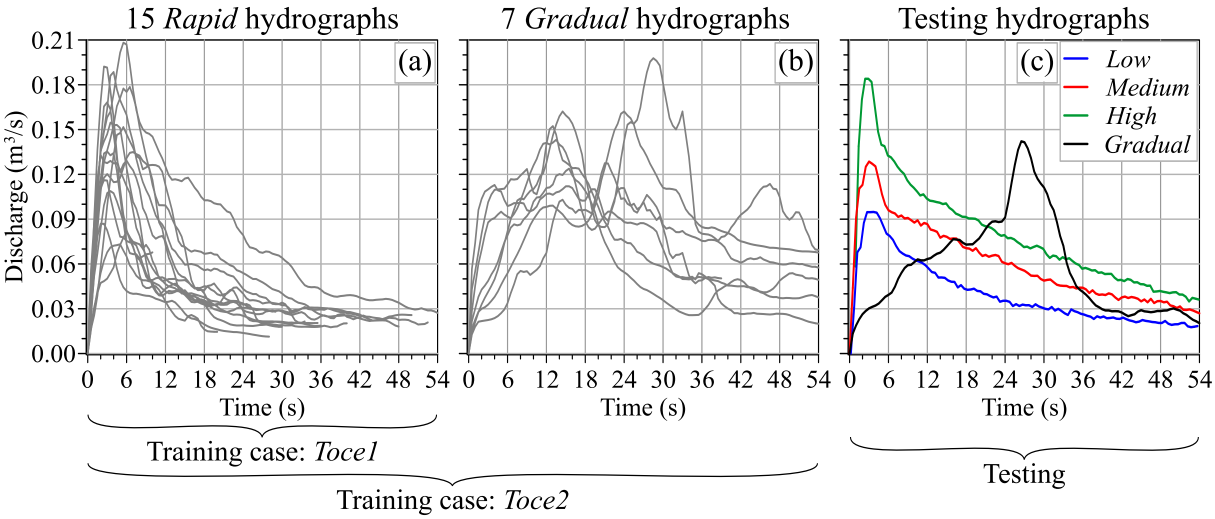

Figure 4. Inflow hydrographs of the Toce River case study. **(a)** 15 *Rapid* hydrographs used to generate the training dataset for the *Toce1* case. **(b)** 7 additional *Gradual* hydrographs utilized for creating the *Toce2* training dataset. **(c)** 3 recorded hydrographs (*Low*, *Medium*, and *High*) and one synthetic hydrograph (*Gradual*) used to generate the testing dataset.

The presence of these elements significantly complicates the dynamics of flood propagation, which requires fully 2D hydraulic models for accurate simulation (Dazzi et al., 2021a). Therefore, the ground-truth maps were obtained using the 2D SWE solver PARFLOOD.

A 2 m resolution DTM of the study area was created by merging LIDAR and bathymetric surveys. To alleviate the computational load, the original DTM was downsampled to resolutions of 10 m and 20 m, leading to grids with approximately 5.2M and 1.3M cells, respectively (see Table 1). These spatial resolutions were deemed suitable for accurately simulating flood propagation in the Po River region, given the main channel's width exceeding 200 meters. To ensure accurate representation of levee overtopping, the crest elevations of the main and minor embankments were retained in the downsampled grids, utilizing data from terrestrial surveys. All numerical simulations assume the presence of non-erodible embankments. Therefore, defended floodplains are inundated only after the overtopping of minor levees.

The upstream boundary condition was imposed at the Casalmaggiore section, while a rating curve was applied at the Borgoforte section downstream. To construct the datasets, 22 historical flood events occurring between 2000 and 2021 were considered. The corresponding water levels, recorded by the Casalmaggiore gauge station, were converted into discharge values using a rating curve (Figure 5a,c). Additionally, 6 synthetic events with peak discharge exceeding 8,000 m^3/s were considered (Figure 5b). The temporal resolution of all inflow hydrographs and output water depth maps was 3 hours, which is deemed suitable to describe the rather slow propagation of the Po River floods. For each numerical simulation, a water depth map obtained from steady-flow discharge values ranging from 500 m^3/s to 2,500 m^3/s was considered as initial condition.

The FS model was trained three times (see Table 2) with the purpose of: (a) analyzing the influence of the training dataset (type and severity of flood events) on the ability of the model to generalize on unseen floods, and (b) examining how the accu-

https://doi.org/10.5194/hess-2024-176
Preprint. Discussion started: 5 August 2024

racy and computational times are influenced by the spatial resolution (i.e. number of cells) used to discretize the domain. For the first purpose, we compared training cases *Po1* and *Po2*. The first one (*Po1*) was conducted employing a dataset exclusively comprised of 19 historical flood events. The corresponding inflow hydrographs, depicted in Figure 5a, exhibited peak discharges ranging from approximately 2,000 m^3/s to 13,000 m^3/s, with total flood duration spanning between 11 and 30 days. Moreover, many events displayed multiple peaks and an oscillatory trend in inflow discharges. Since the observation period is characterized by a scarcity of low-frequency and high-intensity floods, the FS model was also trained using a second dataset (*Po2*) incorporating 6 additional synthetic hydrographs with high peak discharges, as depicted in Figure 5b. The purpose was to determine if a surrogate model trained on a wider and more balanced dataset is characterized by higher accuracy compared to a model trained on historical data only. For the second purpose, we considered a third training case (*Po3*) that exploits inundation maps with doubled the spatial resolution (from 20 m to 10 m). The model was trained with samples obtained from the same flood events as the *Po2* and thus we compared the performance of the last two configurations.

The accuracy in predicting unseen flood events was assessed by simulating three real floods (Figure 5c), which occurred in November 2011, November 2014, and June 2020. These hydrographs exhibited varying discharge peak values ranging between 2,500 m^3/s and 9,700 m^3/s, with flood event duration spanning from 13 to 22 days. Furthermore, the maximum extent of the inundation differed quite significantly across these events. For the June 2020 flood, which had the lowest peak discharge, flooding was confined in the main channel. In contrast, the November 2011 flood affected open floodplains due to the higher peak discharge value. The most severe event, the November 2014 flood, caused inundation of most of the defended floodplains. These diverse testing events thus allow for a comprehensive assessment of the surrogate model's performance across different types of floods, ranging from minor to severe events. The testing dataset includes unseen flood events and remains consistent across all the three training processes (Table 2).

Please notice that the roughness coefficients of the numerical model were calibrated by simulating the three real flood events of the testing dataset. Two distinct Manning coefficients were employed for the main channel and floodplains, set at 0.03 $\mathrm{sm}^{-1/3}$ and 0.045 $\mathrm{sm}^{-1/3}$, respectively. These values were determined to minimize the differences between simulated and recorded water levels at the Boretto gauge station, located few kilometres downstream the Casalmaggiore section (see point G1 in Figure 3b). The results of the calibration procedure are shown in Section 3.2.

To facilitate a comprehensive comparison between the results obtained with the numerical and surrogate models, water depths were extracted at 6 predefined control points situated along the main channel and within the defended floodplains (named G1−G6 in Figure 3b). Notably, point G1 is positioned at the Boretto gauge station, enabling the additional direct comparison with recorded measurements.

2.5 Performance indicators

The accuracy of the FS model is evaluated by comparing the water depth maps predicted by the surrogate model with those simulated by the hydrodynamic model, assumed as ground truth. Two metrics are employed for assessment: the root-mean square error (RMSE), which is a regression metric quantifying differences between target and predicted maps, and the F1 score, a classification metric providing an estimation of the surrogate model's capability to predict the extent of flooded areas.

https://doi.org/10.5194/hess-2024-176
Preprint. Discussion started: 5 August 2024

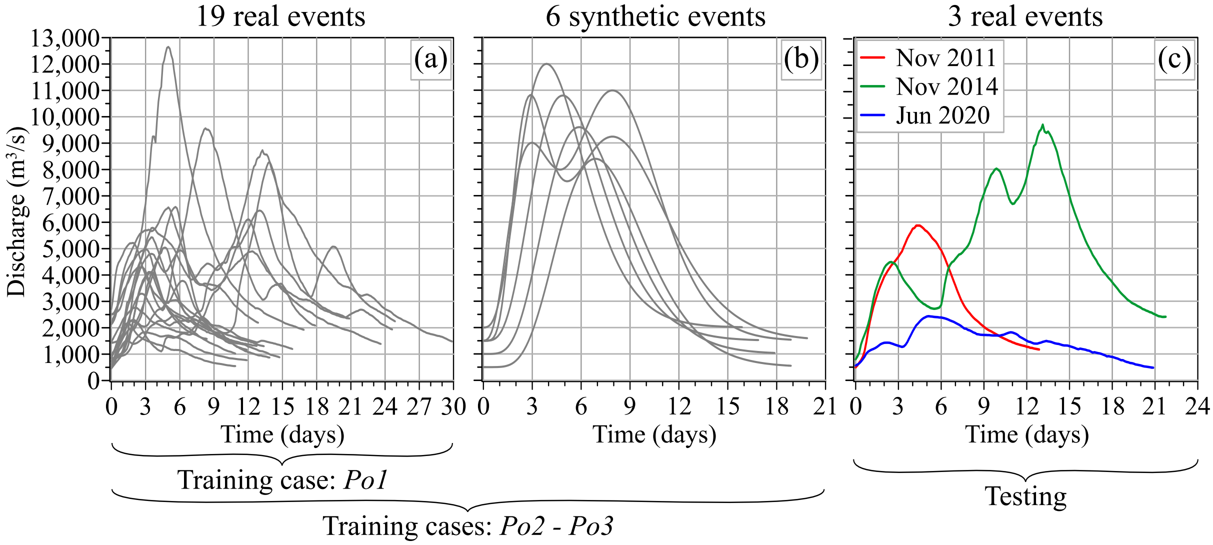

Figure 5. Inflow hydrographs of the Po River case study. **(a)** 19 real events occurring between 2000 and 2021 used to create the training dataset for the *Po1* case. **(b)** 6 additional synthetic scenarios in the training dataset for the *Po2* and *Po3* cases. (c) November 2011, November 2014, and June 2020 recorded events (unseen during training processes) used to generate the testing dataset.

The two metrics are defined as follows:

$$\text{RMSE}_t = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_t^n - \hat{y}_t^n)^2} \quad \text{for} \quad t \in [1,T] \tag{12}$$

$$\text{F1} = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{13}$$

where y_t^n and $\hat{y}_t^n$ denote the ground-truth and predicted water depths in the n-th wet cell at instant t, respectively. N represents the number of wet cells (i.e., cells with water depth higher than a specified threshold ϵ_{wet}) in the map. T indicates the total number of temporal frames considered in the analysis. In Eq. 13, TP denotes the number of cells correctly predicted as flooded (i.e., true positives), FN represents the count of cells wrongly predicted as non-flooded (i.e., false negatives), and FP is the number of cells wrongly predicted as flooded (i.e., false positives).

As anticipated, to differentiate between wet and dry cells in both the predicted and ground-truth maps during metrics calculation, a water depth threshold ϵ_{wet} was employed. The threshold value should be a small fraction of the maximum water depth expected in the dataset, denoted as H_{max}. For the case studies here considered, the ratio $\frac{\epsilon_{wet}}{H_{max}}$ ranged between 0.60% and 0.75% (see Table 1).

To facilitate the comparison across different flood magnitudes, a non-dimensional RMSE was also computed by dividing the RMSE of a specific instant t by the average water depth of the ground-truth map at the same instant:

$$\text{RMSE\_ND}_t = \frac{\text{RMSE}_t}{\frac{1}{N}\sum_{n=1}^{N} y_t^n} \tag{14}$$

https://doi.org/10.5194/hess-2024-176
Preprint. Discussion started: 5 August 2024

3 Results

In this Section, we present the outcomes of the autoregressive forecasting performed by the FloodSformer model. As regards the FS test, whose purpose is to check the accuracy of the trained model in predicting only the maps at time $I+1$ (see Section 2.1.2), results are presented in Appendix A.

Once the data-driven model is trained and tested, it can be applied to forecast unseen sequences of water depth maps of flood events with an autoregressive procedure. The flood events of the testing dataset were also used to assess the model's accuracy for this procedure, which starts with only 1 frame as initial condition (i.e., $P = 1$). This means that the FS model can predict the water depth maps of an entire flood event starting from just one map of initial conditions and from the discharge hydrograph at the upstream boundary condition, making the model suitable for real-time forecasting applications.

Please notice that the FS model provides water depth maps as output. However, in order to ease the analysis of results, only selected examples of maps are shown in this Section, while a more in-depth discussion is based on the water depth values at selected control points. Furthermore, it is important to note that the two cases analyzed have different spatial and temporal scales. Consequently, the results and errors are characterized by different orders of magnitude.

3.1 Toce River flood

For the Toce River case study, the PARFLOOD code was calibrated by comparing the simulated water depths at 9 gauge points (refer to Figure 3a for locations) for the three experimental hydrographs (i.e., *Low*, *Medium*, and *High*) with the corresponding observed values obtained from experiments conducted in the physical model by Testa et al. (2007). This comparison is fundamental to verify the reliability of the ground-truth maps used for training the surrogate model. Figure 6 shows the results for 3 control points (P3, P7, P10), while the remaining points are presented in Figures S1-S4 in the Supplement. The results show that, although some discrepancies are observed within the urban area (e.g., gauge P7 in Figure 6), which are common to other numerical models (e.g., Xia et al., 2017), the PARFLOOD code captures the general flood dynamics quite accurately. Consequently, the datasets used for the surrogate model training and testing were generated numerically with this hydrodynamic model, as detailed in Section 2.4.1.

The surrogate model was trained twice, with different training configurations (i.e., *Toce1* and *Toce2* in Table 2). The prediction accuracy of the FS models was evaluated by forecasting the four flood events of the testing dataset (Figure 4c). The number of future frames (F) recursively predicted by the surrogate model was set to 106, corresponding to a lead time of 53 s.

Table 3 summarises the average metrics computed for the autoregressive predictions of the different testing scenarios. Generally, RMSE values depend on the type of scenario (*Rapid* or *Gradual*), as well as on the dataset used for training the surrogate model. The extremely high F1 score confirms the FS model's high accuracy in predicting the temporal variation of the flood extent throughout the entire duration of the flood scenario.

Initially, we focused on the FS model trained using the *Toce1* configuration. As already mentioned, the dataset for this training case comprises only *Rapid* hydrographs (i.e., discharge peaks occurring within the initial 6 s of the flood). Figure 6 shows the water depths extracted from the ground-truth (dashed black lines) and predicted (red lines) maps in three control

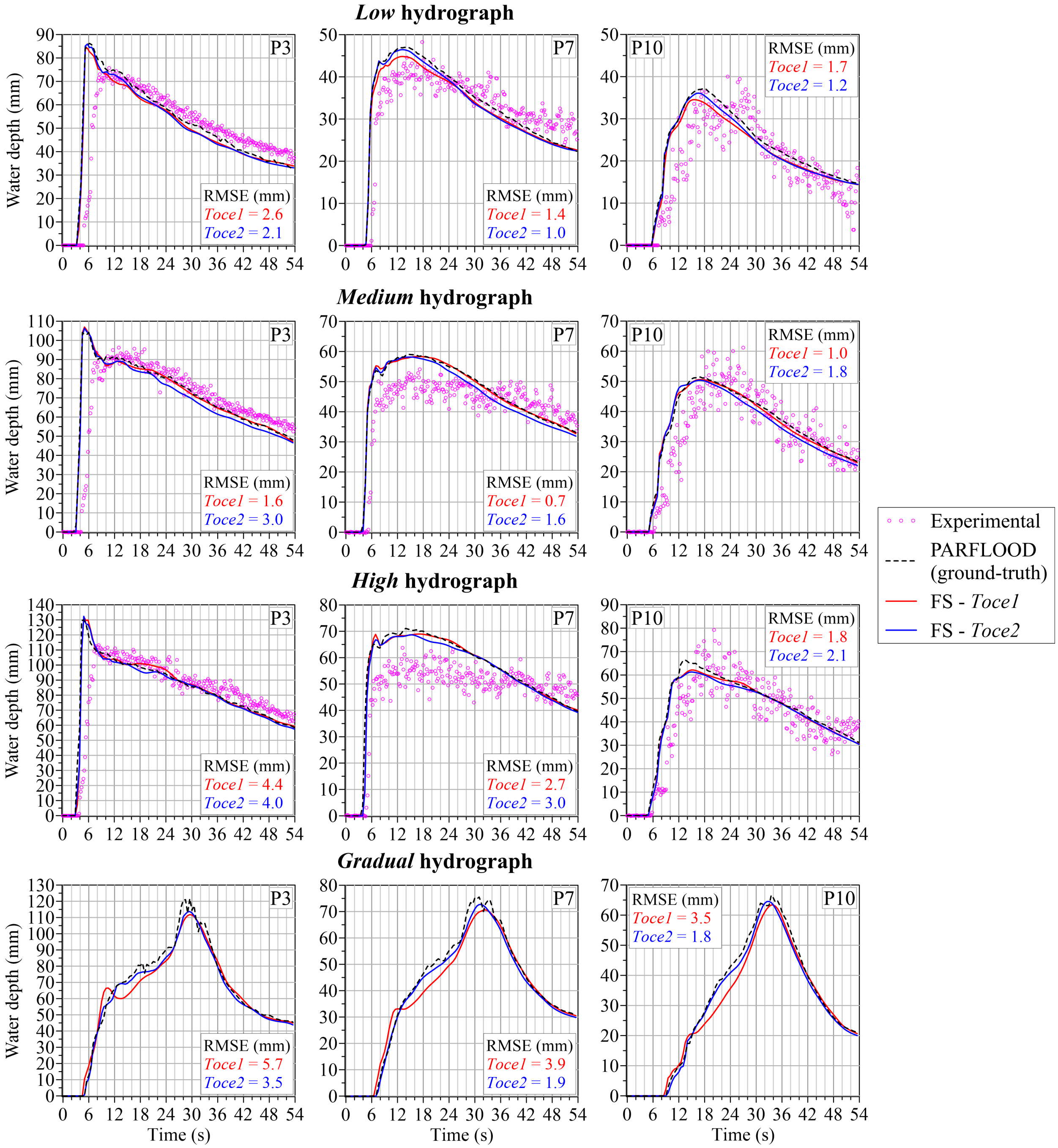

Figure 6. Comparison of recorded and simulated water depths at 3 control points (P3−P7−P10) for the Toce River case study. Each row corresponds to one of the four inflow hydrographs in the testing dataset (i.e., *Low*, *Medium*, *High*, and *Gradual*). The magenta circles represent the recorded water depths from the experimental analysis in the physical model (Testa et al., 2007). Each graph includes the RMSE values computed by comparing the time series of ground-truth and predicted water depths for both the *Toce1* and *Toce2* training configurations.

Table 3. Toce River case study: average RMSE, RMSE_ND (Eq. 14), and F1 score computed for the 106 recursively forecasted maps across the four flood events.

Training case	RMSE (mm)				RMSE_ND (-)				F1 (-)			
	Low	*Medium*	*High*	*Gradual*	*Low*	*Medium*	*High*	*Gradual*	*Low*	*Medium*	*High*	*Gradual*
Toce1	2.1	2.0	3.5	4.8	0.061	0.044	0.067	0.123	0.994	0.995	0.991	0.988
Toce2	1.8	2.6	3.2	3.6	0.056	0.060	0.061	0.083	0.995	0.995	0.991	0.991

points (P3, P7, P10; see Figure 3a for locations), for all testing events, while the results for the other control points are plotted in Figures S1-S4 in the Supplement. For the *Low*, *Medium*, and *High* hydrographs, the surrogate model shows remarkable accuracy in predicting the arrival time of the flood at each point. Additionally, the model accurately reproduces the water depth peaks, particularly notable for point P3, situated just upstream of the urban district. Among these hydrographs, the FS model exhibits the highest accuracy in predicting the *Medium* event. Comparing the time series of ground-truth and predicted water depths at control points, the average RMSEs are found to be lower than 1.6 mm. If the average RMSE of a control point (e.g., 1.6 mm at P3) is made non-dimensional with the maximum water depth expected at the same location (e.g., 104 mm at P3), we obtain a relative error lower than 2%. For the *Low* and *High* hydrographs, the surrogate model tends to slightly underestimate the water depths in some control points. Nevertheless, the above-defined relative error remains lower than 5% confirming the good performances of the FS model.

As expected, the FS model trained with the *Toce1* configuration generates accurate results in forecasting the flood propagation of the three experimental hydrographs, as their trend is similar to those of the *Rapid* hydrographs in the training dataset. In contrast, the surrogate model exhibits lower accuracy in predicting the flood generated by the *Gradual* hydrograph of the testing dataset. This discrepancy arises due to significant differences between the characteristics of this inundation scenario and the *Rapid* flood events in the *Toce1* training dataset. Specifically, the *Gradual* hydrograph features a discharge peak occurring approximately 26 s after the flood begins, contrasting with *Rapid* hydrographs, with discharge peaks within the initial 6 s. For this case, the FS model predicts the water depth peak with lower accuracy and shows an incorrect trend in forecasting water depths during the rising limb of the flood. For instance, focusing on point P3 in Figure 6, the predicted water depth begins to decrease around $t = 10.5$ s, contrary to the expected trend of monotonic increase. Similar discrepancies, although less pronounced, are observed at points P7 and P10.

Next, we analyze how the accuracy changes when the surrogate model is re-trained using the *Toce2* dataset (see Table 2), including both *Rapid* and *Gradual* types of floods (see Figure 4). In Figure 7, the predicted water depth maps for some representative instants of the *Medium* flood event are compared with the ground-truth maps derived from the hydrodynamic model. Overall, the surrogate model shows high accuracy in predicting the entire flood event. Errors are generally below 10% of the average water depth of this scenario (see Table A2) across most of the study domain. The main discrepancies are associated with the uncertainty in predicting the position of the wet/dry front of the flood, where differences are in the range $10-20$ mm. Despite these disparities, the surrogate model effectively captures the dynamics of the flood event.

https://doi.org/10.5194/hess-2024-176
Preprint. Discussion started: 5 August 2024

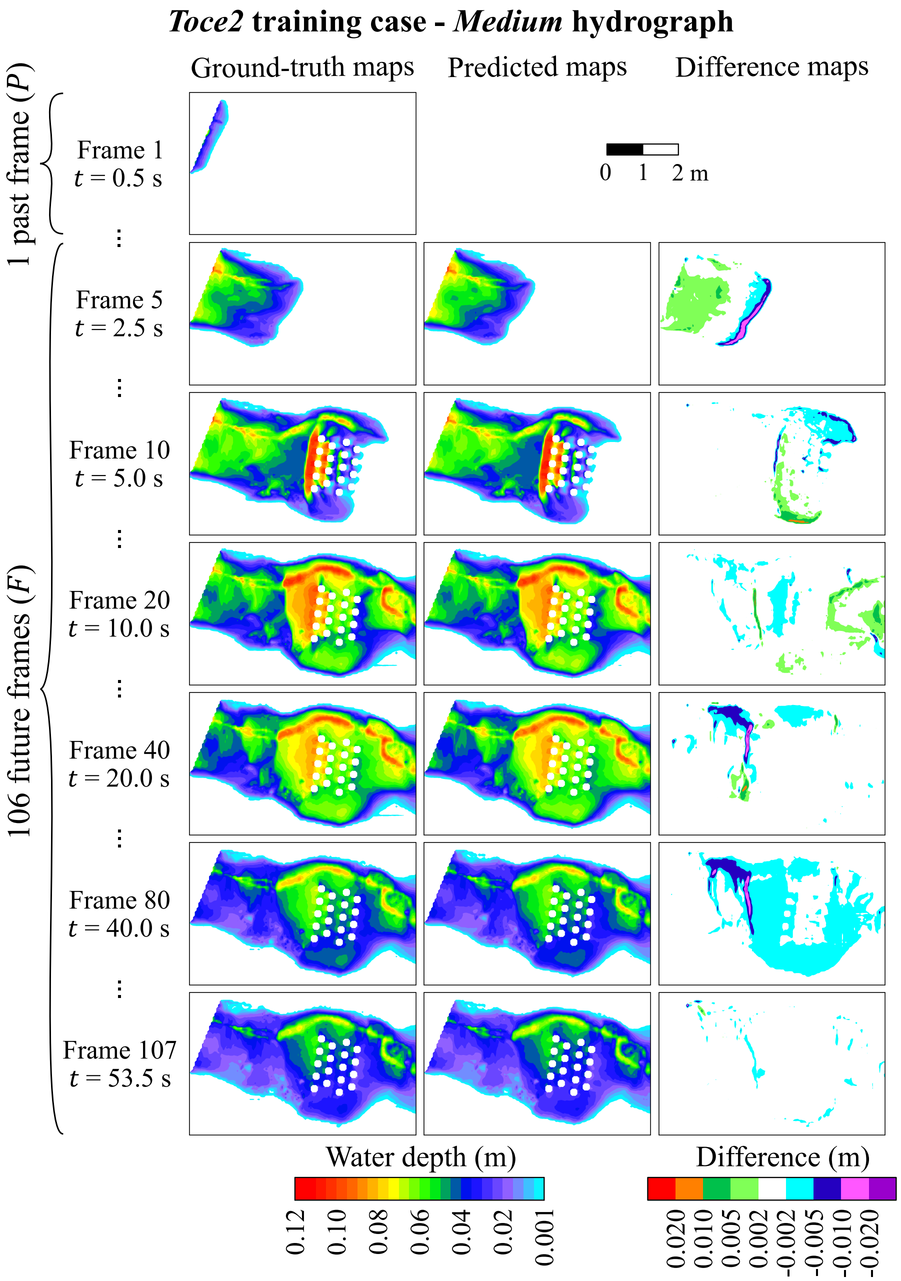

Figure 7. Toce River case study: real-time forecasting of the *Medium* hydrograph using the FloodSformer model trained with the *Toce2* configuration. The columns represent, respectively, the ground-truth maps obtained from the hydrodynamic model, the maps predicted by the surrogate model, and the difference maps between predicted and ground-truth maps. Only selected representative instants are shown.

505 The comparison between the water depths extracted at control points for the two training configurations is shown in Figure 6 and Figures S1-S4 in the Supplement, with the *Toce2* results depicted as blue lines. Focusing on the *Gradual* hydrograph, the

surrogate model trained using the *Toce2* configuration exhibits higher accuracy in predicting water depths at the control points, with average RMSEs approximately halved. Moreover, the predicted water depths exhibit a monotonic increase during the rising limb of the flood, closely following the ground-truth data trend. This improvement indicates that expanding the dataset to include *Gradual* hydrographs helps overcoming the issue of incorrect trends in the rising limb observed in the *Toce1* training configuration. The increased accuracy is further confirmed by comparing the forecasted water depth maps obtained with the two trained models, as shown in Figure S8 in the Supplement.

The addition of gradual hydrographs in the *Toce2* training dataset does not impair the accuracy in predicting *Rapid* floods. Indeed, comparable accuracy for the *Low* and *High* events, and only slight increased RMSEs at some control points for the *Medium* event, are achieved. In general, the ratio between average RMSE and maximum water depth expected at a specific control point is lower than 2.5−3.5%. The comparison of water depth maps predicted using the two training configurations for the three experimental hydrographs is shown in Figures S5-S7 in the Supplement. Additionally, to analyse the accuracy in forecasting water depths across the entire study domain, Figure 8 illustrates the temporal variation of the RMSE computed for all wet cells in the domain for the testing dataset for both the *Toce1* and *Toce2* configurations. Generally, for the *Low*, *Medium*, and *High* hydrographs, the RMSE is relatively high for the first predicted frames due to the underestimation of the wet/dry front of the flood (see Figure 7). This is mainly correlated to the high propagation velocity of the inundation front for the initial part of the event. However, after the flood impacts against the blocks representing the urban district, the errors progressively decrease. Differently, the accuracy of forecasting the *Gradual* hydrograph is strongly correlated with the dataset used for the training procedure. Specifically, the model trained using the *Toce1* dataset exhibits relatively high RMSE_ND values for the first 30−36 s of the flood event. In contrast, the model trained with the *Toce2* dataset generates significantly lower errors in the first 18−24 s of the flood. This reduction in errors is attributed to the use of a dataset containing events more similar to the one considered for testing (i.e., *Gradual* hydrograph). Furthermore, for all events, the higher discrepancies are associated with the hydraulic jump forming in the region upstream of the urban district, in addition to errors near the wet/dry front of the flood for the first predicted frames (see Figures S5-S8 in the Supplement).

3.2 Po River flood

For the second case study, the PARFLOOD model was calibrated comparing the water depths measured by the Boretto gauge station (magenta circles in control point G1 of Figure 9) with corresponding simulated values (dashed black line in control point G1 of Figure 9) for the 2011, 2014 and 2020 flood events (see Figure 5c). Notably, the hydrodynamic code is able to reproduce the flood dynamics along the designated stretch of the Po River. As a result, these findings confirm the reliability of the ground-truth maps used for training the data-driven model.

To assess the impact of various spatial resolution and the type of flood scenarios composing the training dataset, the FS model was trained considering the three configurations detailed in Table 2 and in Section 2.4.2, for which the predictive accuracy in forecasting unseen flood events was evaluated based on the testing dataset (Figure 5c).

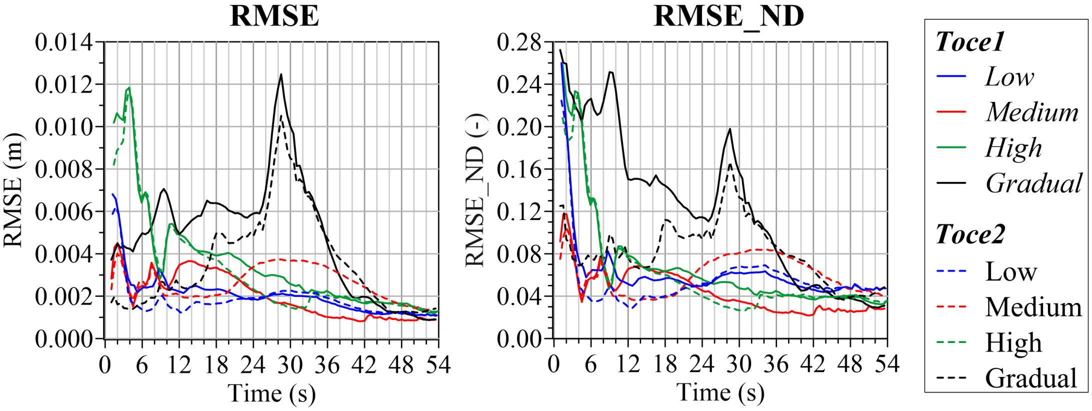

Figure 8. Toce River case study: RMSE and RMSE_ND (Eq. 14) computed for the maps forecasted by the surrogate model for the four hydrographs in the testing dataset. The continuous and dashed lines represent the results for the *Toce1* and *Toce2* training configurations, respectively.

3.2.1 *Po1* training case

In the first training case, named *Po1*, a spatial resolution of 20 m and a training dataset composed of 19 real flood events (Figure 5a) are adopted. The average metrics for the autoregressive prediction are summarized in Table 4. The average RMSE is relatively low (i.e., approximately $0.1-0.2$ m) for the 2011 and 2020 flood events, while it increases to about 0.5 m considering the 2014 flood. To understand the reasons of these discrepancies, the water depth time series at control points extracted from ground-truth (dashed black lines) and predicted (red lines) maps for all the testing scenarios are compared in Figure 9. Generally, the surrogate model shows high accuracy in predicting water depths in the main channel. For the most severe flood scenario in the testing dataset (i.e., the November 2014 event), the average RMSEs at control points G1 and G2 (located in the main channel) are below 0.1 m. Furthermore, the model accurately reproduces the temporal variation of water depths and the arrival time of the flood at control points G3 and G6, situated in two defended floodplains, with average RMSEs lower than 0.11 m. Conversely, the surrogate model exhibits shortcomings in forecasting water depths in other defended floodplains. For instance, the predicted arrival time of the flood at control point G4 (Figure 9a) has a delay of approximately 30 hours, while the maximum water depth is underestimated by approximately 0.7 m. More significantly, the surrogate model completely fails in forecasting the inundation of the defended floodplain where the control point G5 is located. For the November 2014 event, the FS model predicts that this area remains dry, while the numerical code simulates water depths exceeding 4 m. To confirm these results, in Figure 10 the ground-truth and predicted maps for selected instants of the November 2014 flood event are compared. Large differences in some defended floodplains are evident.

The failure of the data-driven model to correctly simulate the flood dynamics in some defended floodplains can be attributed to the types of flood events that are composing the *Po1* training dataset. Specifically, around 80% of the recorded hydrographs have a peak discharge lower than 7,000 m^3/s, while low-frequency and high-intensity floods are scarce (the largest peak dis-

https://doi.org/10.5194/hess-2024-176
Preprint. Discussion started: 5 August 2024

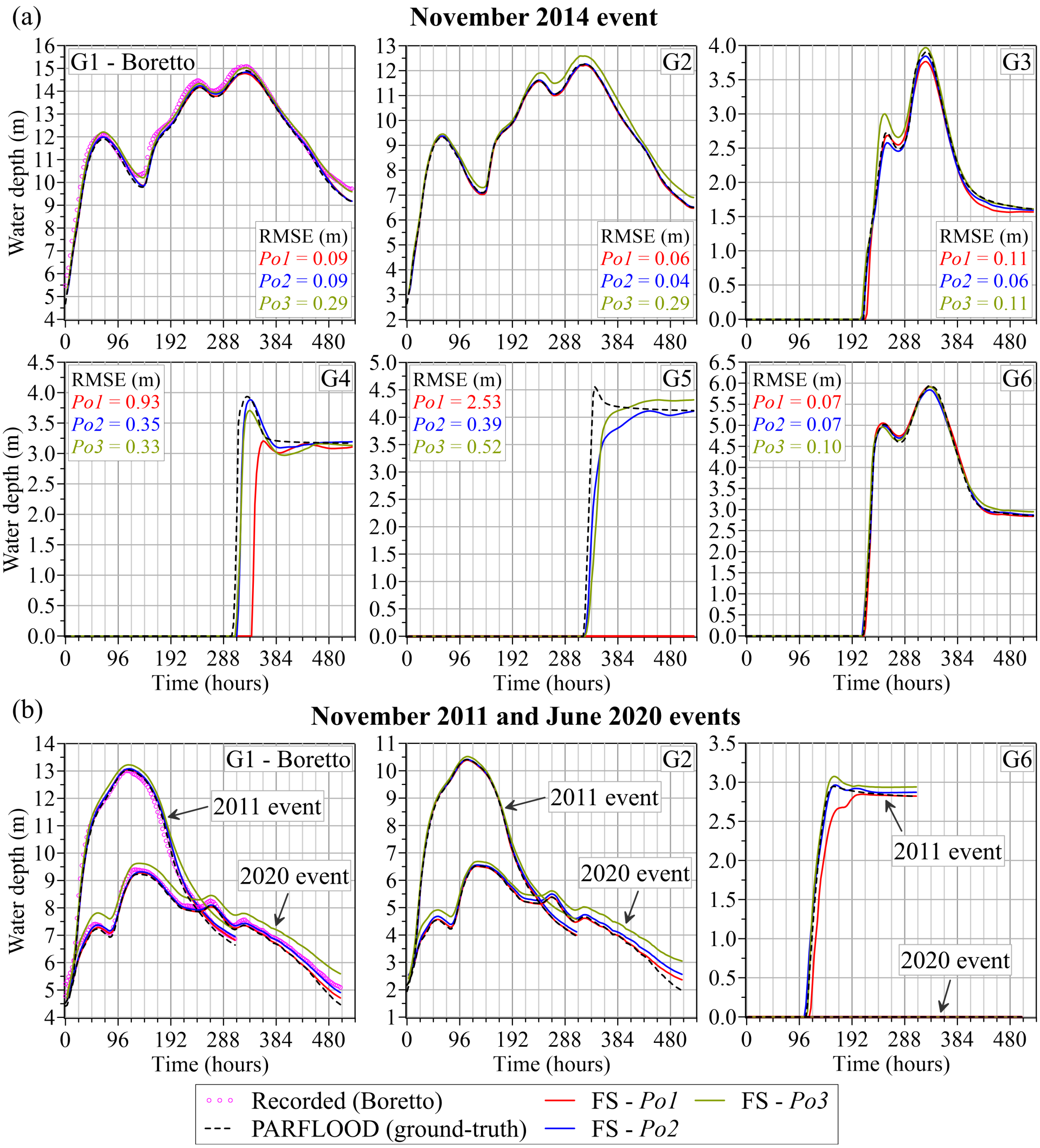

Figure 9. Comparison between recorded and simulated water depths at control points for the Po River case study. The magenta circles represent the recorded water depths at the Boretto gauge station (point G1). **(a)** November 2014 flood event. For each control point, the average RMSE of the time series of ground-truth and predicted water depths using the surrogate model trained with the three configurations (i.e., *Po1*, *Po2*, and *Po3*) is reported. **(b)** November 2011 and June 2020 flood events. For these events, the floodplains with control points G3, G4 and G5 remain dry.

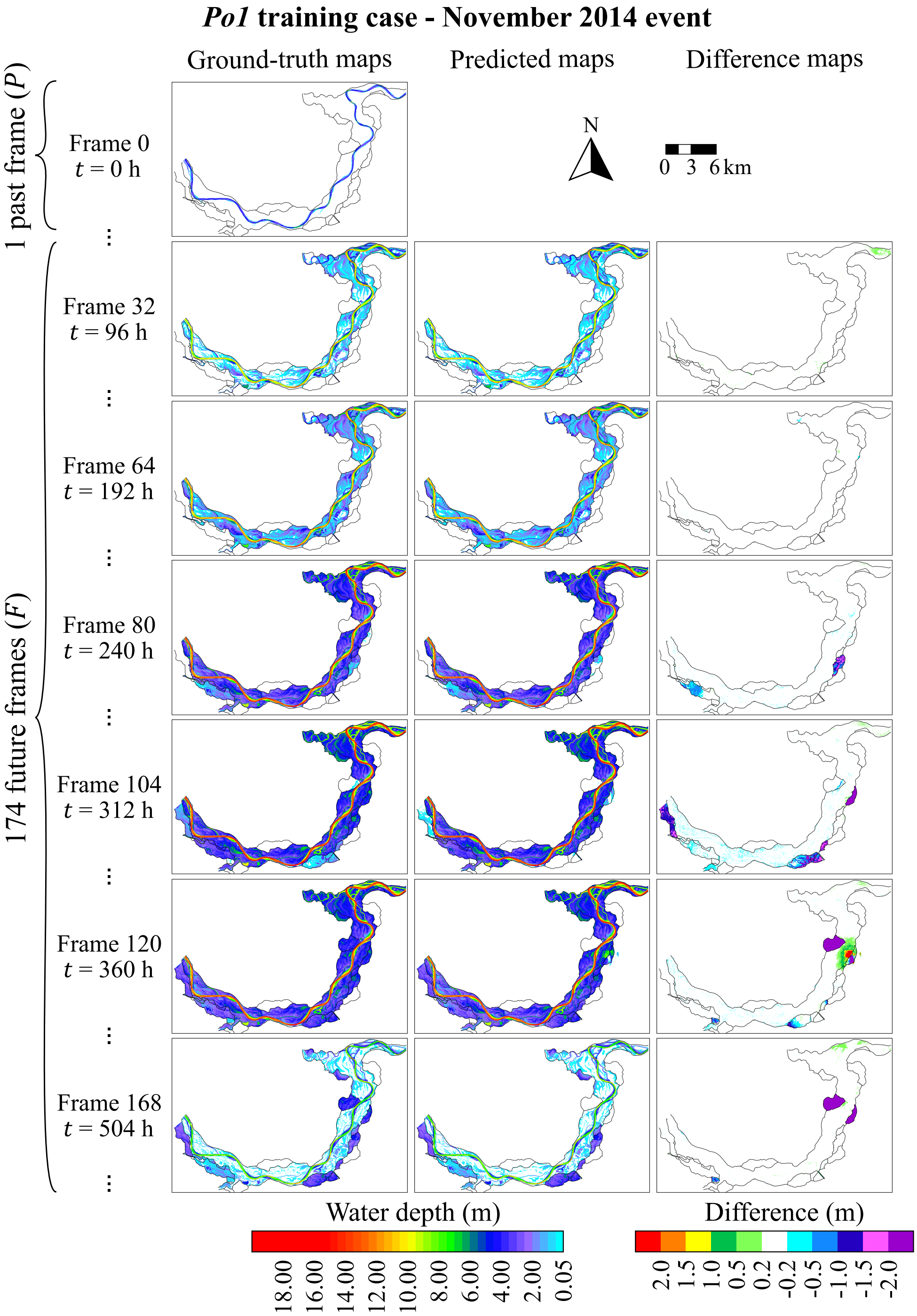

Figure 10. Po River case study: real-time forecasting of the November 2014 flood event using the FloodSformer model with the *Po1* training configuration. The columns represent, respectively, the ground-truth maps obtained from the hydrodynamic model, the maps predicted by the surrogate model, and the difference maps between the predicted and ground-truth maps. Only selected representative instants are shown.

Table 4. Po River case study: average RMSE, RMSE_ND (Eq. 14), and F1 score for all recursively forecasted maps across the three flood events in the testing dataset.

Training case	RMSE (m)			RMSE_ND (-)			F1 (-)		
	2011 event	2014 event	2020 event	2011 event	2014 event	2020 event	2011 event	2014 event	2020 event
Po1	0.21	0.48	0.12	0.064	0.127	0.031	0.972	0.974	0.985
Po2	0.19	0.15	0.16	0.060	0.036	0.041	0.973	0.988	0.982
Po3	0.38	0.31	0.38	0.112	0.073	0.099	0.963	0.979	0.955

charge is up to 13,000 m^3/s, as shown in Figure 5a). In addition, numerous defended floodplains are inundated only during flood events with inflow peak discharge exceeding 6,000−10,000 m^3/s. Consequently, as only few of the training flood scenarios exhibit a peak discharge and water volume sufficient to inundate the defended floodplains, the model struggles to learn the dynamics of floods in these regions due to the insufficient number of examples in the training dataset. Specifically, focusing on the 2014 flood event (Figure 9a), the surrogate model accurately predicts water depths at control points G3 and G6 as they are located in defended floodplains that are inundated for relatively low-intensity flood events (e.g., control point G6 is also flooded during the November 2011 event, as depicted in Figure 9b). In contrast, control points G4 and G5, in which the water depths present the larger error, are located in regions defended by higher levees and thus less frequently flooded.

3.2.2 *Po2* training case

To overcome the limitation stemming from the underrepresented sampling of high-intensity flood scenarios in the training dataset, the second FS training configuration (*Po2* training case in Table 2) expands the dataset by adding synthetically generated flood events. As detailed in Section 2.4.2, the synthetic events feature a peak discharge exceeding 8,000 m^3/s, resulting in the inundation of most defended floodplains. The water depths obtained from the newly trained model are depicted in Figure 9 (blue lines). Focusing on the 2014 flood event (Figure 9a), the use of a more balanced training dataset enhances the accuracy in forecasting the flooding dynamics in defended floodplains. Notably, training the surrogate model using the *Po2* configuration significantly improves the accuracy in forecasting the temporal variation of water depths in control points G4 and G5. The delay in the arrival time of the flood and the inability to predict flooding of the G5-floodplain are successfully addressed. However, at this control point, the water depths are still underestimated during the initial hours following the flood's arrival. In other control points, the accuracy of the surrogate model remains practically unchanged.

Analysing the two less severe scenarios in the testing dataset (i.e., the 2011 and 2020 events in Figure 9b), the surrogate model trained with the *Po2* configuration shows high accuracy in predicting water depths in both the main channel and in the floodplains. However, it tends to slightly overestimate the water depths in the main channel during the recession limb of the flood. This discrepancy may be attributed to the adoption of a more balanced dataset, which allows the model to better specialize in predicting flood events with high severity at the expense of low-intensity floods. Nevertheless, such errors in predicting the recession limb of low-severity flood events are deemed negligible for practical applications.

The increased accuracy of the model trained with the *Po2* configuration compared to *Po1* is further confirmed by the average RMSE computed for the recursively forecasted maps throughout the entire duration of the flood event. The results for each testing scenario are presented in Table 4. For the 2014 flood events, the RMSE for the *Po2* configuration is reduced by approximately 70% compared to the *Po1* case (i.e., from 0.48 m to 0.15 m). Conversely, the 2020 event is characterised by a 30% higher RMSE (i.e., from 0.12 m to 0.16 m) due to the overestimation of the water depths during the recession limb of the flood, as discussed previously. Still, errors of this magnitude can be considered acceptable in the practice.

In Figure 11, the predicted water depth maps for selected instants of the November 2014 flood event, generated by the FS model trained with the *Po2* configuration, are compared with ground-truth maps derived from the hydrodynamic model. The figure confirms the surrogate model's high accuracy in forecasting flood dynamics in both the main channel and open floodplains, where differences between predicted and target maps remain lower than 0.2 m for all predicted frames, representing an error of less than 5% of the average water depth for the entire flood scenario (see Table A2). Overall, more than 70% of the total number of flooded cells in all predicted maps (approximately 33M cells with a water depth higher than 0.05 m, the threshold used to filter out insignificant water depths in the numerical model's output maps, see Section 2.4) exhibit errors lower than 0.1 m. This percentage increases to 94% when considering errors between 0 and 0.2 m. Furthermore, only 0.5% of cells have errors higher than 1 m, mostly located in the defended floodplains (see Figure 11). These inaccuracies are primarily associated with a slight temporal shift in the inundation arrival time, as previously discussed in Figure 9a. Then, high errors diminish to under 0.2 m in subsequent frames. This consideration is also confirmed by the RMSE computed for each of the 174 predicted maps of the November 2014 flood event, represented in Figure 12. The value of the RMSE computed for the whole flooded domain is lower than 0.2 m during the initial and final stages of the event, while errors increase when defended floodplains begin to be flooded. This can be easily assessed by computing the RMSE for the main river region (i.e., main channel and open floodplains) and defended floodplains separately. As anticipated, the RMSE for the main channel remains below 0.2 m for the whole event. Differently, the metric for defended floodplains is higher, exhibiting various peaks corresponding to instants of flooding onset across different areas.

The average F1 score, computed across all forecasted frames for the three testing events is reported in Table 4. The extremely high F1 values confirm the accuracy of the surrogate model in predicting the temporal evolution of the flood extent.

3.2.3 *Po3* training case

For the last training configuration, named *Po3*, we used maps with a spatial resolution equal to 10 m. Additionally, the training dataset was created using the same flood events as the *Po2* configuration, which includes both real and synthetic flood scenarios (see Table 2). The aim of adopting a different resolution, and thus maps with a significantly larger number of cells, is to assess how the FS model scales for larger model dimension, both in term of accuracy of the results and computational times for training and forecasting. Comparing the results of the *Po3* and *Po2* configurations, which have the same training dataset but different spatial resolution of the maps, Table 4 indicates an increase in the average RMSE by $0.15-0.2$ m, accompanied by a

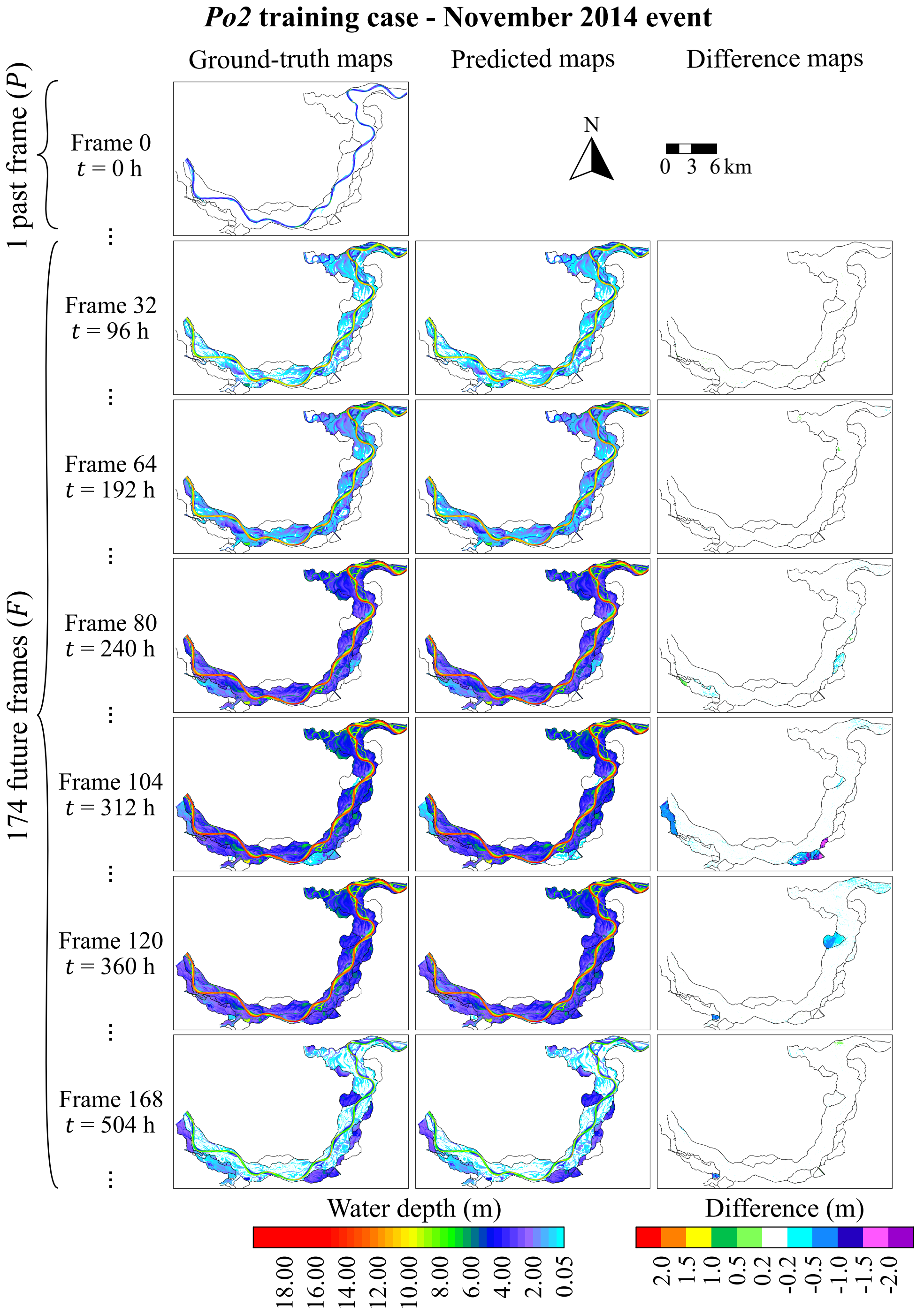

Figure 11. Po River case study: real-time forecasting of the November 2014 flood event using the FloodSformer model with the *Po2* training configuration. The columns represent, respectively, the ground-truth maps obtained from the hydrodynamic model, the maps predicted by the surrogate model, and the difference maps between the predicted and ground-truth maps. Only selected representative instants are shown.

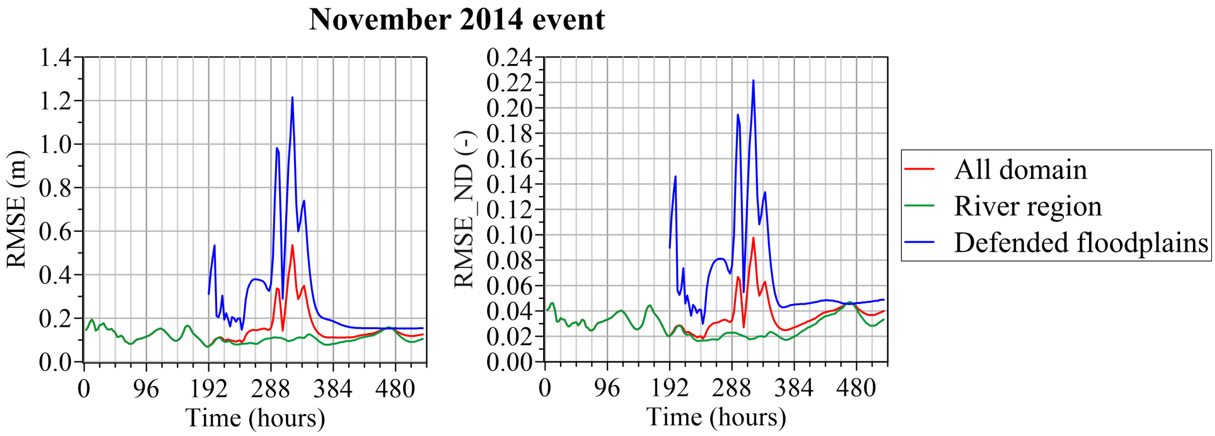

Figure 12. November 2014 flood event: RMSE and RMSE_ND (Eq. 14) computed for the maps forecasted by the surrogate model trained with the *Po2* configuration. RMSE values are also separately calculated for the main river region (including the main channel and open floodplains) and defended floodplains.

decrease in the F1 score due to the higher number of cells. Figure 9 reports the water depths extracted at control points for this configuration with green lines. Generally, the increase in the spatial resolution results in slightly higher discrepancies in some control points. For example, considering the 2014 flood event, the control point G2 exhibits an overestimation of water depths by approximately 0.3 m near the flood peak. Moreover, certain control points in defended floodplains (e.g., G4 and G5) show reduced prediction accuracy compared to the coarser resolution (i.e., *Po2* configuration). Similar considerations hold for the other flood scenarios of the testing dataset (Figure 9b).

This reduced accuracy of the autoregressive procedure may be attributed to the increased model dimension, which implies that, during the training process, the optimization task involves a significantly larger number of parameters compared to the *Po2* configuration. At the same time, the number of samples in the training dataset remains unchanged from the *Po2* training. This possibly leads to a reduced optimization of the *Po3* model, which could have been relieved by adopting a larger training dataset. Nevertheless, in this study, we aimed to compare results across various spatial resolutions using the same dataset, and further analyses are left to future works.

Target and predicted maps for the November 2014 flood event are illustrated in Figure S9 in the Supplement. Analogous to the results obtained with the *Po2* configurations (Figure 11), a good accuracy is achieved in the main channel and in the open floodplains while significant discrepancies primarily affect defended floodplains. Specifically, the surrogate model tends to anticipate flood arrival in certain floodplains, leading to the creation of significant differences at given times, due to sudden increases in water depths.

Despite the reduction in accuracy, predictions generated by the surrogate model trained with the *Po3* configuration maintain acceptable fidelity for real-time flood forecasting purposes. Furthermore, errors in the main river region are comparable to those expected from a physically based model.

3.3 Computational times

In this work, all simulations were performed using NVIDIA A100 GPUs.

For the Toce River case study, the computational time required for the FS model training process is approximately 2 hours when using one GPU. Predicting a flood event in the testing dataset with the autoregressive procedure takes about 10 s. This computational time is comparable to the runtime of the PARFLOOD code, which is very efficient for such a small case study.

For the Po River case study, the computational time of the FS model varies depending on the spatial resolution of the maps. Specifically, when maps with a resolution of 20 m are used, the training time of the surrogate model is approximately 50 hours using 2 GPUs. However, the computational time needed to recursively forecast 522 hours of the November 2014 event is approximately 3 minutes when employing one GPU. In comparison, the PARFLOOD code requires approximately 30 minutes to simulate the entire flood event using the same hardware configuration.

Differently, with a spatial resolution of 10 m, the overall training time of the FS model increases to approximately 71 hours using 4 GPUs. However, once trained, the model forecasts the entire November 2014 event in about 6.5 minutes using one GPU, whereas the PARFLOOD code takes approximately 140 minutes to simulate the entire flood event.

In conclusion, for the Po River case study, the ratio of physical time to the FS model's computational time ranges from 5,000 to 10,000. Furthermore, the surrogate model proves to be approximately $10-20$ times faster than the hydrodynamic model, depending on the spatial resolution adopted.

3.4 Benchmark comparison

In this Section, the performance of the FS model is compared with that of the 1D CNN model (Section 2.3) for both case studies, in order to assess the reliability of the proposed architecture. Specifically, we compared the accuracy of the surrogate models trained using the *Toce2* and the *Po2* training configurations (see Table 2) for the Toce River and Po River case studies, respectively.

For the 1D CNN model, a temporal window size corresponding to 8 timesteps (i.e., $r = 8$ in Section 2.3) was used, following the original implementation of the model (Kabir et al., 2020). Consequently, the input data for the model is a sequence of 9 inflow discharges from $t-8$ to t and the output is the inundation map at time t. This choice is also in line with the past time window used for the FS model ($I = 8$). The training of the convolutional model was conducted with a batch size of 10, the MSE loss function, the Adam optimizer, and a learning rate of 5e-4 and 1e-3 for the Toce River and Po River case studies, respectively. The values of the learning rate and batch size were determined through a trial-and-error process.

Table 5 presents a comparison of the average metrics of the two surrogate models. The FS model outperforms the 1D CNN architecture in terms of RMSEs and F1 scores for all the testing events considered. Focusing on the Toce River case, Figure 13 shows the inundation maps predicted by the convolutional model. The comparison between Figure 13 and Figure 7, which presents the results for the same flood event obtained with the FS model, highlights the higher accuracy of the latter model in predicting the inundation caused by the *Medium* flood event in the Toce River valley.

https://doi.org/10.5194/hess-2024-176
Preprint. Discussion started: 5 August 2024

Table 5. Comparison between the average metrics for the FS model and the 1D CNN model (Section 2.3). Results refer to the *Toce2* and *Po2* training configurations.

Testing event	RMSE (m)		RMSE_ND (-)		F1 (-)	
	FS (our)	1D CNN	FS (our)	1D CNN	FS (our)	1D CNN
Toce River						
Low	0.0018	0.0032	0.056	0.098	0.995	0.916
Medium	0.0026	0.0038	0.060	0.087	0.995	0.906
High	0.0032	0.0057	0.061	0.108	0.991	0.874
Gradual	0.0036	0.0064	0.083	0.157	0.991	0.868
Po River						
Nov 2011	0.19	0.36	0.060	0.120	0.973	0.966
Nov 2014	0.15	0.66	0.036	0.174	0.988	0.964
Jun 2020	0.16	0.66	0.041	0.181	0.982	0.908

Similarly, the FS model demonstrated superior performance in predicting flood events in the Po River region. Considering the most severe event in the testing dataset (i.e., the November 2014 flood), Figure 14 shows some water depth maps predicted by the 1D CNN model. Comparing these results with those obtained using the FS model (Figure 11), it is evident that the 1D CNN has lower accuracy in forecasting the spatiotemporal variation of water depths, especially in defended floodplains. This is mainly due to the inability of the 1D CNN model to account for spatiotemporal correlations between consecutive inundation maps, which is crucial for accurately reproducing complex flood dynamics. These findings are further confirmed by comparing the water depths extracted at selected control points along the Po River region, illustrated in Figure 15. Specifically, the convolutional model fails to accurately predict flood dynamics in most of the defended floodplains (e.g., at points G4 and G5, where results differ significantly from target data), while its accuracy is acceptable for points located in the main channel (i.e., G1 and G2), although still worse than that of FS model.

This benchmark analysis confirms the remarkable performance of the FS model compared to a state-of-the-art DL architecture. The adoption of a large transformer-based model, which considers both temporal and spatial information as input data to predict future frames, introduces a significant advantage compared to simpler models that analyze only upstream hydrograph inflows to predict the spatiotemporal propagation of inundation, as done by the 1D CNN model.

4 Discussion

Differently from the original implementation of the FloodSformer model by Pianforini et al. (2024a), the enhanced version introduced in this study replaces the SA mechanism in the VPTR block with the CA mechanism. This modification allows the

https://doi.org/10.5194/hess-2024-176
Preprint. Discussion started: 5 August 2024

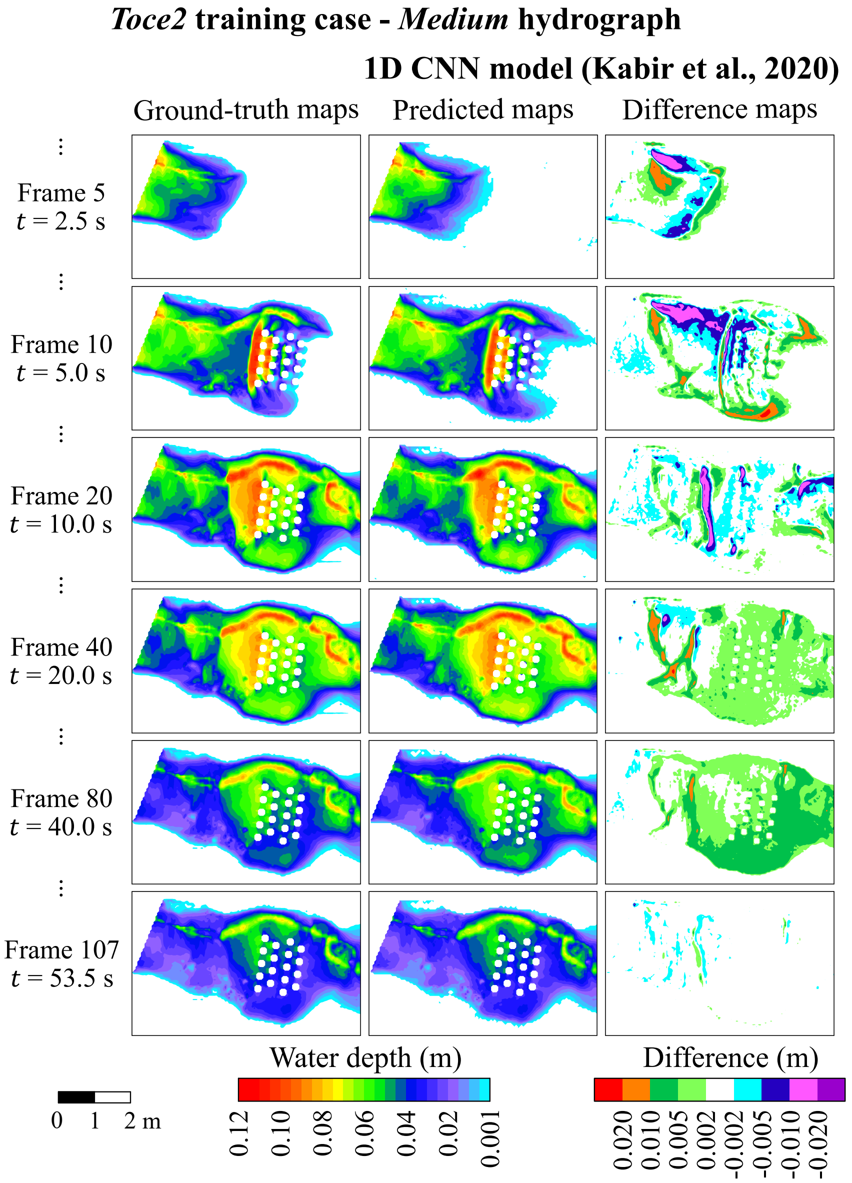

Figure 13. Toce River case study: real-time forecasting of the *Medium* hydrograph using the 1D CNN model (Section 2.3) trained with the *Toce2* configuration. The columns represent, respectively, the ground-truth maps obtained from the hydrodynamic model, the maps predicted by the CNN model, and the difference maps between predicted and ground-truth maps. Only selected representative instants are shown. This Figure can be compared with Figure 7, which shows the results for the same flood event obtained with the FS model.

https://doi.org/10.5194/hess-2024-176
Preprint. Discussion started: 5 August 2024

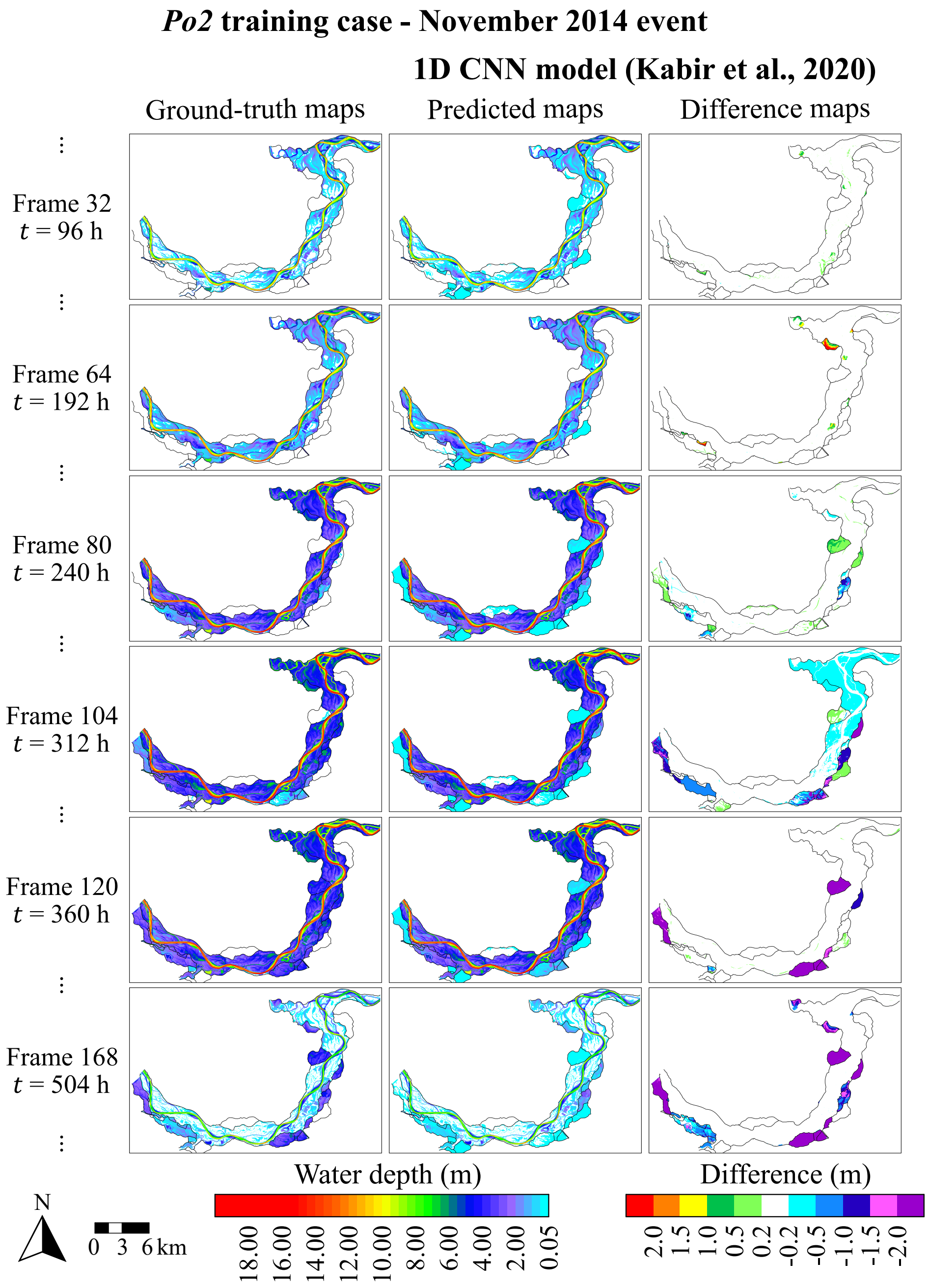

Figure 14. Po River case study: real-time forecasting of the November 2014 flood event using the 1D CNN model (Section 2.3) trained with the *Po2* configuration. The columns represent, respectively, the ground-truth maps obtained from the hydrodynamic model, the maps predicted by the CNN model, and the difference maps between the predicted and ground-truth maps. Only selected representative instants are shown. This Figure can be compared with Figure 11, which shows the results for the same flood event obtained with the FS model.

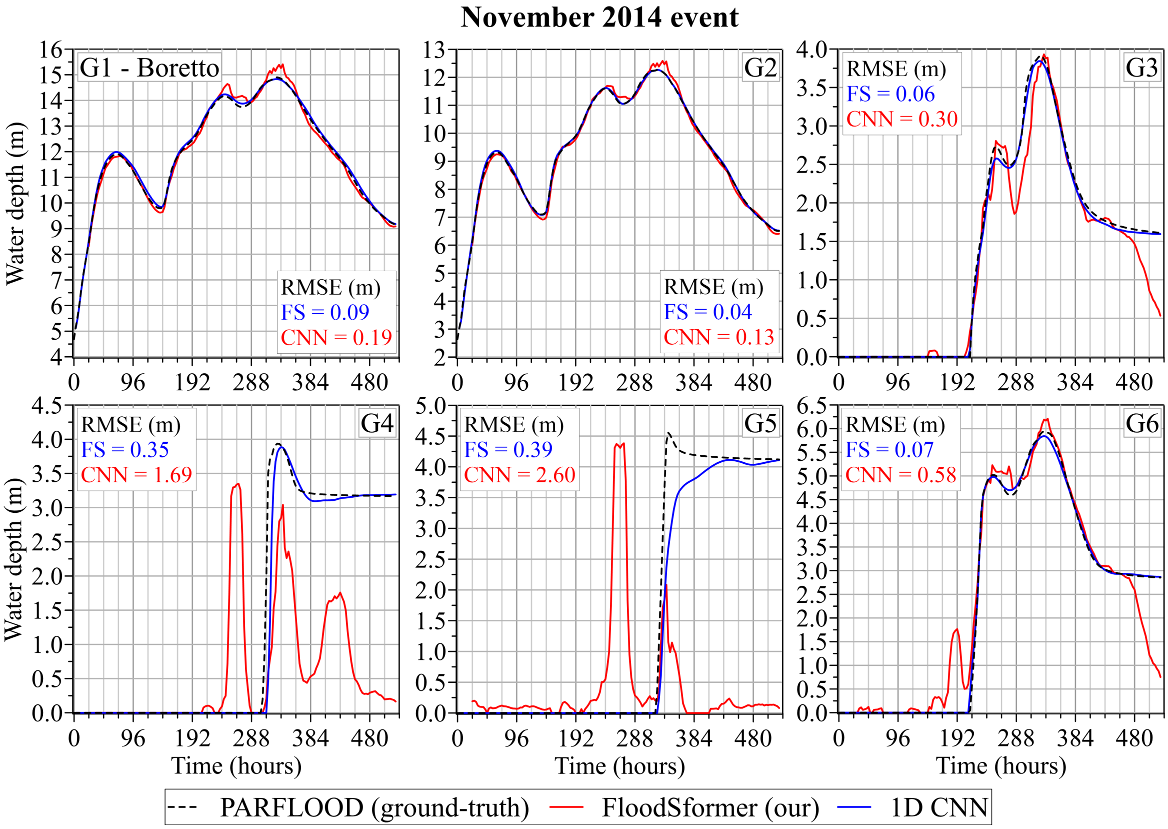

Figure 15. Po River case study: benchmark comparison between the simulated water depths at control points for the November 2014 flood event using the FloodSformer model and the 1D CNN model (Section 2.3), both trained considering the *Po2* configuration.

model to include an upstream boundary condition as additional input information, enabling the surrogate model to be used for real-time river flood forecasts.

Unlike other DL models that rely solely on information from upstream boundary conditions to predict inundation maps (e.g., Kabir et al., 2020), our model also integrates spatiotemporal information from maps of precedent instant. This comprehensive approach leverages both temporal and spatial data, allowing the model to capture the complex interactions between flow dynamics and topographical features, resulting in a higher prediction accuracy compared to other simpler surrogate models (see Section 3.4). This enhancement underscores the importance of considering spatiotemporal correlations in flood modeling to improve predictive performance and reliability.

In view of practical applications, the prediction of a flood event using the autoregressive procedure requires two types of input data. The first is the time series of upstream discharge throughout the flood's duration, typically obtained from meteorological and hydrological model chains. It is important to note that the lead time for predicting upstream inflow discharge may be shorter than the total duration of the flood event, especially for rivers with long flood propagation times, such as the Po River. Despite

https://doi.org/10.5194/hess-2024-176
Preprint. Discussion started: 5 August 2024

this, multiple subsequent run of the FS model's autoregressive procedure can be used for continuous updating of forecasts using newly predicted discharge data from the hydrological model.

The second required input data for the autoregressive procedure is the water depth map representing the initial condition of the flood scenario. A comprehensive analysis of the impact of initial conditions on the FS model predictions is provided in Section 4.1.

As already mentioned, the FS model employs an autoregressive procedure to forecast long sequences of future maps. Typically, such architectures are susceptible to error accumulation, which may limit the total number of maps that can be accurately predicted. However, as shown in Section 3, the FS model consistently produces accurate results even after forecasting tens of future frames. For instance, in the Po River case study, the predicted water depths remain unaffected by error accumulation, as can be seen from the results presented in Sections 3.2.2. We can thus assert that the forecasting lead time of the FS model is solely constrained by the availability of a sufficiently extended period of the inflow hydrograph, which serves as input information for the autoregressive prediction.

The generation of a dataset that accurately describes all the potential flood scenarios within the designated study area constitutes a critical phase in deploying DL architectures. Generally, data-driven models necessitate large datasets to effectively generalize and achieve optimal performance. The use of observed data describing inundation dynamics at suitable spatial and temporal resolutions is prevented by the unavailability of such information for actual flood events. For example, the adoption of satellite images is influenced by their low frequency of acquisition and their susceptibility to meteorological conditions (Bentivoglio et al., 2022). Conversely, numerical simulations can generate potentially limitless data, encompassing both real and synthetic events. However, the quality of the training and testing data is contingent upon the accuracy of the physically based model in correctly reproducing the flood dynamics. For this reason, a calibration process of the hydrodynamic model is essential to correctly set up a data-driven model. Accordingly, in this study, we employed the 2D SWE solver PARFLOOD code that was extensively validated for challenging case studies (e.g., Dazzi et al., 2021a). For each case study analyzed in the present work, the numerical model was calibrated using observed data. This procedure serves to validate the reliability of the simulated maps used to train the surrogate model.

Another aspect of the dataset generation that might influence the final accuracy of the data-driven model is the inclusion of flood events of different type (e.g. hydrographs with different steepness of the rising limb and/or spanning from low to high peak discharges) in the training dataset. As shown in Section 3.2, the use of an imbalanced training dataset could result in a drastic reduction of the fidelity of the surrogate model prediction. Furthermore, it is well known that data-driven models typically perform well at interpolating information within the range of training data. Conversely, accuracy diminishes when the model must extrapolate data beyond the training range (Fraehr et al., 2024). It is fundamental to ensure ample representation of flood events in the training dataset to guarantee optimal performance in real-word applications. This entails, for example, that high-intensity flood scenarios must be included in the training dataset to allow the surrogate model to appropriately learn the dynamics associated with extreme events.

A drawback associated with the proposed surrogate model is its sensitivity to the adopted spatial resolution, as evidenced by the increase in RMSE of the predicted maps with higher resolutions (see Section 3.2.3). While higher spatial resolutions provide more detailed information, they necessitate the use of a more complex surrogate model with an increased number of parameters. Such models are more challenging to train and require larger datasets. Therefore, achieving an optimal balance between spatial resolution and model complexity is crucial to ensure the reliability and robustness of the surrogate model for real-time river flood forecasting applications. Nonetheless, the results obtained for the Po River case study, trained at the finest resolution, remain acceptable for the purpose of real-time river flood forecasting.

The most significant advantage of the proposed data-driven architecture lies in its ability to drastically reduce the computational time compared to the physically based models. For instance, in the Po case study, the autoregressive procedure of the FS model enables the forecasting of a flood event lasting about 3 weeks in just a few minutes, achieving ratios of physical to computational time up to 10,000. Furthermore, the surrogate model exhibits a speedup of about 10−20 times compared to the hydrodynamic model. This speedup depends on the spatial resolution adopted and escalates with the increase in the number of cells adopted to discretize the domain. It is worth noting that the numerical model employed in this study, namely the PARFLOOD code, was efficiently implemented to leverage the capabilities of GPU architectures. Consequently, its computational time is already significantly reduced compared to other serial codes (Vacondio et al., 2014). If the surrogate model was to replace a less efficient code, its advantage in terms of computational time would be even more evident. The good computational performance of the FS model promotes its employment for real-time forecasting of floods.

4.1 Initial condition sensitivity analysis

In Section 3, we showed that the FS model is able to forecast flood scenarios using only one past frame (representing the initial condition) to start the autoregressive procedure. This map is derived from a steady state simulation performed with the hydrodynamic model. However, in real-time flood forecasting, the necessity of conducting a numerical simulation of a steady flow at the start of the flood event can pose a computational bottleneck. Therefore, in this Section, we analyze the influence of initial conditions on the results of flood predictions obtained with the FS model.

For this sensitivity analysis, we considered the November 2014 flood event (see Figure 5c) and the FS model trained with the *Po2* configuration (see Table 2). The “real” initial condition was obtained considering a steady flow with a discharge of 500 m^3/s. This value is similar to the initial value of the November 2014 hydrograph. Additionally, two distinct steady flow conditions with discharge values of 1,500 m^3/s and 2,500 m^3/s were simulated. The maps representing water depths for the three steady flows considered were used as past frames for the FS model predictions.

Figure 16a shows the comparison between water depths extracted at the Boretto gauge station (G1) for the different configurations examined. Remarkably, the surrogate model tends to disregard information about the initial condition after a few time steps. For instance, the forecasted water depths tend to align about 24 hours after the event’s onset regardless of the adopted initial condition. Similar results are observed for other cells in the study area (not shown). This consideration is supported by the analysis of RMSEs computed on each predicted map, reported in Figure 16b. Specifically, RMSEs differ for the very first frames depending on the initial conditions, then gradually converge to identical values for the rest of the event. The convergence is more rapid as the assumed initial frame gets closer to the “actual” initial condition.

https://doi.org/10.5194/hess-2024-176
Preprint. Discussion started: 5 August 2024

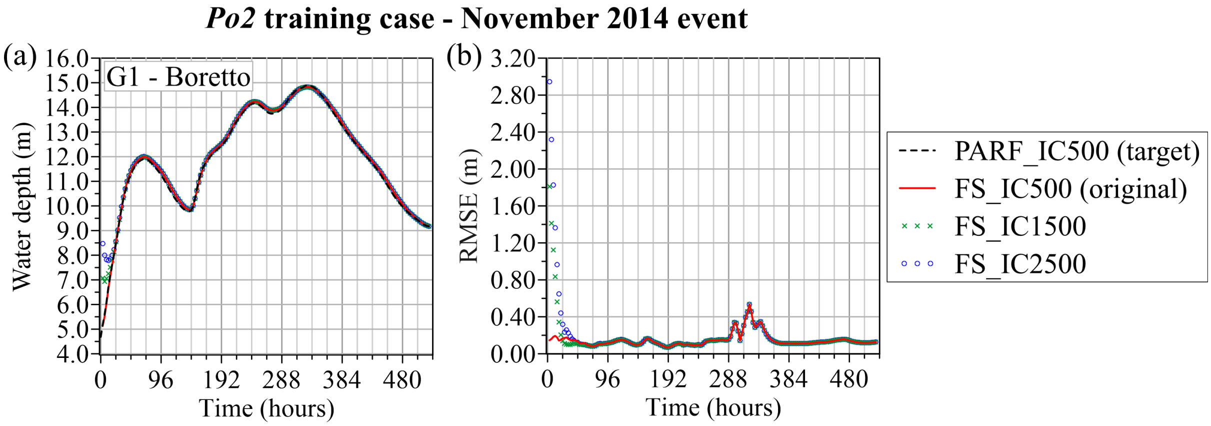

Figure 16. Comparison of results for different initial conditions. The surrogate model forecasts consider different past frame maps. **(a)** Comparison of water depths extracted at the Boretto gauge station. **(b)** RMSE computed on the forecasted maps with the three initial conditions analyzed.

These findings confirm that the prediction of the FS model is independent of the initial condition considered except for a relatively short warmup. Consequently, for practical applications, it is advisable to create a small database containing water depth maps for some steady flow conditions that can serve as initial conditions (past frame) for the autoregressive prediction of the FS model. The map corresponding to the initial discharge value closest to the real-event conditions can then be selected. This procedure enables the use of the FS model for flood prediction avoiding any preliminary numerical simulation. The number and discharge values to be considered depend on the case study. For example, for the Po case study, discharge values in the order of $500 - 3{,}000\ \mathrm{m}^3/\mathrm{s}$ can be adopted for the creation of the database.

5 Conclusions

In this study, the FloodSformer model, a transformer-based data-driven model originally proposed for real-time forecasting of dam-break scenarios (Pianforini et al., 2024a), has been modified to predict river flood inundations with a negligible computational time.

The results demonstrate the FS model's capability to accurately forecast the spatial and temporal evolution of water depths in river floods, relying solely on an initial water depth map and on the hydrograph describing the inflow discharge for the entire event duration, which can be obtained from meteorological/hydrological models. Prediction errors generally align with the uncertainty observed in physically based models. For example, in the Po case study, the average RMSE is lower than 20 cm. Overall, more than 90% of flooded cells exhibit errors lower than 20 cm. Furthermore, the autoregressive procedure ensures acceptable prediction accuracy even after forecasting tens of maps, promoting the prediction of long-lasting flood events. The performance of the proposed model was also compared against a state-of-the-art 1D CNN model, demonstrating superior accuracy in forecasting flood events across all case studies analyzed.

The FS forecasts remain independent of the water depth map used as initial condition for the autoregressive procedure. This finding promotes FS model's adoption for real-time forecasting, eliminating the need for preliminary numerical simulations to generate exact initial conditions.

Finally, the FS model exhibits remarkable computational efficiency in predicting real flood events. For instance, for a Po River flood scenario lasting approximately 3 weeks, the surrogate model requires only a few minutes to forecast all the water depth maps with a temporal resolution of 3 hours. This corresponds to a ratio of physical to computational time up to 10,000. Furthermore, the surrogate model is 10 to 20 times faster than the hydrodynamic model, although the latter was efficiently implemented to run in parallel on GPU. Consequently, the short computational time of the FS model's autoregressive procedure further emphasizes the advantage of the proposed data-driven approach for real-time flood forecasting. This efficiency not only streamlines the forecasting process but also enhances the model's responsiveness to dynamic flood conditions, ultimately contributing to more effective and timely decision-making in flood management and mitigation efforts.

Appendix A: FloodSformer training results

This Appendix presents the outcomes of the FloodSformer model training. Upon completing the training procedure, the model's accuracy in predicting the map at time step $I+1$ is evaluated using input data sequences of I consecutive maps from the testing dataset. This evaluation procedure, named "FS test", ensures the surrogate model's proper training and suitability for autoregressive forecasting of extended sequences of inundation maps. Table A1 provides a summary of the average RMSEs and F1 scores of the FS tests for both the Toce and Po River case studies. It is noteworthy that these case studies differ in spatial and temporal scales, thereby influencing the resulting metrics.

For the Toce River case, the RMSE and F1 score exhibit minimal variations across different training configurations. The RMSE, approximately 1 mm, is less than $2-3\%$ of the average water depth across all testing scenarios (see Table A2). Moreover, an F1 score close to 1 confirms the FS model's high accuracy in predicting the temporal variation of flood extent one time step ahead.

In the Po River case study, an increase in spatial resolution leads to a higher RMSE. This increase primarily stems from the larger number of surrogate model's parameters, while the dimension of the training dataset remains unchanged (see Section 3.2.3 for more details). Nonetheless, the F1 score remains notably high for this case study.

In summary, the limited errors observed in the FS test validate the application of the surrogate model for predicting long sequences of inundation maps using the autoregressive procedure, as shown in Section 3.

Code and data availability. The dataset employed in this study, along with the trained weights of the FS model, can be accessed at Pianforini et al. (2024b). The Python code repository is accessible at Pianforini et al. (2024c).

Table A1. Average RMSEs of the FS test for the Toce and Po case studies.

	Toce River		Po River		
	Toce1	*Toce2*	*Po1*	*Po2*	*Po3*
RMSE (m)	0.0011	0.0009	0.070	0.068	0.117
F1 (-)	0.995	0.996	0.988	0.989	0.987

Table A2. Average water depths ($\bar{h}$) for the testing scenarios of the Toce and Po case studies.

	Toce River				Po River		
	Low	*Medium*	*High*	*Gradual*	2011 event	2014 event	2020 event
$\bar{h}$ (m)	0.034	0.040	0.052	0.041	3.45	4.06	4.16

Author contributions. Conceptualization, MP, SD, and RV; methodology, MP, SD, AP, and RV; software, MP, and AP; formal analysis, MP; data curation, MP; writing - original draft preparation, MP; writing - review & editing, MP, SD, AP, and RV; visualization, MP; supervision, RV; funding acquisition, SD, and RV.

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

Acknowledgements. This research was granted by University of Parma through the action “Bando di Ateneo 2022 per la ricerca” co-funded by MUR-Italian Ministry of Universities and Research - D.M. 737/2021 - PNR - PNRR – NextGenerationEU. RV and SD acknowledge financial support from the PNRR MUR project ECS_00000033_ECOSISTER. This research also benefits from the HPC facility of the University of Parma. Finally, the Authors acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support (project AMNERIS).

References

Bentivoglio, R., Isufi, E., Jonkman, S. N., and Taormina, R.: Deep learning methods for flood mapping: a review of existing applications and future research directions, https://doi.org/10.5194/hess-26-4345-2022, 2022.

Bentivoglio, R., Isufi, E., Jonkman, S. N., and Taormina, R.: Rapid spatio-temporal flood modelling via hydraulics-based graph neural networks, Hydrology and Earth System Sciences, 27, 4227–4246, https://doi.org/10.5194/hess-27-4227-2023, 2023.

Bertasius, G., Wang, H., and Torresani, L.: Is space-time attention all you need for video understanding?, in: ICML, pp. 813–824, 2021.

Bomers, A.: Predicting Outflow Hydrographs of Potential Dike Breaches in a Bifurcating River System Using NARX Neural Networks, Hydrology, 8, 87, https://doi.org/10.3390/hydrology8020087, 2021.

Bomers, A. and Hulscher, S. J. M. H.: Neural networks for fast fluvial flood predictions: Too good to be true?, River Research and Applications, 39, 1652–1658, https://doi.org/10.1002/RRA.4144, 2023.

Burrichter, B., Hofmann, J., da Silva, J. K., Niemann, A., and Quirmbach, M.: A Spatiotemporal Deep Learning Approach for Urban Pluvial Flood Forecasting with Multi-Source Data, Water, 15, 1760, https://doi.org/10.3390/W15091760, 2023.

Burrichter, B., da Silva, J. K., Niemann, A., and Quirmbach, M.: A Temporal Fusion Transformer Model to Forecast Overflow from Sewer Manholes during Pluvial Flash Flood Events, Hydrology, 11, 41, https://doi.org/10.3390/HYDROLOGY11030041, 2024.

Campolo, M., Andreussi, P., and Soldati, A.: River flood forecasting with a neural network model, Water Resources Research, 35, 1191–1197, https://doi.org/https://doi.org/10.1029/1998WR900086, 1999.

Castangia, M., Grajales, L. M. M., Aliberti, A., Rossi, C., Macii, A., Macii, E., and Patti, E.: Transformer neural networks for interpretable flood forecasting, Environmental Modelling and Software, 160, 105 581, https://doi.org/10.1016/j.envsoft.2022.105581, 2023.

Chaudhary, P., Leitão, J. P., Schindler, K., and Wegner, J. D.: Flood Water Depth Prediction with Convolutional Temporal Attention Networks, Water, 16, https://doi.org/10.3390/w16091286, 2024.

Costabile, P., Costanzo, C., and Macchione, F.: Performances and limitations of the diffusive approximation of the 2-d shallow water equations for flood simulation in urban and rural areas, Applied Numerical Mathematics, 116, 141–156, https://doi.org/https://doi.org/10.1016/j.apnum.2016.07.003, new Trends in Numerical Analysis: Theory, Methods, Algorithms and Applications (NETNA 2015), 2017.

CRED: 2023 Disasters in Numbers: A Significant Year of Disaster Impact, Tech. rep., Centre for Research on the Epidemiology of Disasters (CRED), 2024.

Dazzi, S., Shustikova, I., Domeneghetti, A., Castellarin, A., and Vacondio, R.: Comparison of two modelling strategies for 2D large-scale flood simulations, Environmental Modelling and Software, 146, 105 225, https://doi.org/10.1016/j.envsoft.2021.105225, 2021a.

Dazzi, S., Vacondio, R., and Mignosa, P.: Flood Stage Forecasting Using Machine-Learning Methods: A Case Study on the Parma River (Italy), Water (Switzerland), 13, 1612, https://doi.org/10.3390/w13121612, 2021b.

Dazzi, S., Vacondio, R., Mignosa, P., and Aureli, F.: Assessment of pre-simulated scenarios as a non-structural measure for flood management in case of levee-breach inundations, International Journal of Disaster Risk Reduction, 74, 102 926, https://doi.org/10.1016/j.ijdrr.2022.102926, 2022.

Donnelly, J., Abolfathi, S., Pearson, J., Chatrabgoun, O., and Daneshkhah, A.: Gaussian process emulation of spatio-temporal outputs of a 2D inland flood model, Water Research, 225, 119 100, https://doi.org/10.1016/j.watres.2022.119100, 2022.

Ferrari, A., Viero, D. P., Vacondio, R., Defina, A., and Mignosa, P.: Flood inundation modeling in urbanized areas: A mesh-independent porosity approach with anisotropic friction, Advances in Water Resources, 125, 98–113, https://doi.org/10.1016/j.advwatres.2019.01.010, 2019.

Ferrari, A., Dazzi, S., Vacondio, R., and Mignosa, P.: Enhancing the resilience to flooding induced by levee breaches in lowland areas: A methodology based on numerical modelling, Natural Hazards and Earth System Sciences, 20, 59–72, https://doi.org/10.5194/nhess-20-59-2020, 2020.

Ferrari, A., Vacondio, R., and Mignosa, P.: High-resolution 2D shallow water modelling of dam failure floods for emergency action plans, Journal of Hydrology, 618, 129 192, https://doi.org/10.1016/j.jhydrol.2023.129192, 2023.

Fraehr, N., Wang, Q. J., Wu, W., and Nathan, R.: Development of a Fast and Accurate Hybrid Model for Floodplain Inundation Simulations, Water Resources Research, 59, e2022WR033 836, https://doi.org/10.1029/2022wr033836, 2023.

Fraehr, N., Wang, Q. J., Wu, W., and Nathan, R.: Assessment of surrogate models for flood inundation: The physics-guided LSG model vs. state-of-the-art machine learning models, Water Research, 252, 121 202, https://doi.org/10.1016/j.watres.2024.121202, 2024.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep learning, MIT press, 2016.

Hop, F. J., Linneman, R., Schnitzler, B., Bomers, A., and Booij, M. J.: Real time probabilistic inundation forecasts using a LSTM neural network, Journal of Hydrology, 635, 131 082, https://doi.org/10.1016/J.JHYDROL.2024.131082, 2024.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A.: Image-To-Image Translation With Conditional Adversarial Networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1125–1134, 2017.

Jin, H., Lu, H., Zhao, Y., Zhu, Z., Yan, W., Yang, Q., and Zhang, S.: Integration of an improved transformer with physical models for the spatiotemporal simulation of urban flooding depths, Journal of Hydrology: Regional Studies, 51, 101 627, https://doi.org/10.1016/J.EJRH.2023.101627, 2024.

Kabir, S., Patidar, S., Xia, X., Liang, Q., Neal, J., and Pender, G.: A deep convolutional neural network model for rapid prediction of fluvial flood inundation, Journal of Hydrology, 590, 125 481, https://doi.org/10.1016/j.jhydrol.2020.125481, 2020.

Kao, I. F., Liou, J. Y., Lee, M. H., and Chang, F. J.: Fusing stacked autoencoder and long short-term memory for regional multistep-ahead flood inundation forecasts, Journal of Hydrology, 598, 126 371, https://doi.org/10.1016/j.jhydrol.2021.126371, 2021.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrology and Earth System Sciences, 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

Li, W., Liu, C., Hu, C., Niu, C., Li, R., Li, M., Xu, Y., and Tian, L.: Application of a hybrid algorithm of LSTM and Transformer based on random search optimization for improving rainfall-runoff simulation, Scientific Reports, 14, 11 184, 2024.

Liu, C., Liu, D., and Mu, L.: Improved Transformer Model for Enhanced Monthly Streamflow Predictions of the Yangtze River, IEEE Access, 10, 58 240–58 253, https://doi.org/10.1109/ACCESS.2022.3178521, 2022.

Morales-Hernández, M., Sharif, M. B., Kalyanapu, A., Ghafoor, S., Dullo, T., Gangrade, S., Kao, S.-C., Norman, M., and Evans, K.: TRITON: A Multi-GPU open source 2D hydrodynamic flood model, Environmental Modelling & Software, 141, 105 034, https://doi.org/https://doi.org/10.1016/j.envsoft.2021.105034, 2021.

Mosavi, A., Ozturk, P., and Chau, K. W.: Flood prediction using machine learning models: Literature review, https://doi.org/10.3390/w10111536, 2018.

Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., Begelman, G., Nearing, G., Shalev, G., Noga, H., Shavitt, I., Yuklea, L., Royz, M., Giladi, N., Peled Levi, N., Reich, O., Gilon, O., Maor, R., Timnat, S., Shechter, T., Anisimov, V., Gigi, Y., Levin, Y., Moshe, Z., Ben-Haim, Z., Hassidim, A., and Matias, Y.: Flood forecasting with machine learning models in an operational framework, Hydrology and Earth System Sciences, 26, 4013–4032, https://doi.org/10.5194/hess-26-4013-2022, 2022.

Pappenberger, F., Cloke, H. L., Parker, D. J., Wetterhall, F., Richardson, D. S., and Thielen, J.: The monetary benefit of early flood warnings in Europe, Environmental Science & Policy, 51, 278–291, https://doi.org/https://doi.org/10.1016/j.envsci.2015.04.016, 2015.

Pianforini, M., Dazzi, S., Pilzer, A., and Vacondio, R.: Real-time flood maps forecasting for dam-break scenarios with a transformer-based deep learning model, Journal of Hydrology, 635, 131 169, https://doi.org/10.1016/J.JHYDROL.2024.131169, 2024a.

Pianforini, M., Dazzi, S., Pilzer, A., and Vacondio, R.: FloodSformer: River Flood datasets&checkpoints, https://doi.org/10.5281/zenodo.11472228, 2024b.

Pianforini, M., Dazzi, S., Pilzer, A., and Vacondio, R.: FloodSformer: Python code, https://doi.org/10.5281/zenodo.10895200, 2024c.

Rogers, D. and Tsirkunov, V.: Costs and Benefits of Early Warning Systems: Global Assessment Report on Disaster Risk Reduction, Tech. rep., The World Bank, 2011.

Tayfur, G., Singh, V. P., Moramarco, T., and Barbetta, S.: Flood Hydrograph Prediction Using Machine Learning Methods, Water, 10, https://doi.org/10.3390/w10080968, 2018.

Testa, G., Zuccalà, D., Alcrudo, F., Mulet, J., and Soares-Frazão, S.: Flash flood flow experiment in a simplified urban district, Journal of Hydraulic Research, 45, 37–44, https://doi.org/10.1080/00221686.2007.9521831, 2007.

Turchetto, M., Dal Palù, A., and Vacondio, R.: A General Design for a Scalable MPI-GPU Multi-Resolution 2D Numerical Solver, IEEE Transactions on Parallel and Distributed Systems, 31, 1036–1047, https://doi.org/10.1109/TPDS.2019.2961909, 2020.

Vacondio, R., Dal Palù, A., and Mignosa, P.: GPU-enhanced finite volume shallow water solver for fast flood simulations, Environmental Modelling and Software, 57, 60–75, https://doi.org/10.1016/j.envsoft.2014.02.003, 2014.

Vacondio, R., Dal Palù, A., Ferrari, A., Mignosa, P., Aureli, F., and Dazzi, S.: A non-uniform efficient grid type for GPU-parallel Shallow Water Equations models, Environmental Modelling and Software, 88, 119–137, https://doi.org/10.1016/j.envsoft.2016.11.012, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention is All you Need, in: Advances in Neural Information Processing Systems, 2017.

Wallemacq, P. and House, R.: Economic losses, poverty & disasters: 1998-2017, Tech. rep., Centre for Research on the Epidemiology of Disasters (CRED), 2018.

Wang, J.-H., Lin, G.-F., Chang, M.-J., Huang, I.-H., and Chen, Y.-R.: Real-Time Water-Level Forecasting Using Dilated Causal Convolutional Neural Networks, Water resources management, 33, 3759–3780, https://doi.org/https://doi.org/10.1007/s11269-019-02342-4, 2019.

Wei, G., Xia, W., He, B., and Shoemaker, C.: Quick large-scale spatiotemporal flood inundation computation using integrated Encoder-Decoder LSTM with time distributed spatial output models, Journal of Hydrology, 634, 130 993, https://doi.org/10.1016/J.JHYDROL.2024.130993, 2024.

Xia, X., Liang, Q., Ming, X., and Hou, J.: An efficient and stable hydrodynamic model with novel source term discretization schemes for overland flow and flood simulations, Water Resources Research, 53, 3730–3759, https://doi.org/https://doi.org/10.1002/2016WR020055, 2017.

Xia, X., Liang, Q., and Ming, X.: A full-scale fluvial flood modelling framework based on a high-performance integrated hydrodynamic modelling system (HiPIMS), Advances in Water Resources, 132, 103 392, https://doi.org/https://doi.org/10.1016/j.advwatres.2019.103392, 2019.

Xu, Y., Lin, K., Hu, C., Wang, S., Wu, Q., Zhang, L., and Ran, G.: Deep transfer learning based on transformer for flood forecasting in data-sparse basins, Journal of Hydrology, 625, 129 956, https://doi.org/https://doi.org/10.1016/j.jhydrol.2023.129956, 2023.

Ye, X. and Bilodeau, G. A.: Video prediction by efficient transformers, Image and Vision Computing, 130, https://doi.org/10.1016/j.imavis.2022.104612, 2023.

Yin, H., Guo, Z., Zhang, X., Chen, J., and Zhang, Y.: RR-Former: Rainfall-runoff modeling based on Transformer, Journal of Hydrology, 609, 127 781, https://doi.org/10.1016/j.jhydrol.2022.127781, 2022.

Yin, H., Zhu, W., Zhang, X., Xing, Y., Xia, R., Liu, J., and Zhang, Y.: Runoff predictions in new-gauged basins using two transformer-based models, Journal of Hydrology, 622, 129 684, https://doi.org/10.1016/j.jhydrol.2023.129684, 2023.

Zhou, Y., Wu, W., Nathan, R., and Wang, Q. J.: Deep Learning-Based Rapid Flood Inundation Modeling for Flat Floodplains With Complex Flow Paths, Water Resources Research, 58, e2022WR033 214, https://doi.org/10.1029/2022WR033214, 2022.