Earth Systems and Environment (2026) 10:5781–5802
https://doi.org/10.1007/s41748-025-00923-5

ORIGINAL ARTICLE

Enhancing Flood Forecasting with Machine Learning Informed by Integrated ParFlow-CLM Hydrological Modeling

Samira Sadat Soltani[1,2] · Zohreh Mohammadi[3] · Ardalan Izadi[3] · Stefan Kollet[1,2]

Received: 17 May 2025 / Revised: 23 September 2025 / Accepted: 19 October 2025 / Published online: 18 November 2025

Abstract
Reliable hydrologic modeling is essential for sustainable water resource management, accurate projections of future water availability, and the timely forecasting of extreme weather events. The severe flooding that occurred in July 2021, however, highlighted significant limitations in current predictive models, particularly their inadequacy in capturing unprecedented events. These limitations are largely attributed to factors such as sparse observational data, simulation uncertainties, and challenges in model calibration. Another critical factor contributing to flood events is the accumulation of soil moisture following prolonged rainfall, which often leads to soil saturation. Despite its importance, many existing flood forecasting systems inadequately represent soil moisture processes and the complex interactions between surface and subsurface hydrological components. In response to these shortcomings, this study explores the integration of machine learning techniques with hydrologic models to enhance the accuracy of flood predictions. Specifically, we explore the impact of incorporating soil water content (SWC), derived from The integrated surface-subsurface ParFlow/CLM model, into an artificial intelligence (AI) framework for flood forecasting. We developed a Gated Recurrent Unit (GRU) Convolutional model, which initially used eight atmospheric variables (including longwave and shortwave radiation, total precipitation, air temperature, air pressure, specific humidity, wind velocity components) to predict river discharge. In the subsequent stage, SWC was added as an additional input to evaluate its effect on the model's performance. Our comparative analyses revealed that including SWC significantly improved the predictive accuracy of the AI model, highlighting its essential role in discharge estimation. Furthermore, integrating the ParFlow model's outputs with machine learning techniques resulted in a substantial enhancement in flood prediction, outperforming both standalone AI and hydrological models. The AI model with SWC achieved an RMSE reduction of up to 51% and an increase in R^2 from 0.634 to 0.916 at Kordel-Kyll. The findings demonstrate the potential of informing machine learning with integrated hydrologic model output to improve flood prediction systems, offering valuable insights for more effective water resource management.

Graphical Abstract This graphical abstract presents an integrated modeling framework that combines a fully coupled subsurface-surface hydrological model, ParFlow/CLM, with a machine learning algorithm for river discharge prediction. The system was applied to a real flood event in Germany, demonstrating its practical relevance. Atmospheric forcings, including precipitation (APCP), temperature (Temp), specific humidity (SPFH), pressure (Press), wind components (UGRD, VGRD), and radiation (DLWR, DSWR), are collected along with high-resolution soil water content (SWC) data and observed river discharge. The detailed SWC data capture near-saturated soil conditions, which contributed to flooding despite only moderate, but sustained, rainfall. ParFlow/CLM simulates key land surface processes such as snow accumulation, overland flow, and evapotranspiration. Outputs from ParFlow/CLM and meteorological inputs are used as features in a hybrid machine learning model, specifically, a Gated Recurrent Unit (GRU)-based Convolutional Neural Network (CNN), which captures

✉ Samira Sadat Soltani
s.soltani@fz-juelich.de

Zohreh Mohammadi
zohrehmohammadi@alumni.kntu.ac.ir

Ardalan Izadi
a.izadi@kntu.ac.ir

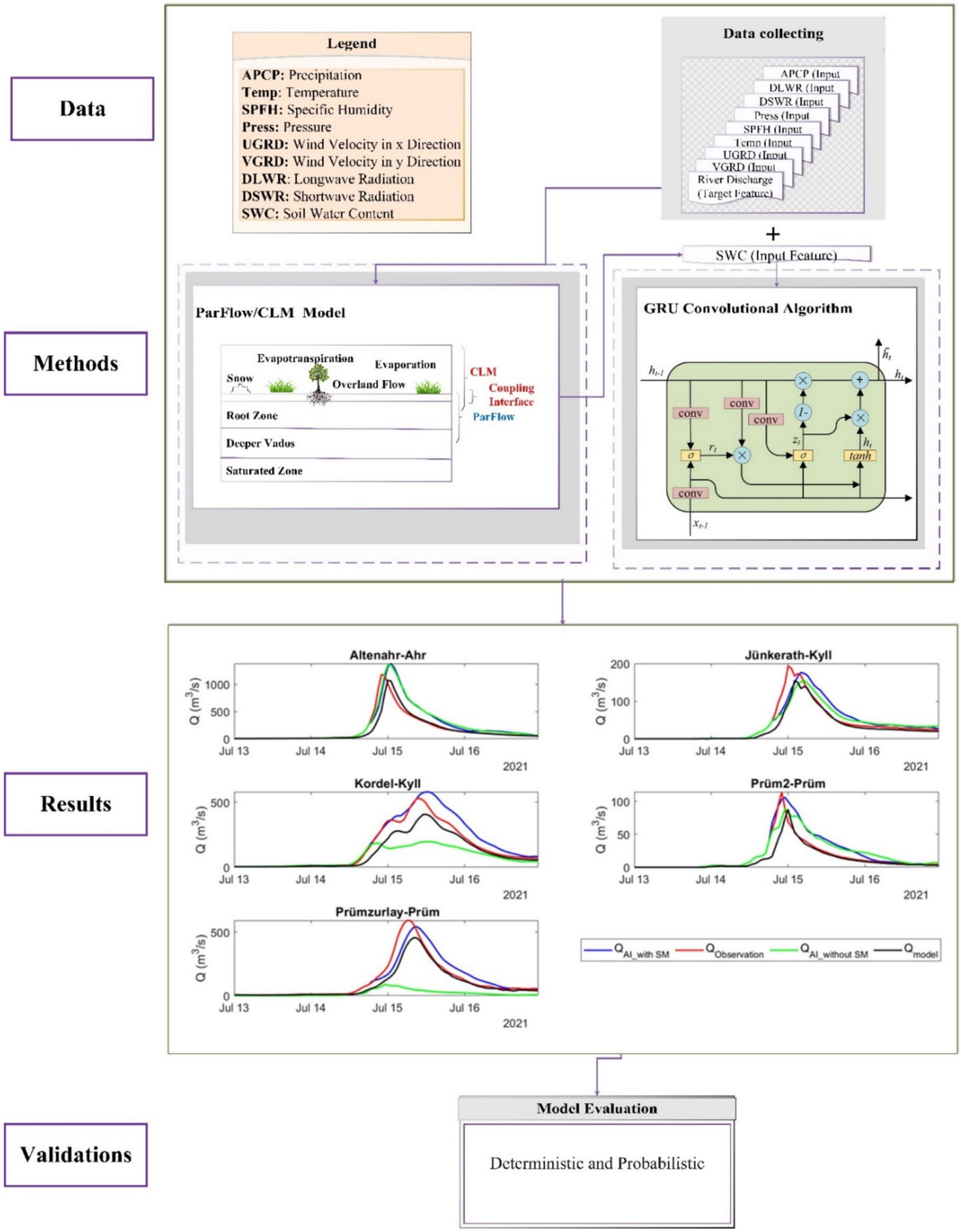
Stefan Kollet
s.kollet@fz-juelich.de

1 Forschungszentrum Jülich GmbH, Institute of Bio- and Geosciences (IBG-3, Agrosphäre), 52425 Jülich, Germany

2 Centre for High-Performance Scientific Computing in Terrestrial Systems, Geoverbund ABC/J, Jülich, Germany

3 Faculty of Civil Engineering, K. N. Toosi University of Technology, Tehran, Iran

temporal dynamics for improved discharge forecasting. A coupling interface links the physical model with the machine learning component. Simulations correspond to observed flood periods and are evaluated using both deterministic and probabilistic metrics. Incorporating hydrological models with machine learning, especially with soil moisture as an input, improves both flood timing and peak prediction. This approach supports early warning systems, where soil moisture forecasts can be used alongside meteorological predictions to enhance model performance, particularly in areas with flatter terrain, where soil moisture plays a more significant role than in steep regions.



Highlights

- GRU model trained on eight key atmospheric variables for discharge prediction.
- Soil Water Content (SWC) from ParFlow/CLM added as a new input feature.
- SWC inclusion significantly boosts flood prediction accuracy.
- Machine learning boosts hydrological modeling in capturing extreme flood events.

Keywords Machine learning · Hydrological modeling · Flood forecasting · Soil water content · Gated Recurrent Unit (GRU) convolutional

1 Introduction

The growing occurrence of intense summer rainfall events is a key manifestation of a changing climate. In recent years, these changes have significantly affected groundwater resources and surface water systems (Badawy et al. 2024), leading to pronounced changes in hydrologic fluxes. This has intensified natural disasters such as floods and droughts (Marshall et al., 2025), disrupted the operations of dams and reservoirs (Tran et al. 2023), and adversely impacted agricultural practices, water quality (Tapas et al. 2024), and sustainable livelihoods (Nguyen et al. 2024), underscoring the urgent need for more reliable predictive models. A striking example is the devastating flood that struck Germany in July 2021, which led to more than 180 deaths and caused widespread damage to residential areas, transportation networks, and industrial facilities. As reported by Munich Re (2022), the disaster inflicted damages estimated at EUR 33 billion, with total losses reaching EUR 46 billion, marking it as the most economically costly natural catastrophe in Germany's history.

Soil water content (SWC) plays a vital role in regulating the exchange of water, energy, and carbon between the land surface and the atmosphere, and it is a key factor in the process of runoff generation. The level of soil moisture directly affects the magnitude of flood events by influencing how rainfall is distributed between surface runoff and infiltration into the subsurface (Brocca et al. 2009; Beck et al. 2009; Tramblay et al. 2012). Hydrological models are frequently employed to estimate this partitioning and to simulate the spatial variability of soil moisture (Soltani 2022). Nonetheless, accurately capturing soil moisture dynamics remains challenging due to the heterogeneity of natural landscapes, including variations in soil composition, land use, and terrain, as well as the intermittent nature of precipitation and the simplifications required by modeling frameworks (Brocca et al. 2017).

Recent research highlights the value of integrated terrestrial modeling platforms (e.g., AquiferFlow-SiB2, CATHY, MIKE SHE, ParFlow/CLM, ParFlow/WRF, and PIHM) in effectively simulating the dynamic interactions between surface and subsurface systems (Abbott et al. 1986; Kollet & Maxwell 2006; Maxwell & Miller 2005). These coupled models are particularly useful for analyzing soil water content (SWC) variability, especially when supported by in-situ measurements at the hillslope scale. They have been extensively applied to investigate the factors influencing SWC patterns and their role in modulating rainfall-runoff processes (Cornelissen et al. 2014; Herbst et al. 2006). Among these, ParFlow/CLM is particularly well-suited for this study due to its fully coupled nature, which allows it to explicitly resolve the two-way feedbacks between surface and subsurface water dynamics, a critical process for accurately simulating soil water content.

These studies underscore the capability of fully coupled hydrological models to enhance the understanding of fine-scale soil moisture dynamics at the catchment scale (Soltani et al. 2022a, 2025b; Soltani 2025). However, achieving high accuracy and reliability in estimating critical variables such as soil moisture and streamflow remains a challenge for prognostic models, largely due to uncertainties associated with hydrometeorological inputs, model parameters, boundary and initial conditions, and structural assumptions. Despite these limitations, free-running models still provide valuable insights, particularly for applications like flood risk assessment and management (Abbaszadeh et al. 2019).

One promising approach to overcome these limits is the application of Machine Learning (ML) techniques, which have demonstrated exceptional capabilities in handling complex, high-dimensional datasets (Drogkoula et al. 2023). ML models excel at learning intricate patterns from historical data, enabling them to provide accurate predictions even in the face of variable and nonlinear relationships. For flood forecasting, hydrological variables such as rainfall, river discharge, and snowmelt, along with environmental factors like SWC and terrain features, serve as critical inputs. By leveraging these data, ML models can produce reliable predictions, contributing to early warning systems and effective disaster management strategies (Kumar et al. 2023; Izadi et al. 2024).

A variety of machine learning methods have been applied to flood prediction. Recent studies demonstrate promising results using techniques such as Long Short-Term Memory (LSTM) networks (Le et al. 2019; Rajab et al. 2023), Artificial Neural Networks (ANNs) (Tayfur et al. 2018), combination of KNN and Random Forest (RF) (Tang et al. 2023)

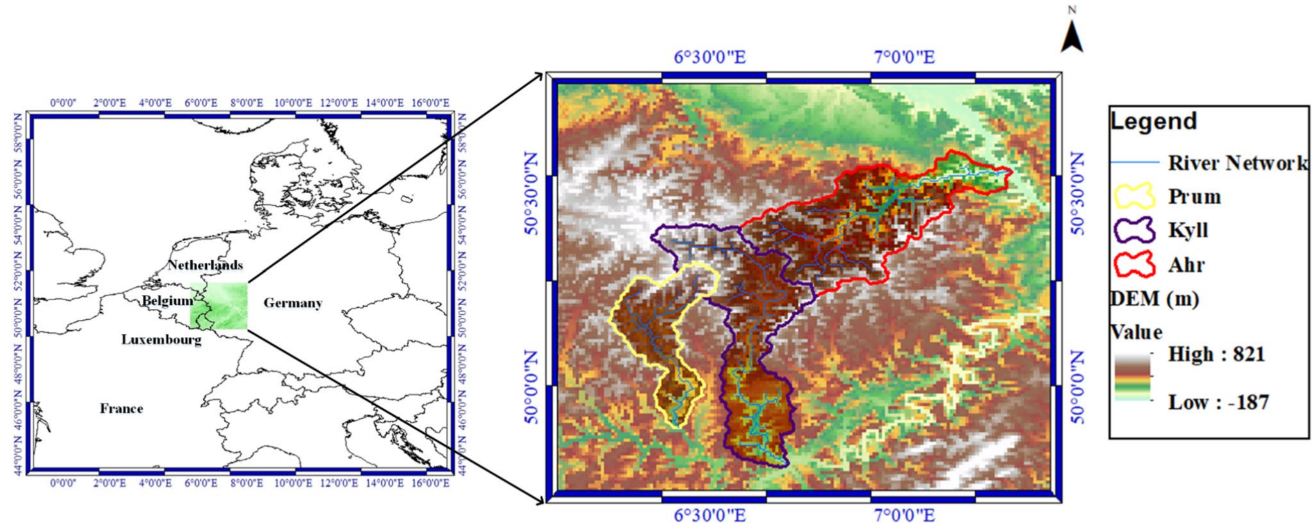


Fig. 1 Overview map of the study area with European context (left), Digital Elevation Model (DEM) overlaid by the main rivers (in blue) and basins (right)

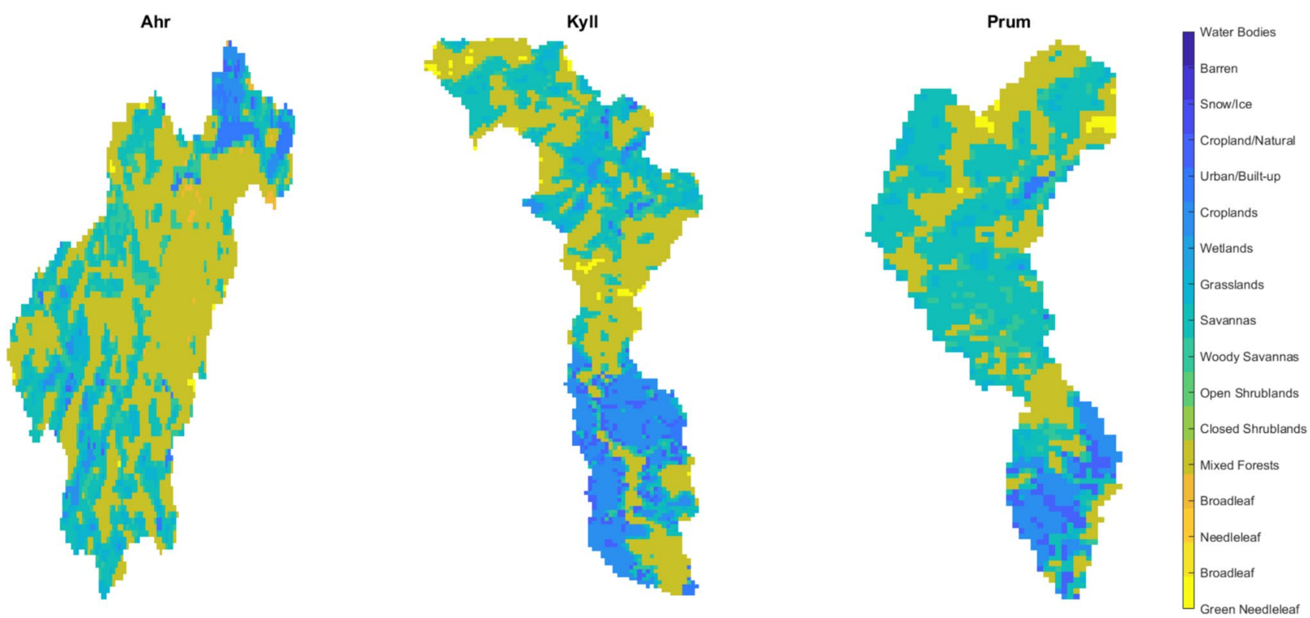


Fig. 2 Land cover classification of the basin Kyll, Prum and Ahr based on MODIS Land Cover Type Product (MCD12Q1)

or combination CNN with Conv-GRU (Miau and Hung 2020). Approaches that combine multiple machine learning techniques are becoming increasingly popular (eg. Moishin et al. 2021, Kordani et al. 2024, Chang et al. 2025). These models leverage the strengths of various machine learning techniques for predictive accuracy, creating robust frameworks for real-time flood forecasting (Zhang et al. 2023). The benefits of AI-based models include handling large datasets, addressing dynamic and nonlinear relationships, integrating data from various sources, reducing manual intervention, and adapting through learning from new data.

However, challenges persist in the accuracy and generalizability of these models. Human activities, such as land-use changes and water conservation, can diminish model efficiency, while regional factors like climate and topography may hinder generalization (Liu et al. 2021). Additionally, AI models often struggle with predicting seasonal and intermittent stream flows (Forghanparast and Mohammadi 2022), and a key area of ongoing research is the development of hybrid systems that can better integrate physical understanding for sustainable resource management (Marshall et al. 2025). Furthermore, without a physical understanding of hydrological processes, AI models may fail to capture the complex interactions driving flooding events.

To overcome these limitations, research suggests leveraging AI-based models and physical models in tandem to improve predictive accuracy and interpretability. AI models can enhance the accuracy and robustness of physical models by processing real-time data and improving parameter estimation. Conversely, physical models can provide the necessary physical understanding to guide AI model training and improve its interpretability. Such integration can lead to more reliable flood forecasting systems that are both accurate and grounded in physical reality (Cho and Kim, 2022; He et al. 2023).

Soil moisture, as one of the key physical parameters, still remains as one of challenges in modeling and predicting runoff estimations and floods (Brunner et al. 2021). Researchers who have incorporated soil moisture into their analyses show enhanced predictive capabilities (Wanders et al. 2014; Alvarez-Garreton et al. 2014). Traditionally, soil moisture measurements have been captured and assimilated in hydrological modeling relying on reanalysis or remotely sensed soil moisture estimates from datasets like ASCAT (Advanced SCATterometer), ERA5, AMSR-E (Advanced Microwave Scanning Radiometer - Earth Observing System), NASA's Soil Moisture Active Passive (SMAP), SMOS (Soil Moisture and Ocean Salinity), which may suffer from very coarse resolution which is not sufficient for accurate local predictions (Meng et al. 2017). Therefore,

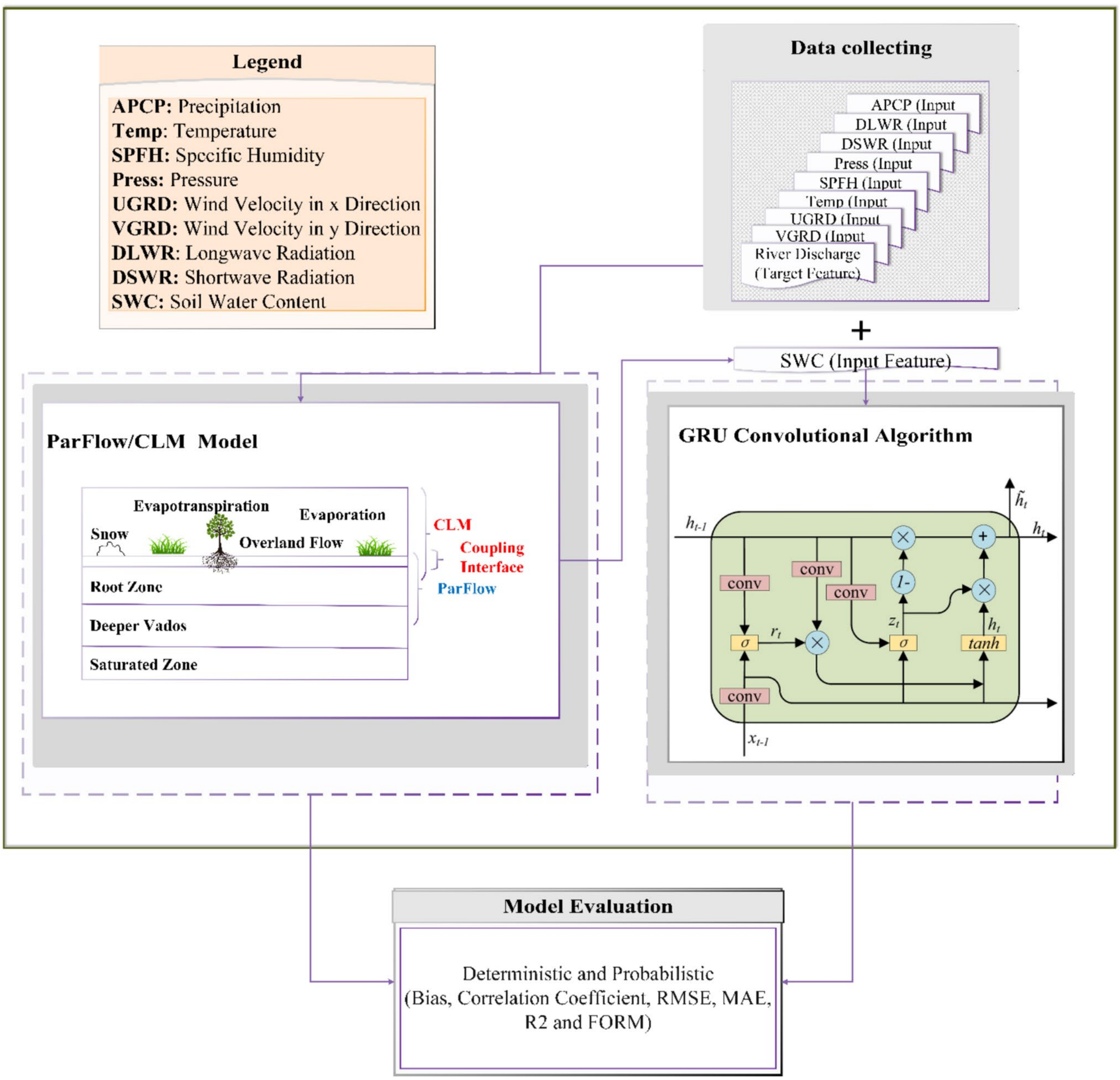


Fig. 3 Flowchart of methodology followed in this study

the more reliable and higher-resolution soil moisture data becomes available, the better the prediction of runoff generation and flood predictions can be expected. In addition, the representation of soil moisture at a fine scale allows for a better understanding of the interactions between surface and subsurface water flow.

The primary objective of this research is to highlight the synergy between machine learning and physics-based modeling, demonstrating how their integration can significantly enhance the reliability of flood forecasting systems. This study employs the ParFlowCLM model to generate high-resolution soil moisture data, crucial for better flood prediction, while also incorporating an AI-driven approach for river discharge forecasting. By developing a robust framework for real-time flood prediction, this integrated model aims to offer valuable insights for water resource managers and policymakers, ultimately supporting the mitigation of flood-related damages and the protection of vulnerable communities. Additionally, the study addresses three key research questions:

1. How can ParFlow/CLM be utilized for flood simulation? This question examines the effectiveness of ParFlow/CLM in modeling flood events by assessing its ability to represent hydrological and land surface processes accurately.

2. How does ParFlow/CLM compare to AI-based methods in flood simulation? This question explores the strengths and limitations of ParFlow/CLM relative to AI-based approaches, evaluating their accuracy, efficiency, and applicability in flood modeling.

3. Can the integration of AI and physics-based modeling enhance flood simulation? This question investigates whether combining AI techniques with ParFlow/CLM can improve model performance, assessing whether these approaches can complement each other for more accurate flood prediction and risk assessment.

2 The July 2021 Flood in Western Europe

The study focuses on the Eifel-Ardennes mid-mountain region, which experienced severe flooding during the July 2021 event. This region, known for its steeply incised valleys, extends across parts of western Germany, eastern Luxembourg, and southeastern Belgium (refer to Fig. 1). The flooding impacted an extensive area, exceeding 20,000 km^2 and reaching locations east of the Rhine River (Mohr et al. 2022). The dominant land cover consists mainly of forests and agricultural lands, with only a small portion classified as urban, as documented by the MODIS Land Cover Type Product (MCD12Q1; Friedl et al. 2002, 2010; see Fig. 2). Climatically, the region is temperate with a strong oceanic influence, receiving annual rainfall between 700 and 1,080 mm, and exhibiting aridity indices that range from 0.5 to 0.9 (Saadi et al. 2023).

Schröter et al. (2015) identified two typical precursors to major flood events in Germany: intense rainfall following dry conditions and moderate rainfall after sustained wet periods. Prior to the July 2021 flood, much of the region had experienced a wetter-than-average month, as reflected by the Antecedent Precipitation Index (API), indicating elevated soil moisture levels (Mohr et al. 2022). Junghänel et al. (2021) further documented frequent rainfall in the three weeks before the event. Soil moisture storage was particularly low, often under 10 mm, in areas like the southern Eifel, the Ardennes, and the Wupper region, while other zones recorded slightly higher levels between 10 and 30 mm, occasionally peaking at 75 mm, though still below long-term averages.

Luxembourg experienced similarly moist antecedent conditions (AGE 2021). Although soil moisture levels were not unusually high, they contributed to flood development through saturation excess. The heavy rainfall further promoted rapid runoff generation by increasing infiltration and enhancing overland flow.

Table 1 Overview of the river basin and gauge characteristics, including water level (W) and streamflow (Q) data. Historical flood statistics and estimates for the July 2021 flood are also included (Mohr et al. 2022)

		Previous historical extreme		Statistical extreme	Flood event in July 2021				
Gauge name	Basin size (km^2)	Max. W (cm)	Max. Q (m^3/s^{-1})	HQ$_{100}$ (m^3/s^{-1})	Peak time (UTC)	Max. W (cm)	Max. Q (m^3/s^{-1})	Peak factor (-)	Operator and data provider
Ahr river basin									www.lfu.rlp.de
Altenahr	749.0	371	236	241	15 July 00:00	984–1019	1000	4	
Kyll river basin									www.lfu.rlp.de
Jünkerath	175.6	266	129	118	14 July 22:45	370	200	1.7	
Kordel	816.3	481	218	248	15 July 08:30	600	600	2.5	
Prüm river basin									www.lfu.rlp.de
Prüm 2	53.2	126	43.5	51.6	14 July 21:15	330	120	2.3	
Prümzurlay	576.1	492	252	278	15 July 05:30	700	600	2	

3 Methodology and Data

In this study, we developed and compared three different modeling approaches to simulate streamflow across five study locations. These included two machine learning models, with and without the inclusion of Soil Water Content (SWC), and a physics-based hydrological model.

First, the ParFlow/CLM model is used to simulate hydrological processes using physical principles and environmental inputs. In parallel, machine learning models are trained on observed and simulated data to learn patterns and enhance prediction performance. Finally, both approaches are integrated to capitalize on the strengths of each, merging the physical interpretability of ParFlow/CLM with the adaptive learning capabilities of AI models. In the final integrated approach, Soil Water Content (SWC) extracted from ParFlow/CLM are incorporated as one of the input features into the AI models. This integration allows the AI models to leverage physically derived soil moisture information, thereby improving their ability to predict streamflow and runoff more accurately.

To effectively address the research questions, Fig. 3 illustrates the flowchart of methodology followed in this study. Initially, the required data is collected including river gauge and forecast data (such as meteorological longwave and shortwave radiation, total precipitation, air temperature, air pressure, specific humidity, wind velocity components). This data is then preprocessed to remove outliers in data training. The flood simulation is conducted through three main steps: 1) flood simulation using ParFlow/CLM, 2) comparison with AI-based methods and 3) integration of AI and physics-based modeling. Finally, the accuracy of the models is evaluated and compared using multiple assessment metrics. In the following, a brief description of each material and method used is provided.

3.1 Data

3.1.1 Land Surface Data

ParFlow/CLM incorporates 18 land cover classifications established by the International Geosphere-Biosphere Programme (IGBP), which are further converted into Plant Functional Types (PFTs) through specific parameterization. The model utilizes the Corine Land Cover (CLC2018 v20) dataset, which provides Europe-wide land cover information at a 100-meter resolution. Initially, the 44 land cover categories defined in CLC2018 are reclassified into the 18 IGBP land cover types, following the methodology outlined in Belleflamme et al. (2023).

3.1.2 Atmospheric Forcing

ParFlow/CLM requires eight key atmospheric variables to simulate surface and near-surface hydrological processes: longwave and shortwave radiation, total precipitation, air temperature, atmospheric pressure, specific humidity, and the two components of wind velocity (Belleflamme et al. 2023). To supply these variables, the model uses data from the High-Resolution (HRES) deterministic medium-range forecast provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Owens and Hewson 2018). These forecasts are initially available at a spatial resolution of 0.1° × 0.1° (approximately 11 km), and are subsequently refined to a 0.0055° × 0.0055° grid (roughly 0.611 km) through bicubic interpolation, as outlined by Belleflamme et al. (2023). This downscaling approach ensures the delivery of high-resolution, temporally coherent atmospheric forcing for ParFlow/CLM simulations.

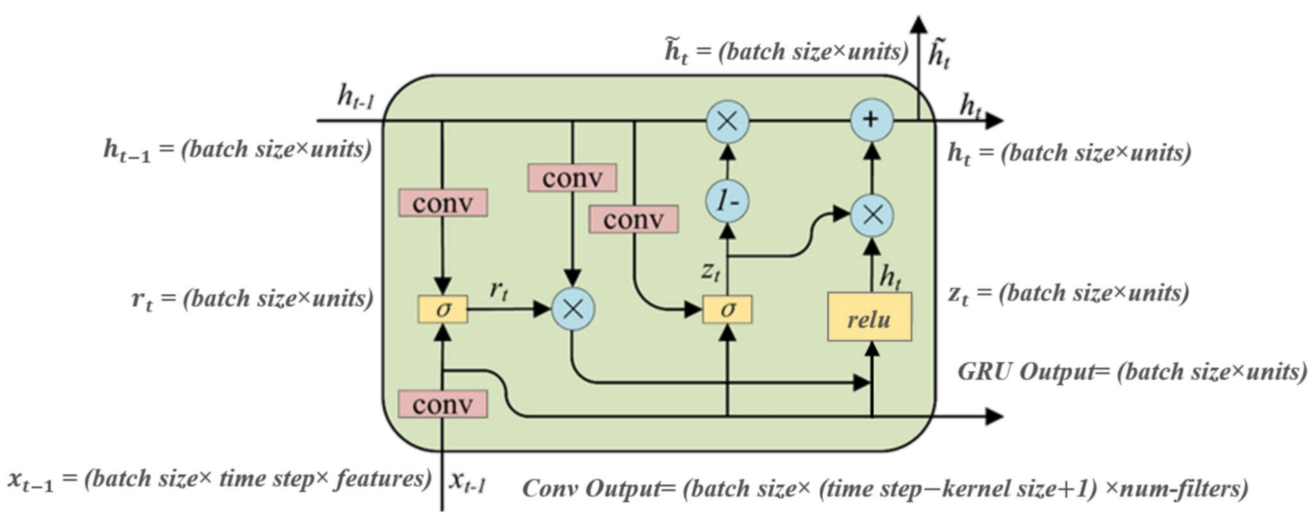


Fig. 4 GRU-Convolutional model architecture. Text labels inside the diagram indicate the tensor dimensions at each major stage of processing

3.1.3 Soil types and Hydraulic Parameters

ParFlow relies on soil hydraulic properties to numerically solve water flow in variably saturated porous media. Essential parameters include porosity, saturated hydraulic conductivity, and those governing the Van Genuchten soil-water retention model. The texture of the upper soil layer was derived from the SoilGrids250m dataset (Hengl et al. 2017), following the methodology outlined by Belleflamme et al. (2023).

3.1.4 Evaluation Data: River Gauge Data

Water level (W) and streamflow (Q) data were collected from multiple river gauges distributed across the study area. The selection process was designed to meet several key criteria: (a) Ensuring broad geographic representation, spanning from the Wupper River in the east to the Amblève River in the west; (b) Including a diverse range of basin sizes, with the smallest at gauge Prüm 2/Prüm (53.2 km^2) and the largest at gauge Kordel/Kyll (816 km^2); and (c) Strategic placement along rivers, aiming to include two gauges per river, one in the headwater region and another closer to the confluence, where feasible.

Water levels were measured directly, while streamflow values were typically calculated from those measurements using site-specific water level–discharge (W-Q) relationships. When water level data were unavailable, most often because floods damaged measuring instruments, river stages exceeded the established W-Q range, or riverbed morphology changed, streamflow was estimated using hydraulic models or data from nearby gauging stations. All hydrological data were sourced from relevant water management authorities, including the federal state of Rhineland-Palatinate and the Wupperverband. A detailed summary of the collected gauge data is provided in Table 1.

3.2 ParFlow/CLM Model

In this study, we utilize the Integrated Hydrological Model (IHM) ParFlow/CLM, an advanced, fully coupled subsurface-land surface model optimized for parallel computing. The model, grounded in partial differential equations (PDEs), simulates the complex interactions of unsaturated and groundwater flow, alongside surface water movement, within a unified continuum framework (Maxwell et al. 2009; Kuffour et al. 2020). ParFlow applies the Richards equation to model water flow in three-dimensional, variably saturated porous media (Ashby and Falgout 1996).

ParFlow has established itself as a leading tool in integrated, physics-based hydrological modeling, with applications across diverse domains and research contexts. It has been employed in theoretical hydrodynamic studies, sensitivity analyses, and basin hydrodynamics research (Schalge et al. 2019; Maina et al. 2020; Soltani et al. 2025a, 2022b, 2025b; Youssefi et al. 2025). Additionally, ParFlow is integral to hydrological scaling studies (Fang et al. 2015; Soltani et al. 2024) and the Terrestrial Systems Modelling Platform (TSMP), which is used for climate-scale assessments and sensitivity analyses (Furusho-Percot et al. 2022).

For this research, we adopt the ParFlow/CLM configuration outlined by Belleflamme et al. (2023), designed to monitor and predict subsurface water resources. This configuration has also been applied to study the catastrophic flood in western Germany and eastern Belgium in July 2021. Saadi et al. (2023) examined the effect of different radar-based precipitation datasets on hydrological simulations, specifically stream discharge. The integration of the Common Land Model (CLM), developed by Dai et al. (2003) and incorporated into ParFlow (Kuffour et al. 2020), allows for a detailed coupling of water and energy exchanges between the surface and subsurface, with near-surface atmospheric conditions as boundary constraints.

Table 2 Evaluation metrics: definitions, equations, ranges, and units

Evaluation Indexes	Equation	Unit	Range	Optimal value
Absolute Bias	$\frac{\left\lvert\sum_{i=1}^{N}(y_i-Y_i)\right\rvert}{N}$	M^3/S	$[0 \sim \infty]$	0
Root Mean Square Error (RMSE)	$\sqrt{\frac{\sum_{i=1}^{n}(Y_i-y_i)^2}{n}}$	M^3/S	$[0 \sim \infty]$	0
Correlation Coefficient	$\frac{\sum_{i=1}^{N}(y_i-\overline{y})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{N}(y_i-\overline{y})^2}\times\sqrt{\sum_{i=1}^{N}(Y_i-\overline{Y})^2}}$	NA	$[-1 \sim 1]$	1
Mean Absolute Error (MAE)	$\frac{\sum_{i=1}^{N}\lvert y_i-Y_i\rvert}{N}$	M^3/S	$[0 \sim \infty]$	0
R-Squared	$1-\frac{\sum_{i=1}^{N}(y_i-Y_i)^2}{\sum_{i=1}^{N}(Y_i-\overline{Y})^2}$	NA	$[0 \sim 1]$	1

3.3 Machine Learning Modelling

Machine learning (ML), which is a subset of AI, enables models to identify patterns and predict based on data and process complex relationships in large datasets. AI has been widely applied in various fields, including hydrology, where it enhances the accuracy of streamflow prediction and flood forecasting. In this research, we integrated a sophisticated physic-based hydrological model, ParFlow/CLM, with a machine learning approach, the Gated Recurrent Unit Convolutional (GRU-Convolutional) model, to predict river discharge. What makes GRU-Convolutional suitable for this research is that Several studies have demonstrated that deep learning models, specifically GRU-based models, outperform traditional models and even GRU-Convolutional models often outperform other deep learning architectures, such as Long Short-Term Memory (LSTM) networks, in flood forecasting and hydrological modeling (Xu et al. 2024; Miau and Hung 2020; Wegayehu and Muluneh 2021; Zhao et al. 2024).

The Gated Recurrent Unit (GRU) Convolutional model is a deep learning architecture that integrates a Convolutional Neural Network (CNN) with a GRU. This combination leverages the feature extraction capabilities of CNN with the sequential processing strength of GRU. The GRU-Convolutional model consists of two main components including Convolutional Layers (CNN), extracting spatial features and local patterns from the input hydrological variables and GRU, capturing long-term dependencies in sequential data, handling the temporal evolution of river discharge (He et al. 2023; Cea and durán L. 2024).

At the first step, the input hydrological time-series data are first passed through convolutional layers, which apply 1D convolutional filters to extract spatial relationships among the hydrological variables. The convolutional layers detect short-term dependencies and local features within the dataset. Then, the output from the convolutional layers is fed into GRU units, which capture long-term dependencies by maintaining information from previous time steps while filtering irrelevant data using gating mechanisms (Fig. 4). Lastly, the final output of the GRU network is passed through a fully connected (dense) layer to predict river discharge (Miau and Hung 2020).

The convolution operation is represented by the following equation:

$$y_i = \sigma\left(\sum(W_i * x_i + b_i)\right) \tag{1}$$

Where x_i represents the input data, y_i denotes the output data, W_i is the convolution kernel weight, and b_i is the convolution kernel bias. The function σ serves as the activation function. The principles of the processes in GRU are represented by following Equations:

$$z_t = \sigma\left(W_z x_t + U_z h_{t-1} + b_z\right) \tag{2}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{3}$$

$$\widetilde{h}_t = tanh(W_c x_t + U_c r_t h_{t-1} + b_h) \tag{4}$$

$$h_t = (1 - z_t) \times \widetilde{h}_t + z_t \times h_{t-1} \tag{5}$$

Where z represents the update gate, determining how much past information needs to be carried forward, and r denotes the reset gate, deciding how much past information to forget. U, and b are the weight and bias parameters. h is the hidden state, a weighted combination of the previous hidden state and the new candidate memory, while $\widetilde{h}_t$ is the candidate hidden state, computing the new memory content. The functions σ correspond to the sigmoid functions, respectively. The variable t indicates the current time step. Both z and r take values within the range of 0 to 1 (He et al. 2023).

The methodological framework consists of two key phases: initial model training with hydrological data and subsequent enhancement through the incorporation of SWC derived from ParFlow. The primary objective of predicting river discharge in this study is to compare performance of a physical model and machine learning in flood forecasting and also assess how much output of a physical model (here SWC) could affect the accuracy of a machine learning model.

The performance of DL models critically depends on the precise tuning of hyperparameters. As suggested by Izadi et al. 2024, the calibration and validation process is achieved through an iterative training-evaluation approach, outlined in the following steps.

1) Data preparation: Splitting the historical input and streamflow data into training and testing sets.
2) Model training: Training the streamflow prediction model using a specific combination of hyperparameters and operators.
3) Performance evaluation: Assessing the predictive accuracy of the trained model on the testing dataset using defined evaluation metrics (Sec. 3.4.1).
4) Iteration: Repeating steps 2 and 3 with different sets of hyperparameters and operators.
5) Comparison: Analyzing and comparing the outcomes of different iterations and selecting the best-performing hyperparameter combination.
6) Optimal time-lag selection: Identifying the most suitable input time lag

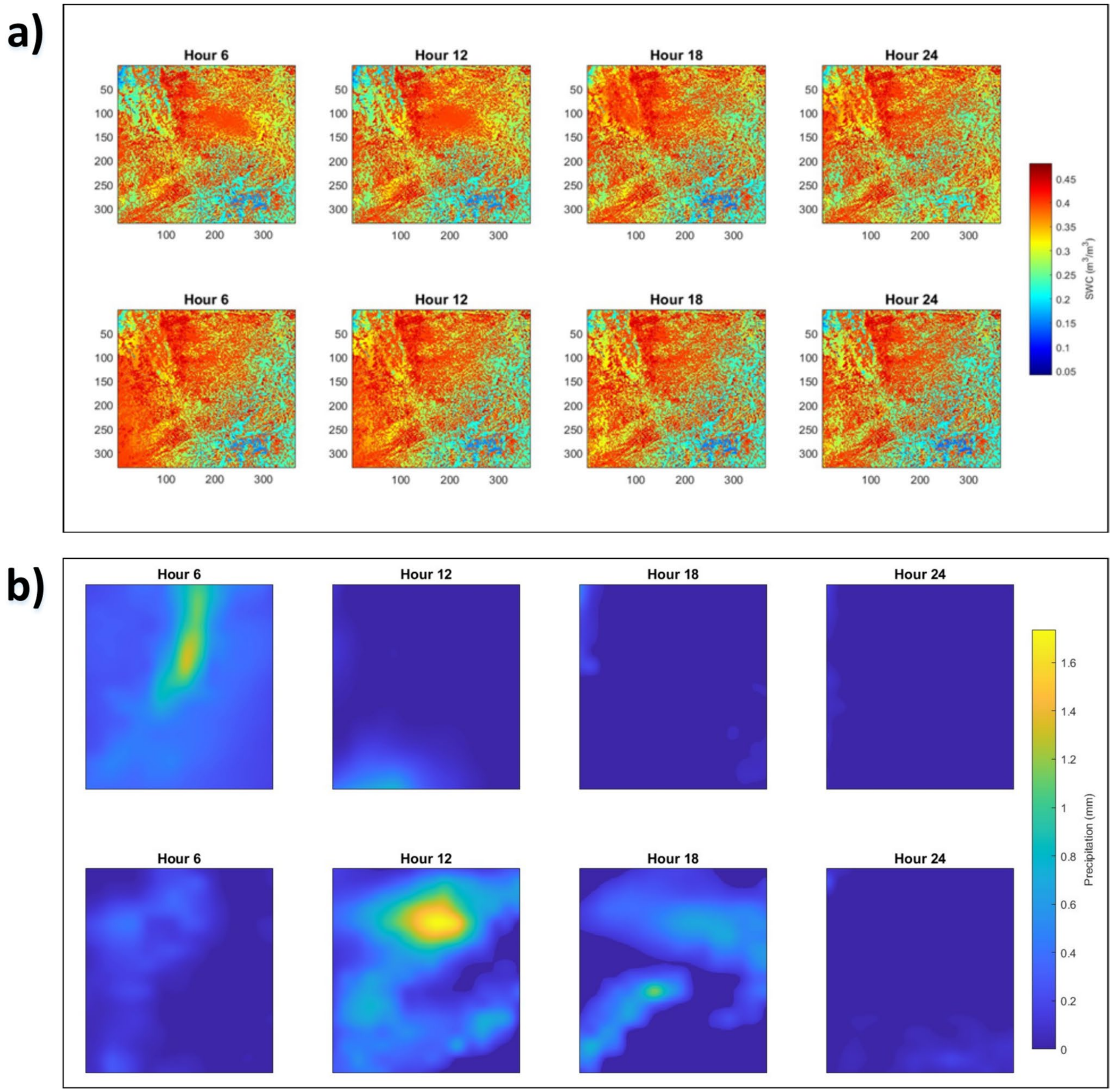


Fig. 5 Spatial pattern of (**a**) SWC and (**b**) precipitation over 14 (first row)−15 (second row) the July flood event, simulated using the ParFlow/CLM hydrological model

3.4 Evaluation metrics

3.4.1 Deterministic

Model performance in simulating hydrological variables and also the GRU Convolutional was assessed using 5 evaluation metrics including Mean Absolute Error (MAE), measuring the average magnitude of errors, Root Mean Square Error (RMSE), quantifying the model's overall predictive accuracy, Absolut Bias, evaluating systematic over- or under-predictions, Correlation Coefficient (R), assessing the strength of the relationship between observed and predicted values and R-squared (R^2), assessing the proportion of variance explained by the model, shown in Table 2. This suite of metrics was chosen to provide a comprehensive evaluation of model performance, with RMSE penalizing large errors, MAE giving a sense of the average error magnitude, R assessing the linear correlation, and R2 quantifying the proportion of variance explained by the model. In this table, $y_i\left(m^3/S\right)$ and $Y_i\left(m^3/S\right)$ represent the predicted and observed river discharge at time step i, respectively. The parameters $\overline{y}\ and\ \overline{Y}$ denote the average values of the predicted and observed river discharge over the total number of time steps, N. The observed streamflow values (Y_i) were obtained from river gauges records, while the predicted river discharge values (y_i) were estimated using the simulating model and GRU Convolutional model.

3.4.2 Probabilistic

The First-Order Reliability Method (FORM) (Madsen et al. 2006) is a widely recognized probabilistic technique used to evaluate the reliability of engineering systems. It offers an approximate yet efficient approach for estimating failure probabilities by transforming complex limit state functions (LSF) into a standardized, normally distributed space. While FORM is commonly applied in structural and mechanical reliability analysis, its utility extends to hydrology, particularly in assessing uncertainties related to water resources, flood risk, and hydrological modeling. Research by Soltani et al. (2020, 2024) has demonstrated the effectiveness of FORM in evaluating probabilistic risks associated with water systems.

At the heart of FORM is the LSF, which delineates the boundary between safe and failure conditions based on uncertain system parameters. The failure condition arises when the system enters an undesirable state. Given that these random variables may follow different probability distributions, FORM transforms them into a standard normal space with independent variables, zero mean, and unit variance. For the purposes of this study, a failure condition is defined as a situation where the simulated peak discharge (Q peak, simulated) is less than 90% of the observed or expected peak discharge (Q peak, observed). The corresponding LSF for this failure condition can be expressed as:

$$LSF = Q_{peak,\ simulated} - 0.9^{*}_{Q\ peak,\ simulated} \quad (6)$$

In this study, failure is defined by a limit state function (LSF) where the simulated peak discharge is less than 90% of the observed value. Q, which inherently reflects the combined uncertainties from the atmospheric input data, the initial and boundary conditions, and the ML model parameters. Both observed and simulated Q values are analyzed, and several candidate probability density functions (PDFs), including Normal, Lognormal, Gamma, Weibull, Exponential, and Generalized Extreme Value (GEV) distributions, are evaluated using different statistical fitting methods (e.g., Kolmogorov–Smirnov, Anderson–Darling, and Akaike Information Criterion (AIC)). Based on standard goodness-of-fit criteria, the Lognormal distribution is identified as providing the best fit for Q. The parameters of the selected Lognormal distribution, μ (mean of ln Q) and σ (standard deviation of ln Q), are determined through statistical estimation from the combined observed and simulated Q dataset. This approach ensures that the probabilistic assessment incorporates all relevant sources of uncertainty in a statistically consistent manner.

A key component of FORM involves identifying the Most Probable Point (MPP), the point on the LSF closest to the origin in the standard normal space, which represents the most likely failure scenario (Soltani et al. 2020, 2024). This is determined through an optimization process, typically solved using the Hasofer-Lind-Rackwitz-Fiessler (HLRF) algorithm, an iterative gradient-based method (Liu and Der Kiureghian 1991; Papaioannou and Straub 2021). The Euclidean distance from the origin to the MPP defines the reliability index (β), a measure of system safety. Using the standard normal cumulative distribution function ($\Phi\left(-\beta\right)$), the failure probability is approximated as:

$$P_f = \Phi\left(-\beta\right) \quad (7)$$

Table 3 The final selected hyperparameters in machine learning modeling

Points		Num. of convolutional layers	Filters	Kernel sizes	Activation functions	Optimizer	Learning rate	Batch size	Num. of epochs	Num. of GRU units
Kordel-Kyll	With SWC	1	32	3	relu	rmsprop	0.001	105	125	45
	Without SWC	1	50	3	relu	rmsprop	0.001	120	120	40
Jünkerath-Kyll	With SWC	1	32	3	relu	rmsprop	0.001	120	130	30
	Without SWC	1	25	5	relu	rmsprop	0.001	130	120	35
Altenahr-Ahr	With SWC	1	32	3	relu	rmsprop	0.001	100	120	30
	Without SWC	1	50	3	relu	rmsprop	0.001	100	120	35
Prüm2-Prüm	With SWC	1	50	3	relu	adam	0.001	120	120	40
	Without SWC	1	50	5	relu	adam	0.001	100	120	40
Prümzurlay-Prüm	With SWC	1	32	3	relu	rmsprop	0.001	120	150	40
	Without SWC	1	28	3	relu	rmsprop	0.001	90	140	40

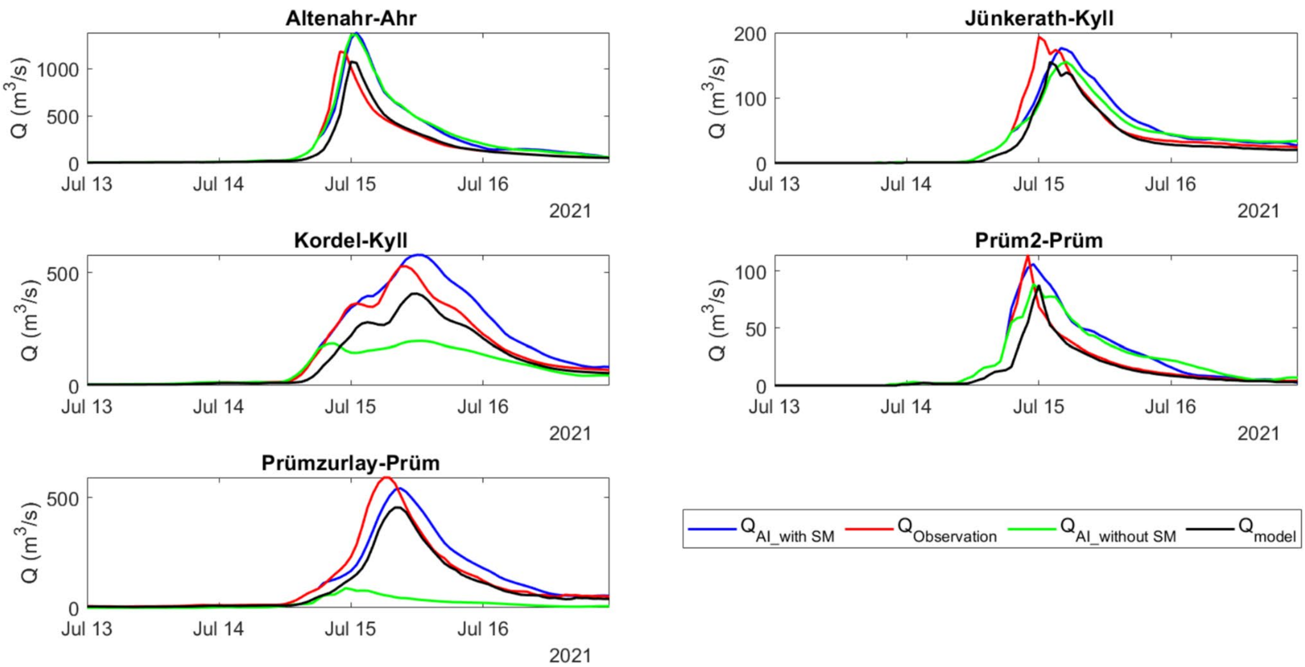


Fig. 6 Simulated hourly hydrograph of by AI-based (with and without SWC), ParFlow/CLM model and observed hourly discharge over flood event period (July 14–15, 2021) at individual stations

4 Results

4.1 Simulated SWC by ParFlow/CLM

Figure 5a shows simulated soil water content (SWC) at 6-hour intervals during July 14–15, showing spatial and temporal variations in response to rainfall inputs. Figure 5b shows precipitation forcing fields at the same time steps, indicating the timing and intensity of rainfall events that drive changes in SWC. The comparison highlights how SWC dynamics follow precipitation pulses, reflecting infiltration, water retention, and subsequent drainage. The simulations were performed with the ParFlow/CLM model at an hourly time step and a horizontal resolution of 0.0055° (~611 m).

4.2 GRU-convolutional model calibration and validation

In this study, five existing hydrometric stations, each located in a distinct basin, were used to capture variations in river discharge during the study period (Table 1). For each station, two GRU-Convolutional based models were developed and validated: one using only atmospheric variables and the other incorporating both atmospheric and hydrological information. In the first model, the input dataset consisted of eight atmospheric predictors, identical to the input variables used in the ParFlow/CLM model, namely near-surface air temperature, specific humidity, meridional and zonal wind components, surface air pressure, total precipitation, and downward shortwave and longwave radiation at the surface (see sect 3.1.2). In the second model, SWC derived from ParFlow/CLM simulations was added as an additional predictor, increasing the number of input features to nine. This integration was designed to evaluate the contribution of hydrological variables to the predictive accuracy of the GRU-convolutional framework and to examine their role in enhancing flood prediction capability. Model outputs were finally compared against observed river discharge at each station during the testing period to assess performance.

To carry out the task using the GRU Convolutional model, the dataset was divided into training and testing sets. Previous studies have shown that the impact of dataset size and train/test splits on model performance can vary widely, and no single ratio is universally optimal (Rácz et al. 2021; Joseph 2022; Joseph and Vakayil 2022). The choice depends on the problem context, model type, and study objectives. In this work, we deliberately adopted a 72%/28% split to strike a balance between training sufficiency and testing reliability. In this partition, the testing set includes major flood peaks to allow meaningful evaluation, while the initial rising stages of floods are retained in the training set to ensure the model can learn the early dynamics of flood development. Temporal dependencies were captured by incorporating a look-back window of 15 time steps as input. The model architecture consisted of a three-layer network, including one convolutional layer, one GRU layer, and one Dense layer.

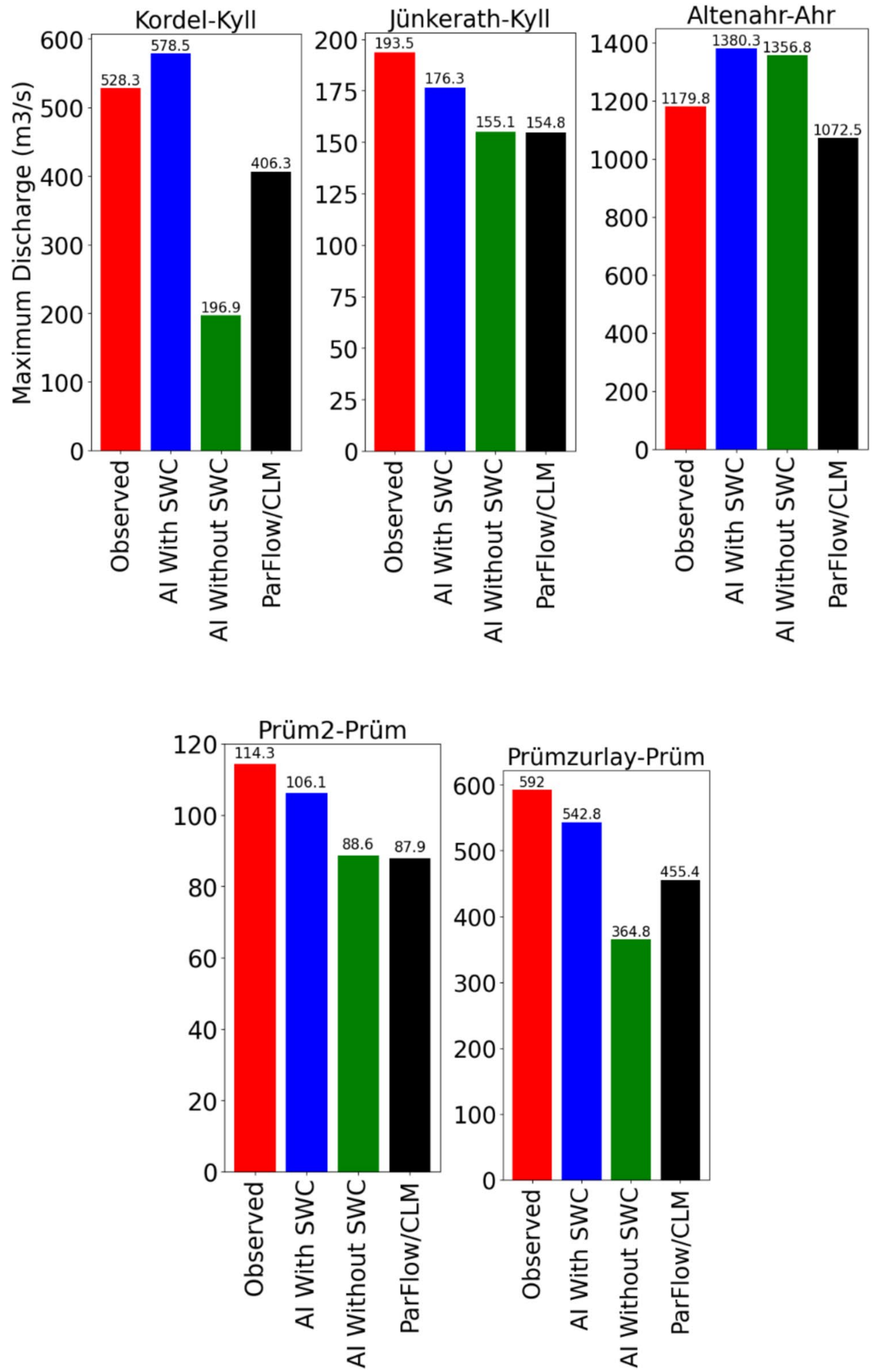


Fig. 7 The comparison of predicted maximum discharge values by three different discharge prediction methods at different stations

Hyperparameter tuning was performed using GridSearchCV with 3-fold cross-validation on the training set, while the testing set was reserved exclusively for final evaluation. The Keras model (keras_model) was used as the estimator, and the hyperparameter search space (param_grid) included key parameters such as learning rate, batch size, number of GRU units, number of convolutional layers, filter sizes, and kernel sizes. The scoring metric was negative mean squared error (neg_mean_squared_error), with mean squared error (MSE) as the loss function. Several additional

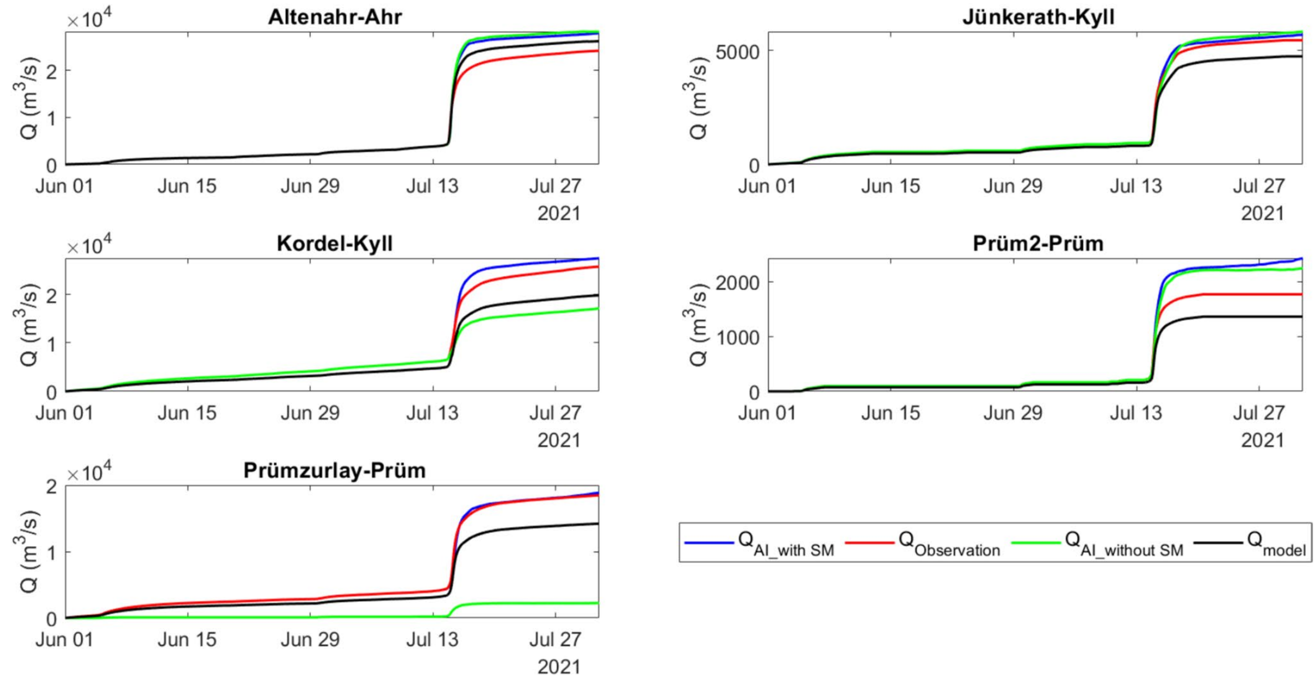


Fig. 8 Simulated Accumulated hourly discharge by AI-based (with and without SWC), ParFlow/CLM model and observed hourly discharge over flood event period (July 14–15, 2021) at individual stations

Table 4 A comparative analysis of three modeling approaches: AI-With SWC, AI-Without SWC, and ParFlow/CLM for predicting streamflow across five locations. The models are evaluated based on Absolute Bias, RMSE, R, MAE, and R^2

Points		Absolute Bias	RMSE	Correlation	MAE	R^2
Kordel-Kyll	AI-With SWC	4.1929	27.7371	0.9828	11.0216	0.91630
	AI-Without SWC	20.9559	57.9778	0.9661	21.2242	0.63433
	ParFlow/CLM	17.15	37.293	0.984	17.15	0.969
Jünkerath- Kyll	AI_With SWC	0.6007	9.3555	0.9444	3.4461	0.8834
	AI_Without SWC	0.9208	9.5128	0.9386	3.2556	0.8794
	ParFlow/CLM	3.658	11.813	0.958	3.785	0.917
Altenahr-Ahr	AI_With SWC	9.0267	71.7738	0.9359	22.0865	0.7406
	AI_Without SWC	9.8128	70.1750	0.9473	22.7961	0.752
	ParFlow/CLM	7.241	70.522	0.920	16.089	0.847
Prüm2-Prüm	AI_With SWC	1.6971	5.4026	0.9703	1.9168	0.7879
	AI_Without SWC	1.1665	5.2094	0.9288	1.9288	0.8166
	ParFlow/CLM	1.5	6.917	0.914	1.744	0.836
Prümzurlay- Prüm	AI_With SWC	1.1060	28.8903	0.9463	10.2930	0.8932
	AI_Without SWC	8.0246	35.4170	0.9490	10.7674	0.8401
	ParFlow/CLM	13.011	37.481	0.97	13.046	0.94

hyperparameters were fine-tuned through a trial-and-error process (see section 3.3).

The entire model was implemented in Python using the TensorFlow and Keras libraries. The final network architecture and the optimized hyperparameters are summarized in Table 3.

4.3 Streamflow Analysis

In the first phase of this study, SWC and streamflow were estimated. In the second phase, streamflow was simulated using the GRU Convolutional model under two scenarios, one incorporating SWC (as output from ParFlow/CLM) and one without SWC.

Figure 6 shows Hourly hydrograph of five stations, Altenahr-Ahr, Jünkerath-Kyll, Kordel-Kyll, Prüm2-Prüm and Prümzurlay-Prüm during the July (14–15 th) 2021 flood event. It compares observed and simulated discharge using different approaches, highlighting the impact of soil moisture data on flood prediction accuracy. Streamflow simulated by ParFlow/CLM performs well but slightly underestimates peak flows. The AI model with soil moisture closely follows the observed data, effectively capturing both the peak and recession of the flood. In contrast, the AI model without soil moisture underestimates the peak at Prümzurlay-Prüm and overestimates it at Altenahr-Ahr, demonstrating the crucial role of soil moisture data. The improved performance of the AI model with SWC, particularly in capturing the rising limb of the hydrograph, can be directly attributed to its ability to leverage the antecedent moisture conditions simulated by ParFlow/CLM (Fig. 5), which showed near-saturated soils prior to the main rainfall event. The AI model that incorporates soil moisture provides significantly better predictions, making it a valuable tool for flood forecasting and risk management. This finding suggests that including soil moisture conditions in flood prediction models can improve preparedness and response efforts for extreme weather events. In terms of hydrograph timing, all models showed a slight delay in reaching peak discharge relative to the observed flood hydrographs. While this lag is minimal, it remains operationally relevant, as precise peak-flow timing is critical for effective flood warnings, emergency decision-making, and response planning. Importantly, the fact that the delay is very small highlights the overall robustness of the models. Among them, the AI-based model with SWC stands out, as it reproduced the timing of peak discharge with the highest accuracy—matching observed peaks at four out of the five stations.

For a more detailed investigation of the flood peak and volume, Fig. 7 and 8 are provided. Figure 7 compares the performance of three discharge prediction methods: the AI-based model with SWC (blue), the AI-based model without SWC (green), and the ParFlow/CLM model (black), against observed values (red). The AI-based model with SWC consistently provided more accurate estimates across locations, with slight over-prediction at Altenahr-Ahr and under-prediction at some sites such as Prümzurlay-Prüm and Jünkerath-Kyll. In contrast, the AI model without SWC showed systematic underestimation, especially at Kordel-Kyll and other Prüm locations. The ParFlow/CLM model performed better than the AI model without SWC but still underestimated discharge in most cases. Overall, incorporating SWC clearly improved prediction accuracy, particularly in regions where soil moisture strongly influences streamflow.

From a flood-warning perspective, a slight overestimation of discharge is generally more advantageous than underestimation. Over-prediction ensures that local authorities and communities are alerted to a potentially higher level of risk, which promotes stronger preparedness and precautionary measures. This proactive bias reduces the likelihood of overlooking critical flood hazards, thereby enhancing the reliability and safety of early-warning systems. In practice, such a conservative approach increases the margin of safety for decision-making, emergency planning, and resource mobilization, ultimately contributing to more effective risk management during extreme hydrological events.

Figure 8 illustrates the accumulated hourly discharge at five selected stations during the flood event of July 14–15, 2021. From the perspective of total flood volume passing through each station, the AI-based model with SWC shows the closest agreement with the observed cumulative hydrographs, particularly at Prümzurlay-Prüm and Jünkerath-Kyll. This highlights the critical role of soil moisture information, which substantially improves the simulation of runoff generation and flood propagation. Nevertheless, at Kordel-Kyll, Prüm2-Prüm, and Altenahr-Ahr, the AI-based model with SWC slightly overestimates flood volumes. This tendency may be related to the smoother recession limb of the simulated hydrographs (see Fig. 6), which declines more gradually than the observed hydrographs across all stations. The AI model without SWC, by contrast, considerably underestimates discharge at Kordel-Kyll and Prümzurlay-Prüm while overestimating it at the other three stations. Unlike the models incorporating SWC, which exhibit more

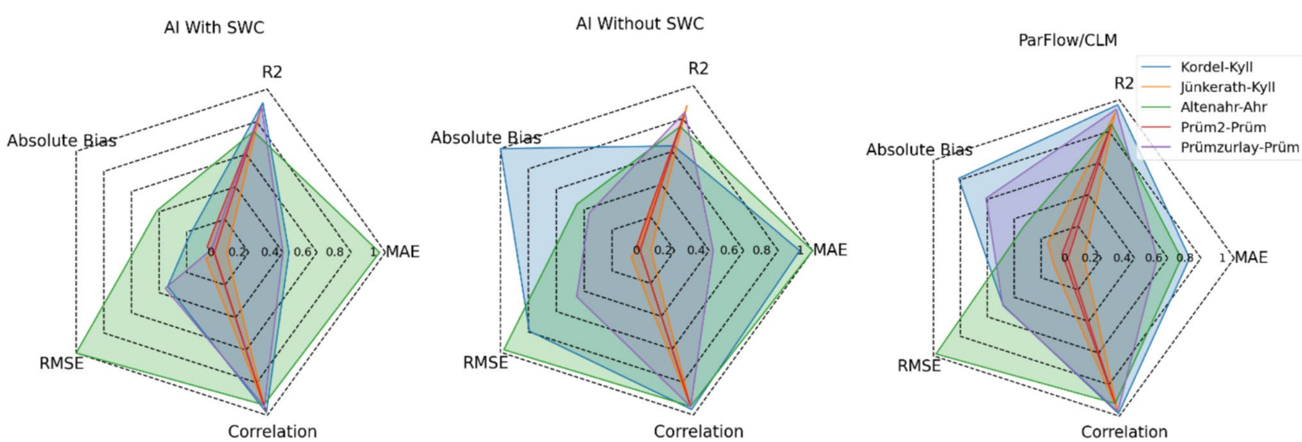


Fig. 9 Spider diagrams comparing the performance of the three different discharge prediction methods Across the five study locations by key evaluation metrics: Absolute Bias, RMSE, MAE along with Correlation, and R^2

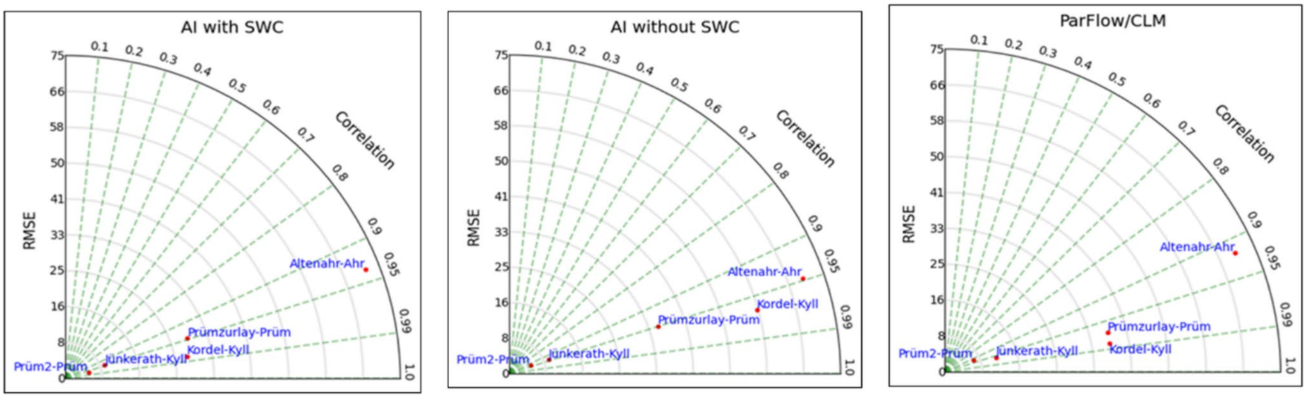


Fig. 10 Taylor diagram comparing the performance of the three different discharge prediction methods across five study stations

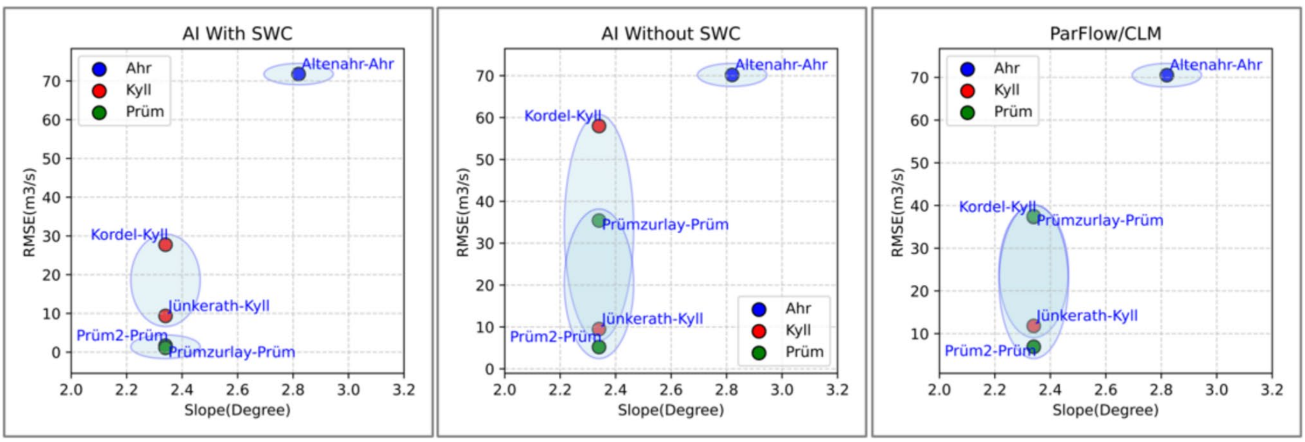


Fig. 11 Scatter plot showing relationship between slope and RMSE for three catchments (Ahr, Kyll, and Prüm) under different modeling configurations: AI with SWC, AI without SWC, and ParFlow/CLM. Colored markers represent individual gauging stations, while shaded ellipses highlight catchment-specific groupings

consistent behavior across all locations, this model shows lower generalizability, as it does not account for the influence of soil moisture and other physics-based parameters. These deviations indicate that neglecting SWC leads to less accurate simulations of flood volume and runoff dynamics. The ParFlow/CLM model generally underestimates discharge compared to observations, except at Altenahr-Ahr. Nevertheless, it successfully reproduces the overall discharge trends at all stations, demonstrating the reliability of physics-based models for simulating flood events. The minor underestimations suggest that further model calibration or higher-resolution input data may be necessary to improve predictive accuracy.

Table 4 presents a comparative analysis of three modeling approaches: AI-With SWC, AI-Without SWC, and ParFlow/CLM, for predicting streamflow across five stations during the flood event July 2021. Consistent with the previous results, an important finding is that AI-With SWC consistently outperforms AI-Without SWC across the majority of locations. This is evident in the lower RMSE and MAE values, which indicate that incorporating SWC enhances the model's ability to simulate streamflow accurately. Additionally, the higher correlation (R) and R^2 values for AI-With SWC suggest a stronger agreement between predicted and observed values. For example, at Kordel-Kyll, AI-With SWC achieves an RMSE of 27.7371, significantly lower than the 57.9778 recorded for AI-Without SWC. This demonstrates the importance of SWC in improving AI-based hydrological modeling.

ParFlow/CLM, a physics-based hydrological model, performs well overall but exhibits higher Absolute Bias and RMSE in certain locations compared to AI models. In some cases, it achieves higher R and R^2 values, reflecting strong agreement between observed and simulated streamflow. For instance, at Jünkerath-Kyll, ParFlow/CLM achieves an R^2 of 0.917, outperforming both AI models. However, at Altenahr-Ahr, while it has a lower correlation coefficient (R = 0.920) than AI models, it achieves a significantly lower MAE (16.089) compared to AI models (~22), suggesting that it may provide more reliable estimates under certain conditions.

Table 5 The performance and reliability of different models based on the FORM

	P_f		
Gauges	AI_With SWC (%)	AI_Without SWC (%)	ParFlow/CLM (%)
Kordel-Kyll	2.90	20	6.45
Jünkerath-Kyll	2.58	6.129	6.45
Altenahr-Ahr	5.16	4.83	2.90
Prüm2-Prüm	2.25	7.09	7.41
Prümzurlay-Prüm	2.58	6.45	7.41

The performance of each model also varies across different locations. At Kordel-Kyll and Prümzurlay-Prüm, the inclusion of SWC in AI models leads to the most significant improvements, highlighting the critical role of soil moisture in runoff generation in these areas. In contrast, at Jünkerath-Kyll, the performance difference between AI-With SWC and AI-Without SWC is minimal, suggesting that other factors, such as rainfall characteristics or land use, may have a stronger influence on streamflow.

At Altenahr-Ahr and Prüm2-Prüm, the AI model without SWC slightly outperformed the model with SWC in terms of RMSE and R2. While this could partly be due to statistical variability, it may also reflect the hydrological characteristics of these basins. Altenahr-Ahr is relatively steep and fast-responding catchment, where flood peaks are driven more directly by high-intensity rainfall than by antecedent soil moisture conditions. In such environments, the inclusion of ParFlow-simulated SWC may introduce additional noise or bias, stemming from uncertainties in the soil moisture simulation, that can offset its potential benefits. This highlights that the value of SWC information is likely to be basin-dependent, with the greatest benefits observed in flatter, slower-responding catchments.

The findings suggest that incorporating SWC into AI models improves their predictive accuracy and helps reduce errors in streamflow simulation. However, ParFlow/CLM remains a robust alternative, particularly in locations where physical processes dominate streamflow dynamics. Given the strengths of both approaches, a modeling framework that integrates AI techniques with physics-based hydrological modeling may provide the most accurate and reliable flood predictions.

Figure 9 illustrates the spider diagram for the five locations providing a comparative visualization of the three modeling approaches based on five performance metrics: Absolute Bias, RMSE, Correlation, MAE, and R^2. Each axis represents a metric and since these metrics have different scales, a normalization process was applied to ensure fair comparisons, Hence, Absolute Bias, RMSE and MAE are normalized using min-max normalization to rescale the values between 0 and 1. As R^2 and correlation are between 0 and 1, they remain the same. Generally, the spider diagram represents that the AI-based model with the inclusion of SWC generally leads to a more reliable model.

A detailed analysis of the spider diagrams and evaluation metrics indicates that all three models at Prüm2-Prüm and Jünkerath-Kyll exhibit lower normalized Absolute Bias, RMSE, and MAE (close to 0), along with higher R^2 and correlation. In contrast, at Altenahr-Ahr, all models show higher normalized Absolute Bias, RMSE, and MAE, coupled with lower R^2 and correlation.

Figure 10 demonstrates Taylor diagrams. An interesting result obviously captured from the Taylor diagram is that all three models exhibit similar performance patterns across the study locations. Notably, weakest model performance is seen in Altenahr-Ahr across all models, while Prüm2-Prüm and Jünkerath-Kyll demonstrate the strongest model performance. Meanwhile, the models perform moderately well in Kordel-Kyll and Prümzurlay-Prüm.

The performance variations observed in the Taylor diagram (Fig. 10) can be explained by key hydrological and geomorphological factors, as illustrated in SM. Figs. 1–3. One of the primary factors influencing model performance is slope (SM. Figure 1). The Ahr basin exhibits a steeper slope compared to the other study locations, and as shown in Fig. 11, Altenahr-Ahr consistently falls in the high-slope, high-RMSE region across all three modeling configurations (AI with SWC, AI without SWC, and ParFlow/CLM). Steep slopes lead to faster runoff responses and higher hydrological variability, making streamflow predictions more challenging. This could explain why Altenahr-Ahr shows the weakest model performance across all three models. Another important factor is stream order (SM. Fig. 2). Higher stream orders typically correspond to more consistent flow regimes, as they integrate larger contributing areas and exhibit more consistent hydrological behavior. SM. Fig. 2 compares the stream network structures and stream order lengths for the Ahr, Kyll, and Prüm basins, showing how these differences can shape flood hydrographs. The Ahr basin has a large proportion of first-order streams combined with moderate second- and third-order channel lengths, suggesting a mixed response: rainfall on the smaller tributaries produces a quick rise in discharge, while flow routing through longer higher-order channels delays part of the runoff, leading to a moderately flashy hydrograph with an extended recession. The Kyll, in contrast, shows a relatively balanced distribution of stream lengths across orders 1–3, which tends to spread runoff contributions over time. This usually produces a smoother hydrograph with a less pronounced peak and longer lag time. The Prüm basin is dominated by first-order streams, with minimal second-order contributions, indicating rapid concentration of runoff. This configuration favors steep rising limbs, sharp flood peaks, and short response times, especially during small to moderate storms. Thus, the hydrograph behavior across these basins reflects how the organization of stream orders controls both flood timing and magnitude. Lastly, the topologic width function (SM. Fig. 3) further supports these observations. The width function is essentially a histogram of contributing stream links as a function of their distance to the outlet. Peaks in the curve indicate where many tributaries occur at similar distances, meaning synchronized flow arrivals at the outlet, which can produce sharper flood peaks in the hydrograph. The topologic width function for Ahr shows a more irregular distribution compared to Kyll and Prüm, indicating more variable flow accumulation patterns, which increases the complexity of streamflow prediction in the Ahr basin. In contrast, the more uniform width functions observed in Prüm and Kyll basins suggest a more predictable hydrological response, aligning with their stronger model performance.

Table 5 provides insights into the performance and reliability of different models based on the FORM: AI-based models with and without SWC and the ParFlow/CLM mode across various locations. The AI-based model with SWC consistently shows the lowest probability of failure (P_f), indicating the highest reliability in discharge prediction. Altenahr-Ahr (5.16%) has the highest Pf within this model, suggesting greater uncertainty compared to other locations. The AI model without SWC shows significantly higher P_f, especially at Kordel-Kyll (20%), highlighting the importance of soil moisture data in improving model accuracy. ParFlow/CLM generally performs better than AI without SWC but still exhibits moderate failure probabilities, particularly at Prümzurlay-Prüm (7.41%). Kordel-Kyll has the highest variation in P_f across models, reinforcing the necessity of incorporating SWC for reliable predictions. Overall, AI_with SWC is the most robust, while AI_without SWC struggles with under-prediction, and ParFlow/CLM remains moderately reliable.

This suggests that in steep, fast-responding catchments, the benefits of including detailed soil moisture information may be less pronounced, and model performance might be more sensitive to the temporal resolution and accuracy of the precipitation forcing. In contrast, in flatter, slower-responding catchments, accurate antecedent soil moisture is likely to be a dominant control on flood generation

5 Discussion

5.1 Limitations of the Single-Event Approach

This study's analysis, training, and validation were conducted exclusively on the July 2021 extreme flood event. While this event provided an exceptional test case due to its severity, data availability, and spatial extent, this focus introduces an inherent limitation: the model's performance on more common, less severe flood events and across different hydrological seasons remains unknown. Consequently, the results should be interpreted primarily as a proof-of-concept demonstrating the integration of physics-based soil water content forecasts within a data-driven flood prediction framework for extreme events, rather than as evidence of a fully generalizable system. However, previous studies have shown that physics-informed and AI-based models, especially those based on continuous simulations rather than solely event-based training, can effectively capture hydrological dynamics across a wider range of conditions (Huynh et al. 2025; Duan et al. 2023; Zhang et al. 2024). This suggests promising potential for extending such models beyond single event scenarios.

This limitation reflects a broader challenge in AI-based hydrological modeling. Models trained on long-term datasets typically reproduce seasonal or average flow dynamics well but tend to underrepresent sharp, high-magnitude peaks (eg., Di Nunno et al. 2023, Izadi et al. 2024). Such peaks may result from short, intense convective storms, rapid snowmelt, or prolonged rainfall. Because these extremes occur far less frequently than typical seasonal discharges and exhibit greater variability in historical records, AI-models learn them less effectively, which limits predictive skill for rare but critical flood events. Future work should expand the model evaluation to a broader range of flood events and hydrological regimes, explicitly testing performance across moderate and severe floods, seasons, and rainfall mechanisms. In parallel, integrating adaptive strategies, such as dynamic model switching and continuous learning, could enhance robustness, enabling reliable real-time forecasting under diverse and evolving hydrological conditions.

5.2 Addressing Potential Data Leakage and Feature Synergy

An important consideration in our study is that the machine learning model receives both the original atmospheric variables and soil water content (SWC) derived from the ParFlow/CLM model, which itself is driven by the same atmospheric inputs. This raises the possibility of data leakage, where the SWC is not an independent source of information but a complex, physics-based transformation of the input variables. Rather than a drawback, we view this as a strength, with ParFlow/CLM effectively serving as a physics-based feature engineering step that enriches the input space with hydrologically meaningful variables. This highlights the value of integrating process-based modeling with data-driven approaches to capture system dynamics more comprehensively.

Regarding independent sources of SWC, several well-established satellite-derived and remote sensing datasets are available (e.g., Sentinel-1, GLEAM, and ERA5-Land). Although the scientific value of these products has been demonstrated in numerous studies, their relatively coarse temporal and/or spatial resolution, rendered them unsuitable for inclusion in the final analysis.

Future work should systematically compare the relative benefits of physics-based SWC against satellite observations and other advanced feature engineering techniques applied directly to atmospheric variables. Such comparative studies would help disentangle the unique contributions of each data source and optimize hybrid modeling frameworks for improved flood prediction.

5.3 Understanding Streamflow Observation and Estimation

In this study, direct measurements of water levels provided the foundation for estimating streamflow. Streamflow values were typically derived from site-specific water level–discharge (W-Q) relationships, which allowed for consistent and reliable calculations under normal monitoring conditions. However, challenges arose during periods when water level data were unavailable. These situations commonly occurred when extreme floods damaged monitoring equipment, when river stages exceeded the calibrated W-Q range, or when morphological changes in the riverbed altered flow dynamics. Under such circumstances, streamflow could not be derived directly from W-Q relationships. Instead, estimates were obtained using hydraulic modeling approaches or, when appropriate, through data interpolation from nearby gauging stations. These alternative methods introduced additional uncertainties but ensured the continuity of streamflow records during critical hydrological events.

6 Conclusions and Future Perspectives

This study highlights the effectiveness of informing machine learning with simulation results from a physically based hydrological model (ParFlow/CLM) to improve flood forecasting. The results demonstrate that incorporating soil water content (SWC) significantly enhances the accuracy of river discharge predictions. AI-based models, particularly those utilizing SWC as an input, showed stronger agreement with observed discharge compared to AI models without SWC and the standalone ParFlow/CLM model. Specifically, the AI model with SWC achieved an RMSE reduction of up to 51% and an increase in correlation (R) values by 1.7%–6.4%, depending on the location. The R^2 value improved from 0.634 to 0.916 at Kordel-Kyll when SWC was included, emphasizing the critical role of soil moisture in runoff generation and flood dynamics. These findings reinforce the importance of incorporating SWC into predictive models for improved flood forecasting accuracy.

While machine learning offers promising improvements, challenges remain. The generalization of AI models across diverse catchments, the need for high-resolution hydrometeorological data, and model interpretability are key areas requiring further investigation. Additionally, integrating AI-driven predictions with physically based models like ParFlow/CLM could lead to more accurate predictions.

Future research should focus on refining integrated modeling approaches by enhancing the coupling between machine learning models and physics-based hydrological simulations. For example, such a framework could be used to drive a process-based model to assess how different terrain scenarios impact hydrological simulations or to evaluate future flood risks under various climate projections (Marshall et al. 2025). Expanding observational networks to collect real-time soil moisture and hydrological data will further improve model reliability. Moreover, the development of adaptive AI models capable of learning from new extreme flood events will be essential for improving resilience in flood-prone regions. Additionally, implementing real-time flood forecasting systems that integrate machine learning with hydrological models could provide more accurate and timely flood warnings, ultimately aiding disaster preparedness and water resource management. By addressing these challenges and fostering interdisciplinary collaboration, future advancements in hydrological modeling can better equip society to mitigate the increasing risks of extreme flooding events driven by climate change.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41748-025-00923-5.

Acknowledgement We gratefully acknowledge the support of the German Research Foundation (DFG) and Earth System Modelling Project (ESM) for funding this study by providing computing time on the ESM partition of the supercomputer JUWELS at Jülich Supercomputing Centre (JSC). We gratefully acknowledge Alexandre Belleflamme for his valuable support in stabilizing the DE06 simulation setup. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project number: 320397309.

Author Contributions Samira Sadat Soltani: Conceived and designed the study, performed the analysis, and wrote the initial draft of the manuscript. Zohreh Mohammadi: Conceived and designed the study, performed the analysis, and wrote the initial draft of the manuscript. Ardalan Izadi: Served as the project supervisor, offering guidance and oversight throughout the study and ensuring the integrity and quality of the research, provided critical revisions and insightful feedback. Stefan Kollet: Served as the project supervisor, provided critical revisions and insightful feedback.

Funding Open Access funding enabled and organized by Projekt DEAL. DFG

Data Availability The data and codes supporting this study's findings are available upon a rational request from the corresponding author.

Declarations

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

References

AGE: Hochwasserereignis Juli 2021. l'Administration de la gestion de l'eau (AGE), Esch-sur-Alzette, Luxembourg, https://eau.gouvernement.lu/fr/actualites/2021/07-juillet/ Hochwasserereignis2021.html (last access: 9 May 2022), 19 July 2021.

Abbaszadeh P, Moradkhani H, Zhan X (2019) Downscaling SMAP radiometer soil moisture over the CONUS using an ensemble learning method. Water Resour Res 55(1):324–344

Abbott M, Bathurst J, Cunge J, O'connell P, Rasmussen J (1986) An introduction to the European Hydrological System—Systeme Hydrologique European, "SHE", 2: Structure of a physically-based, distributed modelling system. J Hydrol 87:61–77

Alvarez-Garreton C, Ryu D, Western AW, Crow WT, Robertson DE (2014) The impacts of assimilating satellite soil moisture into a rainfall–runoff model in a semi-arid catchment. J Hydrol 519:2763–2774

Ashby SF, Falgout RD (1996) A parallel multigrid preconditioned conjugate gradient algorithm for groundwater flow simulations. Nucl Sci Eng 124:145–159

Badawy A, Sultan M, Abdelmohsen K, Yan E, Elhaddad H, Milewski A, Torres-Uribe HE (2024) Floods of Egypt's Nile in the 21st century. Sci Rep 14(1):27031

Beck HE, de Jeu RA, Schellekens J, van Dijk AI, Bruijnzeel LA (2009) Improving curve number-based storm runoff estimates using soil moisture proxies. IEEE J Sel Top Appl Earth Obs Remote Sens 2:250–259

Belleflamme A, Goergen K, Wagner N, Kollet S, Bathiany S, El Zohbi J, Rechid D, Vanderborght J, Vereecken H (2023) Hydrological forecasting at impact scale: the integrated ParFlow hydrological model at 0.6 km for climate resilient water resource management over Germany. Front Water 5:1183642

Brocca L, Melone F, Moramarco T, Singh V (2009) Assimilation of observed soil moisture data in storm rainfall-runoff modeling. J Hydrol Eng 14:153–165

Brocca L, Ciabatta L, Massari C, Camici S, Tarpanelli A (2017) Soil moisture for hydrological applications: open questions and new opportunities. Water 9:140

Brunner MI, Slater L, Tallaksen LM, Clark M (2021) Challenges in modeling and predicting floods and droughts: a review. Wires Water 8(3):e1520

Cea JFF durán L 2024 Streamflow forecasting with deep learning models : a side-by-side comparison in Northwest Spain. pp 5289–315

Chang LC, Yang MT, Chang FJ (2025) Flood resilience through hybrid deep learning: advanced forecasting for Taipei's urban drainage system. J Environ Manage 379:124835

Cornelissen T, Diekkrüger B, Bogena HR (2014) Significance of scale and lower boundary condition in the 3D simulation of hydrological processes and soil moisture variability in a forested headwater catchment. J Hydrol 516:140–153

Dai Y, Zeng X, Dickinson RE, Baker I, Bonan GB, Bosilovich MG, Denning AS, Dirmeyer PA, Houser PR, Niu G (2003) The common land model. Bull Am Meteorol Soc 84:1013–1024

Drogkoula M, Kokkinos K, Samaras N (2023) A comprehensive survey of machine learning methodologies with emphasis in water resources management. Appl Sci. https://doi.org/10.3390/app132212147

Duan Q, Ye A, Gong W, Rakovec O, Schäfer D, Kumar R (2023) A Hybrid Physics–AI Model to Improve Hydrological Forecasts. Artificial Intelligence for the Earth Systems. https://doi.org/10.1175/AIES-D-22-0023.1

Fang Z, Bogena H, Kollet S, Koch J, Vereecken H (2015) Spatio-temporal validation of long-term 3D hydrological simulations of a forested catchment using empirical orthogonal functions and wavelet coherence analysis. J Hydrol 529:1754–1767

Forghanparast F, Mohammadi G (2022) Using deep learning algorithms for intermittent streamflow prediction in the headwaters of the Colorado River, Texas. Water 14(19):2972

Friedl MA, McIver DK, Hodges J, Zhang Xy, Muchoney D, Strahler AH, Woodcock CE, Gopal S, Schneider A, Cooper A, Baccini A, Gao F, Schaaf C (2002) Global land cover mapping from MODIS: algorithms and early results. Remote Sens Environ 83(1):287–302

Friedl MA, Sulla-Menashe D, Tan B, Schneider A, Ramankutty N, Sibley A, Huang X (2010) MODIS collection 5 global land cover: algorithm refinements and characterization of new datasets. Remote Sens Environ 114(1):168–182

Soltani SS, Fahs M, Al Bitar A, Ataie-Ashtiani B (2022b) Fully coupled subsurface-land surface hydrological models: a scaling approach to improve subsurface storage predictions, Copernicus Meetings

Furusho-Percot C, Goergen K, Hartick C, Poshyvailo-Strube L, Kollet S (2022) Groundwater model impacts multiannual simulations of heat waves. Geophys Res Lett 49(10):e2021GL096781

He L, Ji S, Xin K, Chen Z, Chen L, Nan J et al (2023) Application of Deep Learning in Drainage Systems Monitoring Data Repair—A Case Study Using Con-GRU Model. Water (Switzerland) 15(8):1–19

Hengl T, Mendes de Jesus J, Heuvelink GBM, Ruiperez Gonzalez M, Kilibarda M, Blagotic A et al (2017) SoilGrids250m: global gridded soil information based on machine learning. PLoS ONE 12:e0169748. https://doi.org/10.1371/journal.pone.0169748

Herbst M, Diekkrüger B, Vanderborght J (2006) Numerical experiments on the sensitivity of runoff generation to the spatial variation of soil hydraulic properties. J Hydrol 326:43–58

Huynh DVK, Renard B, Colleoni F, Monnier J, Roux H (2025) A distributed hybrid physics–AI framework for learning corrections of internal hydrological fluxes in physically based models. Hydrol Earth Syst Sci 29:3589–3610. https://doi.org/10.5194/hess-29-3589-2025

Izadi A, Zarei N, Reza Nikoo M, Al-Wardy M, Yazdandoost F (2024) Exploring the potential of deep learning for streamflow forecasting: a comparative study with hydrological models for seasonal and perennial rivers. Expert Syst Appl 252:124139

Joseph VR (2022) Optimal ratio for data splitting. Stat Anal Data Min 15(4):531–538

Joseph VR, Vakayil A (2022) Split: an optimal method for data splitting. Technometrics 64(2):166–176

Junghänel T, Bissolli P, Daßler J, Fleckenstein R, Imbery F, Janssen W, Kaspar F, Lengfeld K, Leppelt T, Rauthe M, Rauthe-Schöch A, Rocek M, Walawender E, Weigl E.: Hydro-klimatologische Einordnung der Stark- und Dauerniederschläge in Teilen Deutschlands im Zusammenhang mit dem Tiefdruckgebiet "Bernd" vom 12. bis 19. Juli 2021, DWD–Deutscher Wetterdienst, Offenbach, Germany, https://www.dwd.de/DE/leistungen/besondereereignisse/niederschlag/20210721_bericht_starkniederschlaege_tief_bernd.html (last access: 9 May 2022), 22 July 2021

Kollet SJ, Maxwell RM (2006) Integrated surface–groundwater flow modeling: a free-surface overland flow boundary condition in a parallel groundwater flow model. Adv Water Resour 29:945–958

Kordani M, Nikoo MR, Fooladi M, Ahmadianfar I, Nazari R, Gandomi AH (2024) Improving long-term flood forecasting accuracy using ensemble deep learning models and an attention mechanism. J Hydrol Eng 29(6):04024042

Kuffour BN, Engdahl NB, Woodward CS, Condon LE, Kollet S, Maxwell RM (2020) Simulating coupled surface–subsurface flows with ParFlow v3. 5.0: capabilities, applications, and ongoing development of an open-source, massively parallel, integrated hydrologic model. Geosci Model Dev 13(3):1373–1397

Kumar V, Azamathulla HM, Sharma KV, Mehta DJ, Maharaj KT (2023) The state of the art in deep learning applications, challenges, and future prospects: a comprehensive review of flood forecasting and management. Sustainability. https://doi.org/10.3390/su151310543

Le XH, Ho HV, Lee G, Jung S (2019) Application of long short-term memory (LSTM) neural network for flood forecasting. Water. https://doi.org/10.3390/w11071387

Liu PL, Der Kiureghian A (1991) Optimization algorithms for structural reliability. Struct Saf 9(3):161–177

Liu X, Sang X, Chang J, Zheng Y (2021) Multi-model coupling water demand prediction optimization method for megacities based on time series decomposition. Water Resour Manage 35:4021–4041

Marshall SRO, Tran TND, Arshad A, Rahman MM, Lakshmi V. (2025) SWAT and CMIP6-driven hydro-climate modeling of future flood risks and vegetation dynamics in the White Oak Bayou Watershed, United States. Earth Systems and Environment, 1-23

Maina FZ, Siirila-Woodburn ER, Vahmani P (2020) Sensitivity of meteorological-forcing resolution on hydrologic variables. Hydrol Earth Syst Sci 24:3451–3474

Maxwell RM, Miller NL (2005) Development of a coupled land surface and groundwater model. J Hydrometeorol 6:233–247

Maxwell RM, Kollet SJ, Smith SG, Woodward CS, Falgout RD, Ferguson IM, Baldwin C, Bosl WJ, Hornung R, Ashby S (2009) ParFlow user's manual. Int GroundWater Model Cent Rep GWMI 1:129

Meng S, Xie X, Liang S (2017) Assimilation of soil moisture and streamflow observations to improve flood forecasting with considering runoff routing lags. J Hydrol 550:568–579

Madsen HO, Krenk S, Lind NC. (2006) Methods of structural safety, Courier Corporation

Miau S, Hung WH (2020) River flooding forecasting and anomaly detection based on deep learning. IEEE Access 8:198384–198402

Mohr S, Ehret U, Kunz M, Ludwig P, Caldas-Alvarez A, Daniell JE, Ehmele F, Feldmann H, Franca MJ, Gattke C (2022) A multi-disciplinary analysis of the exceptional flood event of July 2021 in central Europe. Part 1: event description and analysis. Nat Hazards Earth Syst Sci Discuss 2022:1–44

Moishin M, Deo RC, Prasad R, Raj N, Abdulla S (2021) Designing deep-based learning flood forecast model with ConvLSTM hybrid algorithm. IEEE Access 9:50982–50993

Munich Re: Hurricanes, cold waves, tornadoes: weather disasters in USA dominate natural disaster losses in 2021 – Europe: extreme flash floods with record losses, Munich Re, Media relations on 10 January 2022: Natural disaster losses 2021, Munich, Germany, https://www.munichre.com/en/company/media-relations/media-information-and-corporate-news/media-information/2022/natural-disaster-losses-2021.html last access: 9 May 2022

Nguyen BQ, Van Binh D, Tran TND, Kantoush SA, Sumi T (2024) Response of streamflow and sediment variability to cascade dam development and climate change in the Sai Gon Dong Nai River basin. Clim Dyn 62(8):7997–8017

Owens, R. G. and Hewson, T (2018) ECMWF forecast user guide, Reading: ECMWF, 10, m1cs7h

Papaioannou I, Straub D (2021) Variance-based reliability sensitivity analysis and the FORM α-factors. Reliab Eng Syst Saf 210:107496

Rácz A, Bajusz D, Héberger K (2021) Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification. Molecules 26(4):1111

Rajab A, Farman H, Islam N, Syed D, Elmagzoub MA, Shaikh A et al (2023) Flood forecasting by using machine learning: a study leveraging historic climatic records of Bangladesh. Water. https://doi.org/10.3390/w15223970

Saadi M, Furusho-Percot C, Belleflamme A, Chen JY, Trömel S, Kollet S (2023) How uncertain are precipitation and peak flow estimates for the July 2021 flooding event? Nat Hazards Earth Syst Sci 23:159–177

Schalge B, Haefliger V, Kollet S, Simmer C (2019) Improvement of surface run-off in the hydrological model ParFlow by a scale-consistent river parameterization. Hydrol Process 33(14):2006–2019

Schröter K, Kunz M, Elmer F, Mühr B, Merz B (2015) What made the June 2013 flood in Germany an exceptional event? A hydro-meteorological evaluation. Hydrol Earth Syst Sci 19:309–327

Soltani S, Belleflamme A, Hammoudeh S, Kollet S (2025a) Enhancing streamflow predictions with a river parameterization in an integrated hydrological model, EGU General Assembly 2025, Vienna, Austria, 27 Apr–2 May 2025, EGU25-11292, https://doi.org/10.5194/egusphere-egu25-11292, 2025a

Soltani SS, Ataie-Ashtiani B, Danesh-Yazdi M, Simmons CT (2020) A probabilistic framework for water budget estimation in low runoff regions: a case study of the central basin of Iran. J Hydrol 586:124898

Soltani SS, Fahs M, Al Bitar A, Ataie-Ashtiani B (2022a) Improvement of soil moisture and groundwater level estimations using a scale-consistent river parameterization for the coupled ParFlow-CLM hydrological model: a case study of the Upper Rhine Basin. J Hydrol 610:127991

Soltani SS (2025) 2 Assimilating GRACE Data. Remote Sensing for Geophysicists, p.21

Soltani SS, Ataie-Ashtiani B, Al Bitar A, Simmons CT, Younes A, Fahs M (2024) Assimilating multivariate remote sensing data into a fully coupled subsurface-land surface model. J Hydrol 641:131812

Soltani SS, Belleflamme A, Georgen K, Kollet S (2025b) Improving real-time flood forecasting: probabilistic validation of assimilated remotely-sensed soil moisture data. Earth Syst Environ J. https://doi.org/10.1007/s41748-025-00726-8

Soltani, S.: Assimilating remote sensing information into a distributed hydrological model for improving water budget predictions, Université de Strasbourg; Sharif University of Technology (Tehran), 2022

Tang Y, Sun Y, Han Z, Soomro S, Wu Q (2023) Journal of hydrology: regional studies flood forecasting based on machine learning pattern recognition and dynamic migration of parameters. J Hydrol Reg Stud 47:101406

Tapas MR, Etheridge R, Finlay CG, Peralta AL, Bell N, Xu Y, Lakshmi V (2024) A methodological framework for assessing sea level rise impacts on nitrate loading in coastal agricultural watersheds using SWAT+: a case study of the Tar-Pamlico River basin, North Carolina, USA. Sci Total Environ 951:175523

Tayfur G, Singh VP, Moramarco T, Barbetta S (2018) Flood hydrograph prediction using machine learning methods. Water (Switzerland) 10(8):1–13

Tramblay Y, Bouaicha R, Brocca L, Dorigo W, Bouvier C, Camici S, Servat É (2012) Estimation of antecedent wetness conditions for flood modelling in northern Morocco. Hydrol Earth Syst Sci 16:4375–4386

Tran TND, Nguyen BQ, Grodzka-Łukaszewska M, Sinicyn G, Lakshmi V (2023) The role of reservoirs under the impacts of climate change on the Srepok River basin, Central Highlands of Vietnam. Front Environ Sci 11:1304845

Wanders N, Karssenberg D, De Roo A, De Jong SM, Bierkens MFP (2014) The suitability of remotely sensed soil moisture for improving operational flood forecasting. Hydrol Earth Syst Sci 18(6):2343–2357

Wegayehu EB, Muluneh FB (2021) Multivariate streamflow simulation using hybrid deep learning models. Comput Intell Neurosci 2021(1):5172658

Xu W, Chen J, Corzo G, Xu CY, Zhang XJ, Xiong L et al (2024) Coupling deep learning and physically based hydrological models for monthly streamflow predictions. Water Resour Res 60(2):1–25

Youssefi F, Soltani SS, Khorrami B, Ali Sh (2025) Integrating fully-coupled hydrological modeling and random forest machine learning model to enhance spatial resolution of GRACE observed water storage across the Rhine basin. Nat Resour Res J. https://doi.org/10.1007/s11053-025-10528-4

Zhang Y, Zhou Z, Van Griensven TJ, Yang SX, Gharabaghi B (2023) Flood forecasting using hybrid LSTM and GRU models with lag time preprocessing. Water (Basel) 15(22):1–18

Zhang Y, Shen C, Appling AP (2024) Enhancing hydrological prediction through physics-informed models: integrating static watershed attributes into LSTM architectures. Water Resour Res 60(5):e2023WR035789. https://doi.org/10.1029/2023WR035789

Zhao X, Wang H, Bai M, Xu Y, Dong S, Rao H et al (2024) A comprehensive review of methods for hydrological forecasting based on deep learning. Water. https://doi.org/10.3390/w16101407

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.