Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Transferable machine learning prediction of urban pluvial flood water depth using multi-resolution digital elevation models

Babak Razdar

babak.razdar@cmcc.it

CMCC: Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici https://orcid.org/0000-0001-8969-2581

Ilenia Manco

CMCC: Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici

Pasquale Schiano

CMCC: Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici

Alfredo Reder

CMCC: Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici

Paola Mercogliano

CMCC: Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici

Research Article

Keywords: Urban pluvial flooding, Random Forest, Flood water depth, Digital elevation model, Model transferability, Uncertainty and confidence analysis

Posted Date: May 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9193376/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Transferable machine learning prediction of urban pluvial flood water depth using multi-resolution digital elevation models

Author 1: Babak Razdar*
CMCC Foundation - Euro-Mediterranean Center on Climate Change, Italy
https://orcid.org/0000-0001-8969-2581
Author 2: Ilenia Manco
CMCC Foundation - Euro-Mediterranean Center on Climate Change, Italy
https://orcid.org/0000-0002-2872-4655
Author 3: Pasquale Schiano
CMCC Foundation - Euro-Mediterranean Center on Climate Change, Italy
https://orcid.org/0000-0002-4439-6444
Author 4: Alfredo Reder
CMCC Foundation - Euro-Mediterranean Center on Climate Change, Italy
https://orcid.org/0000-0002-5782-8170
Author 5: Paola Mercogliano
CMCC Foundation - Euro-Mediterranean Center on Climate Change, Italy
https://orcid.org/0000-0001-7236-010X

***Corresponding author**
E-mail address: babak.radzar@cmcc.it
CMCC Foundation - Euro-Mediterranean Center on Climate Change, Italy

Abstract. Machine learning methods, particularly Random Forest (RF), present an effective solution for predicting urban pluvial flood water depth, especially in data-scarce regions where traditional hydrological and hydraulic observations are lacking. This study proposes a DEM-centric RF framework for predicting spatially continuous urban flood water depth across various precipitation return periods, emphasizing its robustness to terrain resolution degradation and spatial transferability. The model is trained and validated in Milan using a 1 m LiDAR-derived DEM and evaluated with a coarser 5 m EU-DEM to assess sensitivity to reduced resolution. Transferability is tested by applying the trained model to Cologne using independent 1 m LiDAR data. Model performance is assessed across five precipitation scenarios (5–100-year return periods) using absolute error metrics, tolerance thresholds, spatial diagnostics, and confidence-stratified analyses. Results show stable performance, with mean absolute errors below 0.30 m and over 58% of grid cells within ±0.25 m of reference flood depths. Prediction errors are mainly found in hydraulically complex areas, while the majority of the urban domain shows well-constrained predictions. Confidence diagnostics indicate that high- and very-high-confidence predictions comply with tolerance thresholds under both

DEM resolution degradation and spatial transfer. These findings suggest that relative terrain structure governs urban pluvial flooding patterns, enabling reliable flood depth predictions with medium-resolution DEMs. This framework offers a scalable tool for urban flood screening, hazard assessment, and climate adaptation planning in areas with limited data.

Keywords: Urban pluvial flooding; Random Forest; Flood water depth; Digital elevation model; Model transferability; Uncertainty and confidence analysis

1. Introduction

The frequency and severity of natural disasters have increased in recent decades, a trend widely attributed to climate change, rapid urbanisation, and anthropogenic land-use change. Among these hazards, flooding remains the most pervasive and destructive, causing extensive damage to infrastructure, disrupting economic activities, and posing significant risks to human life. Global assessments consistently report rising flood-related losses, with urban areas experiencing disproportionate impacts due to high population density, extensive impervious surfaces, and complex drainage systems (IPCC, 2023; Rentschler et al., 2022; Dottori et al., 2018). Urban flooding is characterised by rapid runoff generation, limited infiltration capacity, and localised water accumulation, often triggered by short-duration, high-intensity rainfall events. These processes are expected to intensify under projected climate change scenarios, which indicate an increase in the frequency and magnitude of extreme precipitation events across many regions (Donat et al., 2016; Westra et al., 2014). Enhancing the capability to predict flood dynamics in urban environments has therefore become a critical priority for disaster risk reduction and climate adaptation planning.

Accurate prediction of flood water depth is particularly important for urban flood risk management. Unlike binary flood extent mapping, water depth provides quantitative information essential for emergency response, infrastructure vulnerability assessment, and damage estimation (Apel et al., 2004; Merz et al., 2010; Büchele et al., 2006). However, reliable prediction of urban flood water depth remains challenging due to the highly nonlinear nature of flood processes. In addition, the limited availability of high-resolution hydrological and hydraulic data constrains robust model calibration and validation (Teng et al., 2017; Costabile et al., 2019; Apel et al., 2009; Liu et al., 2025).

These challenges are further compounded by the scarcity of monitoring networks capable of measuring spatially distributed flood water depths in most cities, particularly in rapidly developing or resource-constrained regions. As a result, empirical evaluation and independent validation of flood models remain limited, reinforcing the need for alternative modelling strategies that can operate under data-scarce conditions (Liu et al., 2025; Mosavi et al., 2018).

Traditionally, physically-based hydrological and hydrodynamic models, such as *SWMM* (U.S. Environmental Protection Agency, 2017), *HEC-HMS* (U.S. Army Corps of Engineers, 2016), *HEC-RAS* (U.S. Army Corps of Engineers, 2016), and *MIKE URBAN* (DHI, 2017), have been employed to simulate urban flood processes. have been employed to simulate urban flood processes. These models are grounded in depth-averaged continuity and momentum equations that explicitly enforce mass and momentum conservation during flood propagation (Bates et al., 2010). When supported by detailed input data, rigorous calibration, and accurate representation of drainage infrastructure, such models can reproduce flood dynamics with considerable physical realism, particularly for complex inundation and flow processes (Neal et al., 2012; Schumann et al., 2018; Costabile et al., 2019). However, when the spatial resolution of the underlying digital elevation model (DEM) or digital terrain model (DTM) is reduced, hydraulic models become increasingly sensitive to terrain quality, particularly where small-scale conveyance features such as canals, culverts, and urban drainage pathways are not explicitly represented or are smoothed by the coarser topographic surface (Sanders, 2007; Cook & Merwade, 2009; Teng et al., 2017). In practice, such hydraulic features are frequently missing or poorly resolved in commonly available terrain datasets. This issue can cause physically-based hydraulic simulations to exhibit non-physical behaviour, including artificial flow blocking, unrealistic ponding, or unintended release of water from portions of the computational domain (Sanders, 2007; Cook

& Merwade, 2009; Teng et al., 2017). Reducing these effects often requires substantial terrain preprocessing, including manual inspection, drainage enforcement, and localised modification of the DTM, which is time-consuming and data-intensive, especially in urban and semi-urban catchments (Cook & Merwade, 2009; Teng et al., 2017). In addition to these preprocessing demands, broader application of physically-based models is commonly constrained by high computational cost, sensitivity to parameter uncertainty, and reduced efficiency when applied to large or complex domains (Neal et al., 2012; Teng et al., 2017). These limitations restrict their suitability for rapid flood assessment, large-scale screening, and near-real-time applications.

In response to these constraints, machine learning (ML) approaches have emerged as efficient alternatives or complementary tools to physically based models in flood research, particularly under conditions of limited hydrological data availability (Mosavi et al., 2018). ML algorithms are capable of learning complex relationships directly from data and can provide rapid predictions once trained, even in data-scarce environments (Shen, 2018; Mosavi et al., 2018; Yaseen, 2021; Bowes et al., 2022). Rather than explicitly solving the governing hydraulic equations, ML models infer empirical relationships between input features, such as precipitation, topography, and land cover, and flood response variables. Flood characteristics, including inundation depth, are commonly predicted on a grid-cell basis without requiring explicit representation of hydraulic connectivity or detailed drainage networks (Shirzadi et al., 2020). Consequently, ML-based approaches reduce dependence on explicit representation of fine-scale hydraulic connectivity and can provide practically useful flood estimates even where local flow pathways are incompletely captured, supporting large-area screening applications (Teng et al., 2017).

ML algorithms are commonly categorised into supervised learning, unsupervised learning, and reinforcement learning (Mosavi et al., 2018; James et al., 2021; Criminisi et al., 2019). Within supervised learning, which dominates most flood prediction and flood hazard applications, deep learning (DL) and ensemble learning represent the two most widely adopted methodological families (Mosavi et al., 2018; Mudashiru et al., 2021; Bentivoglio et al., 2022; Chitwatkulsiri & Miyamoto, 2023). DL models are capable of learning hierarchical feature representations and capturing complex spatial and temporal dependencies when large training datasets are available (Bentivoglio et al., 2022; Zhang et al., 2022). However, their performance is often constrained by high data requirements, computational cost, and model complexity, which limit their applicability in heterogeneous and data-scarce urban flood environments (Mosavi et al., 2018; Nearing et al., 2021; Mudashiru et al., 2021). Ensemble learning methods address these limitations by integrating multiple weak learners to enhance robustness, reduce overfitting, and better manage nonlinear relationships and mixed-type predictors (Belgiu & Drăguț, 2016; Mudashiru et al., 2021). Among these approaches, Random Forest (RF) has been extensively applied in flood-related studies due to its ability to efficiently handle high-dimensional and heterogeneous datasets while maintaining stable performance under limited or noisy training data (Mosavi et al., 2018; Mudashiru et al., 2021). RF models have demonstrated strong predictive accuracy and interpretability in urban flood hazard and inundation studies, often outperforming single decision trees and other classical ML algorithms (Belgiu & Drăguț, 2016; Norallahi & Kaboli, 2021; Seleem et al., 2023; Bentivoglio et al., 2022). Despite these advantages, RF-based applications for spatially continuous urban flood water depth prediction remain relatively limited, with most studies focusing on flood susceptibility mapping, inundation extent delineation, or discrete depth classification (Mosavi et al., 2018; Bentivoglio et al., 2022). Moreover, many ML-based flood depth

models rely on hydrological inputs such as rainfall or discharge data that are often unavailable or highly uncertain in urban settings, limiting model transferability (Merz et al., 2014; Tehrany et al., 2015; Nearing et al., 2021). Recent advances highlight the potential of digital DEM-derived terrain attributes as physically meaningful predictors that capture runoff routing and water accumulation patterns without extensive hydrological calibration, making them particularly suitable for data-scarce urban contexts (Samela et al., 2018; Costabile et al., 2019; El Baida et al., 2024). To address this gap, this study proposes an RF-based framework for urban flood water depth prediction that emphasises DEM-derived hydrological parameters under low-data conditions. The proposed approach aims to balance physical relevance and computational efficiency, offering a practical alternative to computationally intensive hydrodynamic simulations. By focusing on topography-driven predictors and ensemble learning, the framework seeks to enhance the transferability and scalability of urban flood water depth modelling. Unlike most existing data-driven urban flood studies, which are calibrated for single cities or rely on hydrological forcing and location-specific retraining, this study explicitly evaluates both DEM-resolution robustness and cross-city model transferability within a unified modelling framework.

The specific objectives of this study are to:

- **i.** derive and analyse key DEM-based hydrological parameters relevant to urban flood processes;
- **ii.** develop a Random Forest regression model for predicting spatially distributed urban flood water depth;
- **iii.** evaluate model performance using standard statistical and spatial metrics under limited hydrological data availability; and
- **iv.** assess the relative importance of topographic predictors in controlling flood depth variability.

To address these objectives, the proposed framework integrates DEM-derived hydrological and terrain predictors with RF regression to model spatially continuous urban pluvial flood water depth under multiple precipitation scenarios. Model robustness to terrain resolution is examined by comparing predictions obtained using high-resolution LiDAR (Light Detection and Ranging) and coarser EU-DEM data, while spatial transferability is evaluated through application of the trained model, without retraining, to an independent urban case study.

The paper is structured as follows. Section 2 describes the study areas, data sources, DEM-derived predictors, and the overall modelling framework. Section 3 presents model performance, robustness to DEM resolution degradation, and cross-city transferability results, including uncertainty and confidence-based analyses. Section 4 discusses the implications of the findings, model limitations, and applicability for urban flood risk assessment. Section 5 concludes the paper and outlines perspectives for future research.

2. Materials and Methods

2.1. Conceptual framework

The conceptual workflow applied in this study is shown in **Figure 1**. It targets urban environments where detailed hydrometric observations are scarce, and flood information is primarily available through scenario-based modelling outputs.

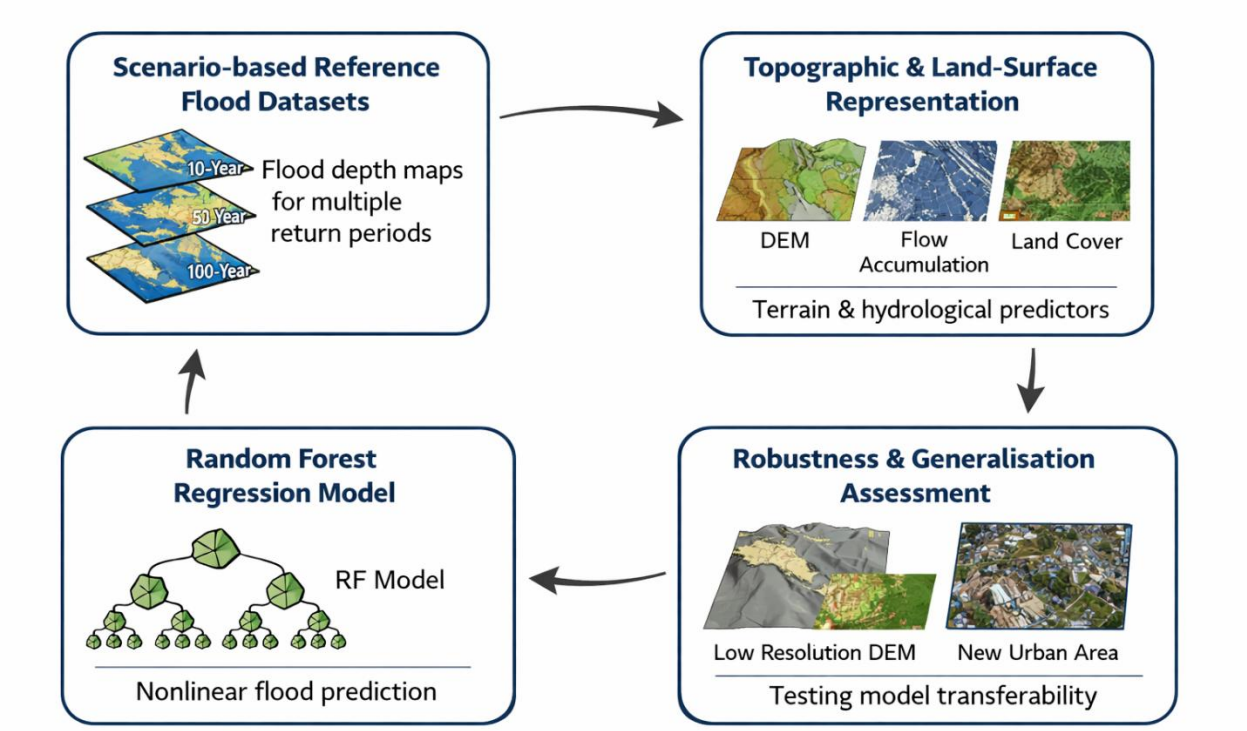


Figure 1. Framework for data-driven flood prediction

Specifically, it consists of four interconnected components:

1. **Scenario-based reference flood datasets.** Spatially explicit flood water depth maps for multiple precipitation return periods serve as learning targets. These datasets provide realistic representations of flood behaviour across a range of rainfall severities, forming the basis for supervised model training.
2. **Topographic and land-surface representation.** DEM-derived hydrological and terrain attributes are extracted to characterise local surface morphology, runoff routing tendencies, and potential storage areas. These predictors capture physically meaningful controls on flood accumulation without explicitly solving hydraulic equations.
3. **Random Forest regression module.** The RF model learns nonlinear relationships between terrain descriptors, rainfall severity, and flood water depth. Its ensemble structure captures complex interactions among predictors while ensuring robustness under heterogeneous terrain inputs and sparse hydrological data.
4. **Robustness and generalisation assessment.** We evaluate the stability of terrain–flood relationships under varying data conditions, including reduced terrain resolution and application to independent urban areas. This component tests whether relative morphological structure provides transferable predictive information beyond the original calibration context.

2.2. Reference flood hazard dataset and case study selection

To implement and evaluate the proposed framework, its application has been restricted to a subset of case studies selected based on data availability and suitability for supervised training of the RF model. Specifically, flood water depths used for training were obtained from a dataset developed within the C3S Pluvial Flood Risk Assessment in Urban Areas Service (https://climate.copernicus.eu/disaster-risk-reduction) by Mercogliano et al. (2021). The data was presented in Essenfelder et al. (2022) for further reference. The dataset provides probabilistic flood hazard maps for five precipitation return periods (5, 10, 25, 50, and 100 years) over 20 European cities. These maps were generated using the Safer_RAIN model (Samela et al., 2020; Mediero et al., 2022), a simplified raster-based filling-and-spilling approach operating on high-resolution DEMs. The model simulates rainfall accumulation and overflow within hierarchically organised surface depressions under prescribed design storms. It adopts a hydrostatic formulation, accounting for terrain-controlled storage and infiltration processes represented through a pixel-based Green-Ampt scheme (Green and Ampt, 1911). Among the cities available within the dataset, we have selected Milan (Italy) and Cologne (Germany) for implementation of the framework. The choice has been guided by two criteria: (i) morphological similarity and low-relief urban structure, which favour terrain-controlled flood dynamics, and (ii) the availability of DEMs at multiple spatial resolutions, including high-resolution LiDAR data at 1 m and the coarser EU-DEM at 5 m. This dual-resolution availability enables systematic assessment of model robustness to terrain resolution degradation while maintaining consistent reference flood information.

The geographic locations and topographic characteristics of the two case study areas are shown in **Figure 2**.

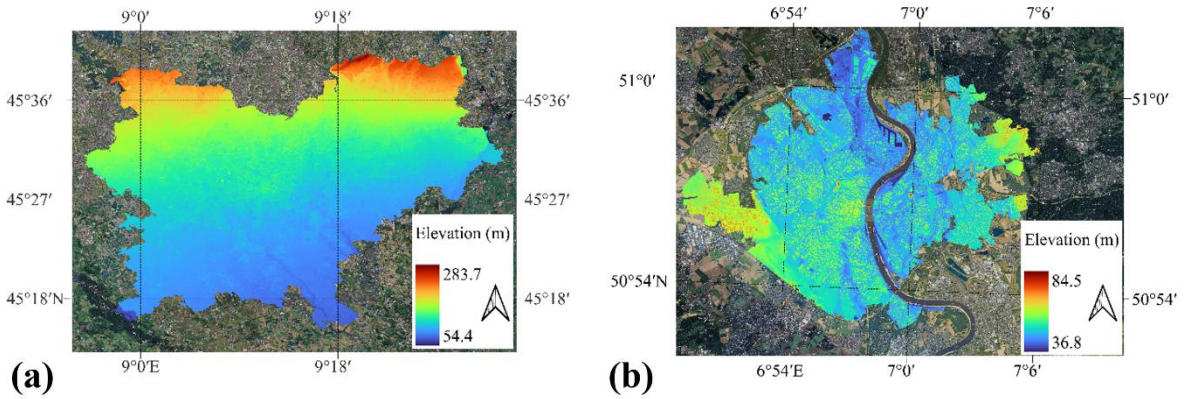


Figure 2. Geographic location and digital elevation models (DEMs) of the two case study areas: (a) Milan (Italy) and (b) Cologne (Germany).

In brief, Milan, located in northern Italy, is one of the country's most densely urbanised metropolitan regions. The administrative area covers approximately 180 km², with elevations ranging from about 98 m to 171 m above mean sea level and a mean elevation of approximately 122 m. The city is characterised by very low topographic relief, which enhances surface water accumulation during short-duration, high-intensity rainfall events and increases sensitivity to subtle elevation gradients and micro-topographic controls on urban ponding. On the other side, Cologne, located in western Germany, covers approximately 143 km², with elevations ranging from about 36.8 m to 84.5 m above mean

sea level and a mean elevation of approximately 51.4 m. Similar to Milan, Cologne exhibits relatively low relief and extensive urban development, but differs in urban layout and drainage configuration. Its morphological comparability, combined with structural differences, provides a suitable setting for evaluating the spatial transferability of the terrain-driven flood water depth prediction framework.

2.3. **Explanatory variables (predictors)**

Explanatory variables (or predictors) were derived exclusively from digital terrain and land-use information. The modelling strategy intentionally minimises reliance on traditional hydrometric observations by using (i) DEM-derived hydrological and terrain parameters, (ii) land-use information, and (iii) scenario-based precipitation inputs corresponding to different return periods.

2.3.1. **DEM-derived hydrological and terrain parameters**

All explanatory terrain variables were derived from DEMs without using absolute elevation as a direct predictor. Instead, relative and process-oriented terrain indices were extracted to represent topographic controls on runoff routing, flow convergence, and surface water accumulation. This choice is particularly relevant in low-relief urban environments, where subtle elevation differences and micro-topographic features strongly influence flood dynamics. All terrain derivatives were computed from the DEM using standard terrain and hydrological analysis tools implemented in QGIS, ensuring methodological consistency across study areas and DEM resolutions (QGIS Development Team, 2024). The following DEM-derived parameters were used as input features:

1) Slope: it represents the local gradient of the terrain and directly influences surface runoff velocity and drainage efficiency (Moore et al., 1991). In flat urban environments, even subtle slope variations can strongly affect surface water retention, flow pathways, and localised ponding during intense rainfall events (Costabile et al., 2019).

2) Topographic Position Index (TPI): it quantifies the relative elevation of a grid cell with respect to its surrounding neighbourhood, allowing the identification of ridges, valleys, and local depressions (Weiss, 2001). TPI has been widely used to characterise micro-topographic controls on surface water convergence and divergence, which are critical for urban flood accumulation processes (Samela et al., 2018).

3) Terrain Ruggedness Index (TRI): it measures terrain heterogeneity by calculating elevation differences between neighbouring cells (Riley et al., 1999). In urban areas, TRI captures surface roughness introduced by both natural and anthropogenic features, influencing flow resistance and contributing to spatial variability in flood water depth patterns (Costabile et al., 2019).

4) Topographic Wetness Index (TWI): it combines local slope and upstream contributing area to describe the spatial tendency for water accumulation (Beven & Kirkby, 1979; Moore et al., 1991). Higher TWI values indicate zones of flow convergence and potential surface saturation, making it a useful proxy for identifying flood-prone locations in urban landscapes (Tehrany et al., 2015).

5) Topographic depressions (sink areas): they represent local sink areas where surface runoff accumulates and is temporarily stored prior to overflow. These features are particularly influential in flat urban terrain and play a key role in controlling localised inundation and urban ponding (Lindsay & Creed, 2005; Leitão et al., 2009). Depression mapping was performed using sink identification and hydrological preprocessing tools available in QGIS.

2.3.2. **Land-use information**

In addition to DEM-derived predictors, land-use information was incorporated to account for the influence of surface characteristics on runoff generation and the spatial variability of flood water depth. Land-use data were obtained from the CORINE Land Cover (CLC) 2018 dataset provided by the Copernicus Land Monitoring Service, which offers harmonised land-use and land-cover information across Europe at 100 m spatial resolution with a six-year update cycle and high thematic consistency (Copernicus Land Monitoring Service, 2018). The same land-use dataset was used for both the Milan and Cologne case studies, ensuring methodological consistency and enabling a controlled assessment of model transferability across different urban contexts. Land-use classes were included as categorical predictors to represent differences in imperviousness, surface roughness, and infiltration capacity across urban fabric, transportation infrastructure, green spaces, and open areas. The importance of land-use and land-cover data in urban flood modelling is well recognised, as such data capture anthropogenic controls on flood response that are not represented by terrain information alone (Lin et al., 2019; Islam et al., 2025). In addition, hybrid approaches combining machine learning with GIS-derived land-cover predictors have demonstrated improved performance in flood susceptibility and risk assessments (Wang et al., 2019). The integration of land-use data, therefore, complements terrain-derived predictors by accounting for both natural and anthropogenic influences on urban flood water depth patterns.

2.3.3. **Scenario-based precipitation inputs**

The precipitation inputs consist of event-based design storm scenarios associated with fixed return periods (5, 10, 25, 50, and 100 years). These scenarios were used as inputs for the reference flood hazard maps developed by Essenfelder et al. (2022) and serve as the rainfall forcing for the Random Forest model. They are derived from intensity-duration-frequency (IDF) curves based on the ERA5@2km dataset, obtained through convection-permitting dynamical downscaling (≈2.2 km) of the ERA5 reanalysis for the period 1989–2018 (Reder et al., 2022). This high-resolution dataset enables improved representation of extreme hourly precipitation over urban areas compared to coarse-resolution reanalysis products. The resulting design rainfall inputs provide spatially coherent and physically consistent extreme precipitation estimates for probabilistic pluvial flood hazard assessment at the city scale.

2.4. RF model structure and training strategy

RF model was employed to reproduce spatial flood water depth using a combination of DEM-derived, land-use, and precipitation-based predictors (see §2.3). In this study, the RF model was trained using a 1 m LiDAR-derived DEM for Milan, together with the corresponding flood water depth outputs and precipitation scenarios.

The flood water depth outputs correspond to scenario-based pluvial flood conditions associated with distinct rainfall severities. These design rainfall scenarios produce spatially varying flood water depth patterns that reflect differences in rainfall intensity and flood extent across return periods, while ensuring consistency across cities and scenarios. The ensemble of scenario-based flood depth outputs was used jointly to train the RF model. This enabled the model to learn flood water depth variability across a wide range of rainfall severities while requiring minimal additional hydrological input or temporal rainfall information.

Given the very large dataset size (approximately 55.2 million grid cells at 1 m spatial resolution), the data were randomly partitioned into 70% for training (38,656,294 samples) and 30% for validation (16,566,984 samples). The random data split was used exclusively for internal validation within the Milan domain, allowing assessment of model performance on unseen samples drawn from the same spatial distribution as the training data. External spatial transferability was evaluated independently through application of the trained model to the Cologne case study, as described in Section 2.6. These two evaluation procedures serve distinct purposes: the former assesses internal predictive reliability, whereas the latter evaluates out-of-domain generalisation.

A 70/30 split is widely adopted in machine learning applications involving large datasets, as it provides sufficient data for robust model learning while retaining an independent subset for reliable performance evaluation (Hastie et al., 2009; Kuhn & Johnson, 2013). In flood-related and geospatial machine learning studies, this partitioning strategy has been shown to offer a stable balance between predictive accuracy and generalisation capability without introducing unnecessary variance in validation results (Mosavi et al., 2018).

To assess the influence of ensemble size on model behaviour, RF models were trained using 50, 100, and 150 decision trees. Following this sensitivity analysis, an ensemble size of 100 trees was selected for subsequent experiments as a compromise between predictive accuracy, computational efficiency, and model stability. Model training employed bootstrap sampling and random feature selection to enhance ensemble diversity and reduce variance, consistent with the original Random Forest formulation (Breiman, 2001). At each split, a random subset of predictors equal to the square root of the total number of features was considered, which is standard practice for RF regression (Hastie et al., 2009; James et al., 2021). A minimum leaf size of five samples was imposed to limit excessive tree complexity and reduce sensitivity to noise, which is particularly important for large, high-resolution geospatial datasets (Genuer et al., 2010; Probst et al., 2019). All trees were trained using bootstrap-resampled subsets of the training data.

Model training was parallelised on a high-performance computing (HPC) system. This exploited the inherently independent and parallelisable structure of Random Forest ensembles, in which individual decision trees can be trained independently without inter-tree communication (Liaw & Wiener, 2002; Criminisi et al., 2012). All RF training and prediction experiments were performed on the CMCC Juno HPC cluster using the LSF workload manager. Model training was executed as single-node CPU jobs with memory allocations ranging from 120 to 170 GB. For the full Milan LiDAR dataset (~55 million grid cells), training required approximately 2.2 hours for 50 trees (peak memory ~82 GB), 4.5 hours for 100 trees (peak memory ~120 GB), and 6.9 hours for 150 trees (peak memory ~157 GB), demonstrating near-linear scaling of computational cost with ensemble size. Validation and spatial prediction for a single precipitation scenario required approximately 2 hours at 1 m resolution and less than 30 minutes when using the 5 m DEM, with peak memory usage of approximately 100–105 GB. Application of the trained model to Cologne exhibited similar runtime and memory characteristics. These results indicate that, although high-resolution city-scale flood prediction is memory-intensive, the computational burden remains manageable on modern single-node HPC systems without requiring distributed computing or GPU acceleration.

2.5. DEM resolution and model transferability experiments

To evaluate robustness under realistic data constraints, the trained framework was also applied using a coarser 5 m EU-DEM for Milan. This resolution experiment quantifies the sensitivity of the proposed approach to terrain-data

degradation, which is particularly relevant where high-resolution LiDAR is unavailable. Model transferability was evaluated by applying the Milan-trained RF model to the city of Cologne (Germany) using predictors derived from a 1 m LiDAR DEM and flood water depth outputs for the same five precipitation return periods. To ensure a rigorous test of spatial generalisation, the model was transferred without retraining, recalibration, hyperparameter adjustment, or fine-tuning; the predictor set, ensemble configuration, and model parameters were kept identical to those used in Milan. This design provides a strict out-of-domain assessment of whether transferable terrain–flood relationships are sufficient to reproduce spatial flood water depth patterns in an independent urban setting.

Previous studies have shown that flood modelling performance can be sensitive to DEM resolution, particularly in low-relief urban environments where micro-topographic features affect flow routing and ponding (Sanders, 2007; Cook & Merwade, 2009; Leitão et al., 2009; Teng et al., 2017). In addition, recent work on data-driven flood prediction has highlighted challenges in transferring models across cities with differing morphology, drainage configuration, and terrain structure (Seleem et al., 2023; Do Lago et al., 2023). By explicitly testing both DEM resolution degradation and cross-city transfer within a unified framework, this study quantifies the robustness and limitations of a DEM-centric machine learning approach and supports its use for regional screening and comparative urban flood hazard assessment.

2.6. Model interpretability and feature contribution analysis

Beyond predictive performance, the internal behaviour of the RF model was analysed to assess predictor relevance, physical consistency, and the robustness of inferred controls on flood water depth. Feature contribution was evaluated using a combination of SHAP (SHapley Additive exPlanations) values and permutation importance. These are widely adopted explainable artificial intelligence (XAI) techniques for interpreting complex, non-linear machine learning models and mitigating the inherent black-box nature of ensemble approaches (Lundberg & Lee, 2017; Molnar, 2022). In flood-related and geospatial machine learning studies, the joint use of SHAP and permutation importance has become increasingly important for enhancing model transparency, identifying dominant physical drivers, and supporting defensible, decision-relevant interpretations beyond predictive accuracy alone (Molnar et al., 2022; Gulshad et al., 2024; Pradhan et al., 2023; Choubin et al., 2025). SHAP quantifies the average marginal contribution of each predictor to model predictions across all samples using a game-theoretic framework, enabling consistent interpretation in the presence of non-linearities and feature interactions (Lundberg & Lee, 2017; Molnar, 2020). Global SHAP importance was computed as the mean absolute Shapley value across all grid cells and normalised to percentage contributions. For a given feature *i*, the global SHAP importance was computed as:

$$I_i^{SHAP} = \frac{1}{N}\sum_{n=1}^{N} | \phi_i(x_n) | \quad (1)$$

where $\phi_i(x_n)$ is the Shapley value of feature *i* for sample *n*. To facilitate comparison among predictors, SHAP importance percentages were obtained by normalising by the total SHAP contribution:

$$P_i^{SHAP} = 100 \times \frac{I_i^{SHAP}}{\sum_{j=1}^{M} I_j^{SHAP}} \quad (2)$$

To complement SHAP-based contribution analysis and assess the sensitivity of model performance to individual predictors, permutation importance was computed by randomly permuting each predictor and quantifying the resulting reduction in model performance (Breiman, 2001; Fisher et al., 2019; Molnar et al., 2020). For feature *i*, permutation importance was defined as:

$$I_i^{perm} = \frac{1}{R}\sum_{r=1}^{R}[L(y, f(X^{\pi_{i,r}})) - L(y, f(X))] \quad (3)$$

where $X^{\pi_{i,r}}$ denotes the dataset with feature *i* permuted in repetition *r*, $L(\cdot)$ is the loss function, and *R* is the number of permutation repetitions. To facilitate comparison among predictors, permutation importance values were normalised by their sum:

$$R_i^{perm} = \frac{I_i^{perm}}{\sum_{j=1}^{M} I_j^{perm}} \quad (4)$$

3. Results

3.1. Sensitivity of model performance to ensemble size

The sensitivity of model performance to RF ensemble size was evaluated prior to subsequent spatial and transferability analyses. RF models were trained using 50, 100, and 150 decision trees. With 50 trees, the model achieved a training coefficient of determination of R^2 = 0.788 and a validation R^2 of 0.623. Increasing the ensemble size to 100 trees produced a marginal improvement in validation performance (R^2=0.626), while training performance remains comparable (R^2=0.780). However, further increasing the number of trees to 150 led to a reduction in both training and validation performance, with R^2 values decreasing to 0.686 and 0.606, respectively. In theory, increasing the number of trees in an RF ensemble reduces variance and improves prediction stability by averaging across a larger number of decorrelated decision trees (Breiman, 2001). However, empirical evidence indicates that accuracy gains often saturate beyond a certain ensemble size, particularly for very large and information-rich datasets (Probst et al., 2019). The results obtained here confirm that increasing model complexity does not necessarily enhance generalisation performance under such conditions. The negligible improvement observed when increasing the ensemble size from 50 to 100 trees, combined with the degradation in validation accuracy at 150 trees, suggests that excessive ensemble size may amplify noise and redundancy rather than extract additional predictive information. Similar behaviour has been reported in large-scale environmental and geospatial applications, where overly complex models can reduce validation performance despite increased computational effort (Genuer et al., 2010; Probst et al., 2019). Based on these findings, an ensemble size of 100 trees was selected as the optimal configuration, representing the best compromise between predictive accuracy, computational efficiency, and model stability. **Table 1** summarizes model performance metrics for each ensemble size, including training and validation error, dataset size, training runtime, and peak memory usage.

Table 1 summarises model performance metrics for each ensemble size, including training and validation error, dataset size, training runtime, and peak memory usage.

Ensemble Size	Dataset	MSE	RMSE	MAE	R^2	Median Absolute Error	Training Time	Peak Memory
50 trees	Train	0.08	0.29	0.14	0.78	0.06	~2.2 hrs	82 GB
	Test	0.14	0.37	0.19	0.63	0.08	~7hrs	55 GB
100 trees	Train	0.08	0.28	0.14	0.78	0.06	~4.5 hrs	120 GB
	Test	0.14	0.37	0.19	0.63	0.08	~7hrs	105 GB
150 trees	Train	0.08	0.28	0.14	0.78	0.06	~6.9 hrs	157 GB
	Test	0.14	0.37	0.19	0.63	0.08	~7hrs	159 GB

3.2. **Flood water depth prediction for Milan (Italy) under different precipitation scenarios**

The performance of the trained RF model was evaluated for the Milan case study across five precipitation return period scenarios (5, 10, 25, 50, and 100 years). For each scenario, spatially continuous flood water depth was predicted at a 1 m resolution. Model predictions were evaluated against high-resolution reference flood simulations covering approximately 29.0 million grid cells per scenario. **Table 2** summarises the statistical and tolerance-based performance metrics of the RF model across the five precipitation scenarios, providing a quantitative assessment of model behaviour under progressively increasing rainfall severity.

Table 2. Statistical and tolerance-based performance of the Random Forest model for flood water depth prediction in Milan under different precipitation scenarios.

Scenario	Precipitation (mm/h)	R^2	RMSE (m)	MAE (m)	Bias (m)	Accuracy ±0.10 m (%)	Accuracy ±0.25 m (%)	Accuracy ±0.50 m (%)
5-YRP	33.4	0.38	0.31	0.15	0.02	63.83	85.40	93.73
10-YRP	38.5	0.40	0.33	0.17	0.05	57.03	81.76	92.48
25-YRP	45.0	0.44	0.36	0.18	0.04	55.62	79.39	91.10
50-YRP	49.9	0.42	0.40	0.20	0.05	55.20	77.53	89.67
100-YRP	54.7	0.42	0.43	0.22	0.05	54.23	76.02	88.40

The coefficient of determination (R^2) ranges from 0.38 to 0.44, with the highest value obtained for the 25-year return period event. Tolerance-based metrics are more informative than global R^2 for operational flood-depth assessment. Root mean square error (RMSE) and mean absolute error (MAE) increase progressively with increasing event severity, reflecting the larger magnitude and spatial variability of flood depths associated with more extreme rainfall. Bias values remain consistently low for all scenarios, indicating the absence of systematic over- or under-estimation of flood water depth. The moderate R^2 values observed here (0.38–0.44) are consistent with previous high-resolution urban flood modelling studies, where predictive skill is constrained by unresolved micro-topography, limited representation of drainage infrastructure, and complex localised hydraulic interactions (Fewtrell et al., 2011; Schubert & Sanders, 2012; Ochoa-Rodriguez et al., 2015; Teng et al., 2017). In addition, flood water depth distributions at 1 m resolution are strongly right-skewed, with a large proportion of shallow or near-zero values and comparatively sparse extreme depths. Under such spatially heteroscedastic and heavy-tailed conditions, where variance is dominated by a small fraction of deep inundation cells, variance-based metrics such as R^2 become highly sensitive to attenuation of

extreme values. Ensemble regression models such as Random Forest reduce variance through aggregation of decorrelated trees (Breiman, 2001; Hastie et al., 2009), which smooths sharp gradients and contracts predictions toward central tendencies at the distribution tails. Consequently, moderate R^2 values are statistically expected in highly skewed flood-depth domains, even when absolute error metrics and tolerance-based agreement indicate strong practical predictive performance.

Tolerance-based accuracy thresholds (±0.10 m, ±0.25 m, and ±0.50 m) were additionally evaluated to assess agreement with reference simulations from an operational and emergency-response perspective. They better reflect the accuracy requirements of hazard mapping and risk management applications (Hunter et al., 2005; Bates et al., 2014; Teng et al., 2017). Accuracy under the strictest tolerance (±0.10 m) decreases modestly with increasing return period, reflecting the growing influence of fine-scale depth variability under more extreme rainfall conditions. In contrast, agreement within ±0.25 m and ±0.50 m remains consistently high across all scenarios, indicating robust model performance at practically relevant tolerance levels. These trends suggest that increasing rainfall intensity primarily affects the model's ability to reproduce fine-scale depth variations, while the overall magnitude and spatial structure of flooding remain well captured. **Figure 3** provides a comparison of tolerance-based accuracy across all scenarios. It summarises model agreement at the three tolerance thresholds and highlights systematic trends with increasing event severity.

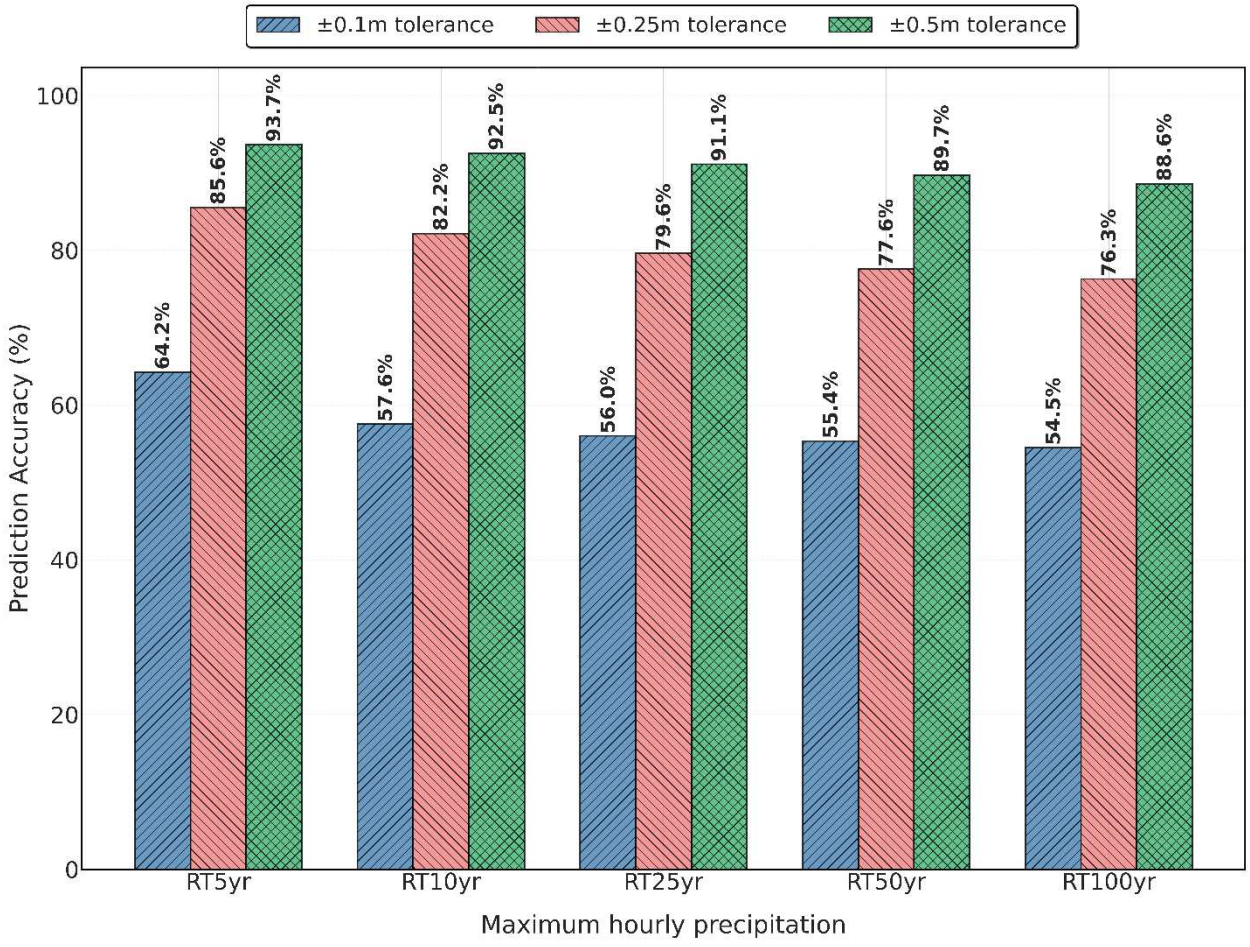


Figure 3. Tolerance-based accuracy of RF-predicted flood water depth across maximum hourly precipitation scenarios (see **Table 2**) in Milan.

While tolerance-based metrics quantify predictive accuracy, they do not directly describe how flood depth magnitude and spatial extent evolve with rainfall severity. **Figure** 4 illustrates the mean reference and RF-predicted flood water depths across Milan for all precipitation scenarios. The figure shows the average values of both observed and predicted flood depths for each scenario across the entire urban domain. The vertical line at each point represents the standard deviation, indicating the variability in flood depths for each scenario.

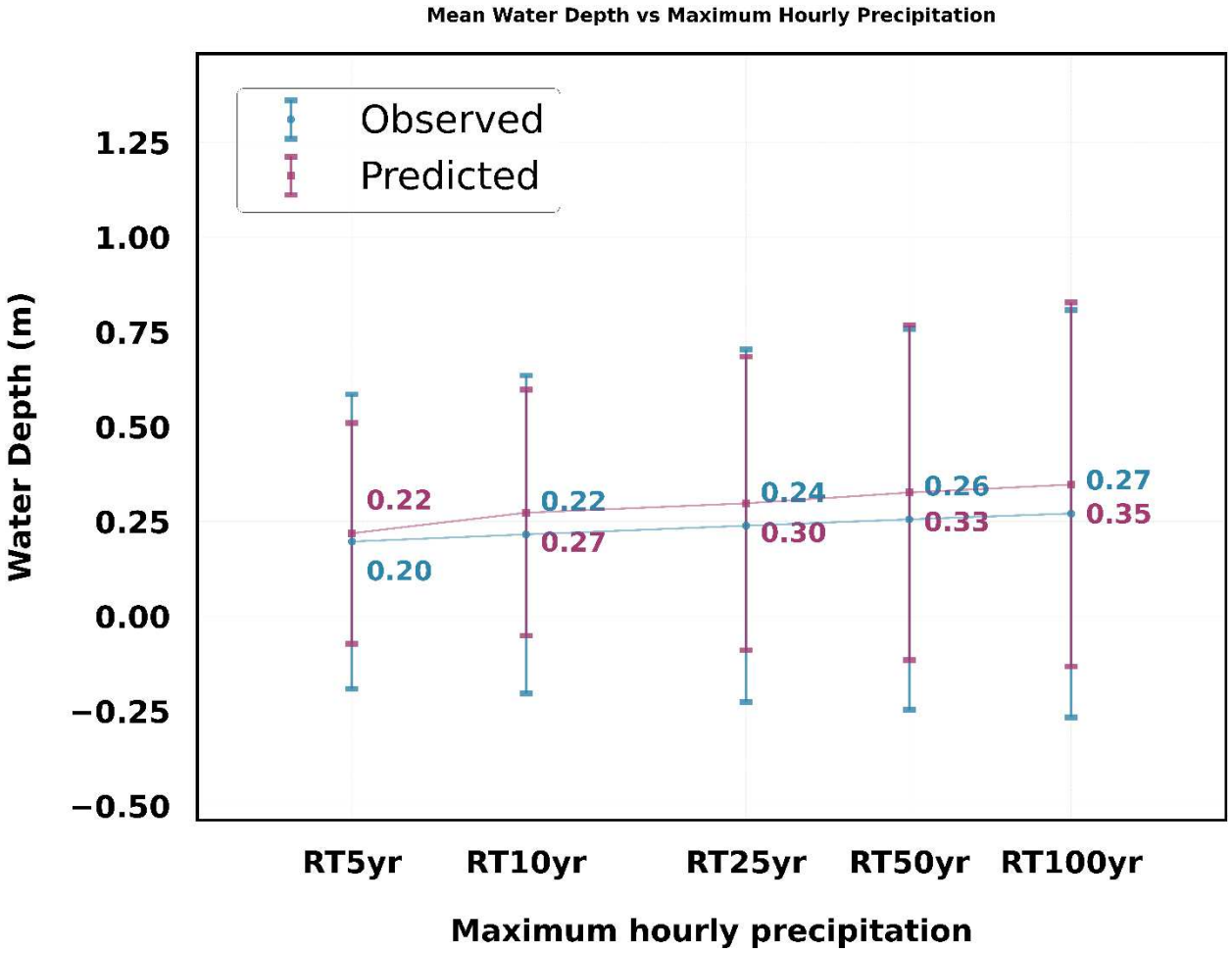
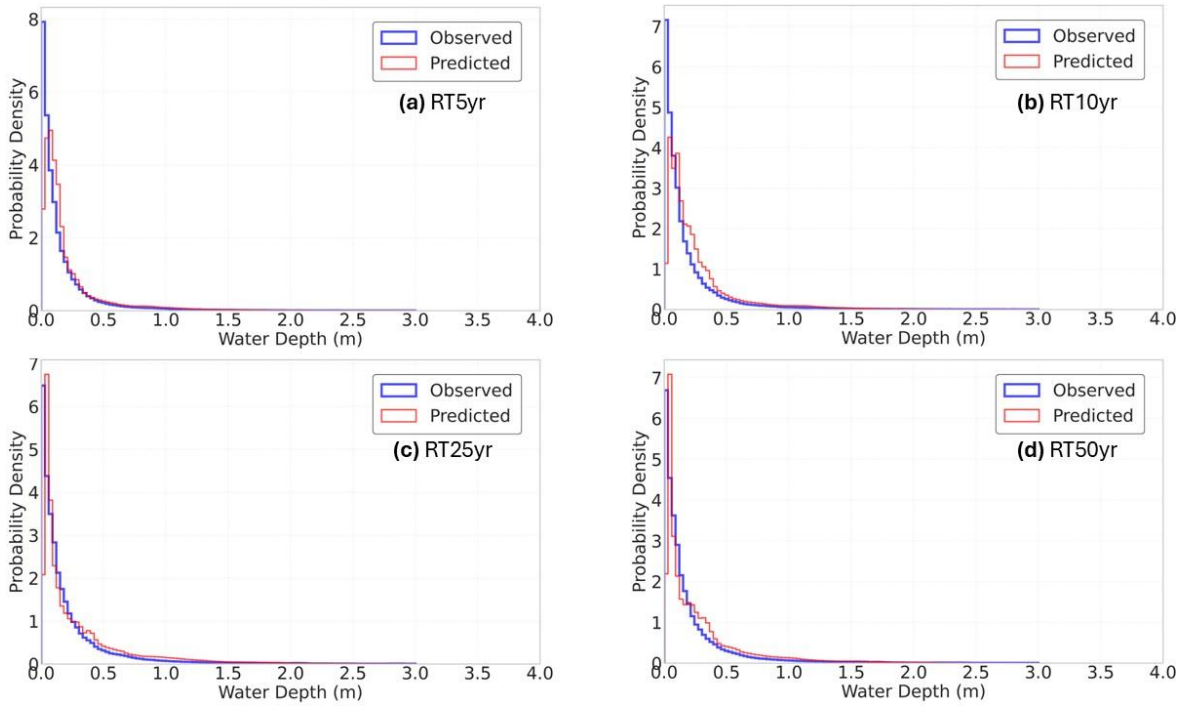


Figure 4 Mean reference and RF-predicted flood water depth across Milan for different precipitation scenarios.

Figure 4 shows a clear and systematic increase in flood magnitude and spatial extent with increasing return period. The gradual increase in mean flood depth further demonstrates that flood severity intensifies not only at localised hotspots but across the broader urban domain, providing a physical explanation for the performance trends observed in **Figure 3** and **Table 2**.

A consistent difference between predicted and reference depth distributions is also evident. Predicted depths show a mild compression of the depth distribution values, where extreme values are attenuated and near-zero depths are slightly elevated. This is reflected in the vertical whiskers, which represent the standard deviation at each scenario. The whiskers for the predicted depths indicate a reduced spread compared to the reference depths, consistent with the variance-reduction and averaging properties of Random Forest regression. These properties tend to smooth sharp gradients and shrink extreme predictions toward central tendencies (Breiman, 2001; Hastie et al., 2009). As a result, mean predicted depths are moderately higher than reference means, particularly for higher return periods. This pattern reflects statistical regularisation effects inherent to ensemble regression rather than a systematic physical overestimation of flood magnitude.

To further characterise model performance beyond tolerance-based accuracy metrics and flood depth magnitude, the distributional behaviour of prediction errors was analysed across all precipitation scenarios (see **Figure 5 and 6**). Figures 5a to 5e show the error probability density functions (PDFs), while Figure 6 presents the corresponding cumulative distribution functions (CDFs), allowing a consistent comparison of central dispersion and tail behaviour as rainfall severity increases. Across all return periods, error distributions remain closely centred around zero, indicating limited systematic bias. With increasing return period, the distributions broaden and the tails become more pronounced, reflecting increased hydraulic complexity and stronger depth gradients under more extreme events. The CDF comparison confirms that most errors remain confined to a narrow range around zero for all scenarios, while progressively higher-return-period events exhibit a gradual increase in the probability of larger deviations.



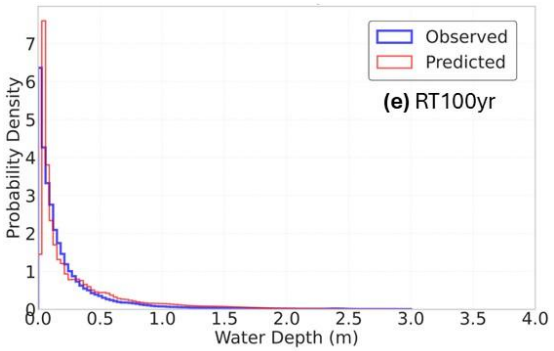


Figure 5. Probability density functions (PDFs) of RF prediction errors for flood water depth across precipitation return period scenarios in Milan.

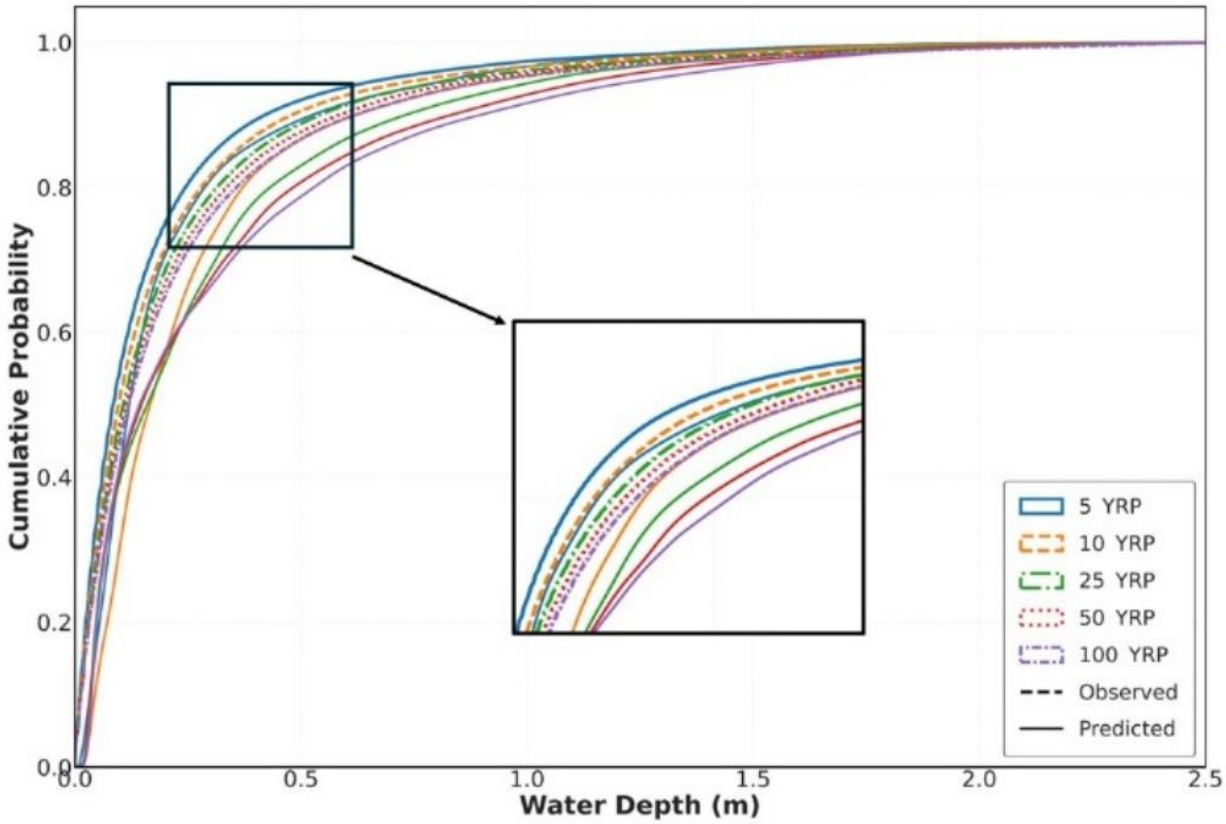


Figure 6. Cumulative distribution functions (CDFs) of RF prediction errors for flood water depth across precipitation return period scenarios in Milan.

Building on these performance diagnostics, prediction confidence classes were derived from the uncertainty estimates and grouped into four categories: Low (>1 m), Medium (0.5-1 m), High (0.25-0.5 m) and Very High (0-0.25 m) confidence

Figure 7 presents the spatial distribution of these confidence classes for all precipitation scenarios.

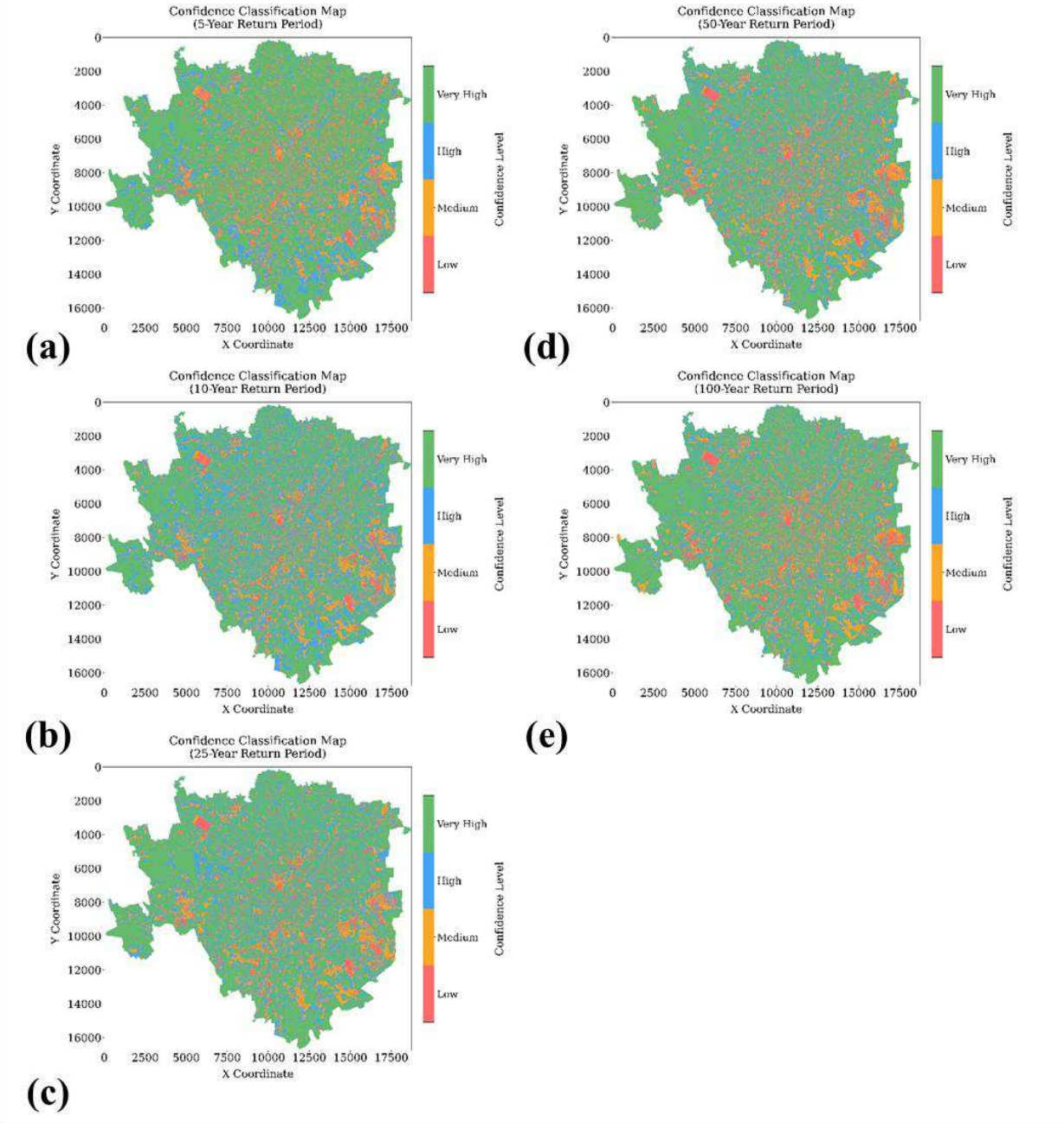


Figure 7. Spatial distribution of RF prediction confidence classes for flood water depth in Milan across multiple return period scenarios: (a) 5-year, (b) 10-year, (c) 25-year, (d) 50-year, and (e) 100-year precipitation events.

High and very high confidence predictions dominate large portions of the study area, particularly in zones where flood behaviour is primarily controlled by local terrain morphology and topographic depressions. In contrast, lower confidence levels are mainly confined to transition areas characterised by more complex interactions between terrain and land use. As event severity increases, the overall spatial organisation of confidence remains broadly stable, while localised expansions of lower-confidence areas become more evident, consistent with increasing flood depth

variability under extreme rainfall conditions. To examine whether these confidence patterns are supported by the spatial distribution of model uncertainty, prediction uncertainty was quantified using the inter-percentile range between the 10th and 90th percentiles of predicted flood water depth.

This inter-percentile range captures the central 80% of the predictive distribution while remaining less sensitive to extreme values and outliers than minimum–maximum spread or variance-based measures. Such percentile-based uncertainty metrics are commonly adopted in hydrological modelling to provide robust characterisation of predictive dispersion under skewed distributions (Wilks, 2011; Montanari, 2007). The resulting spatial patterns are shown in **Figure 8**.

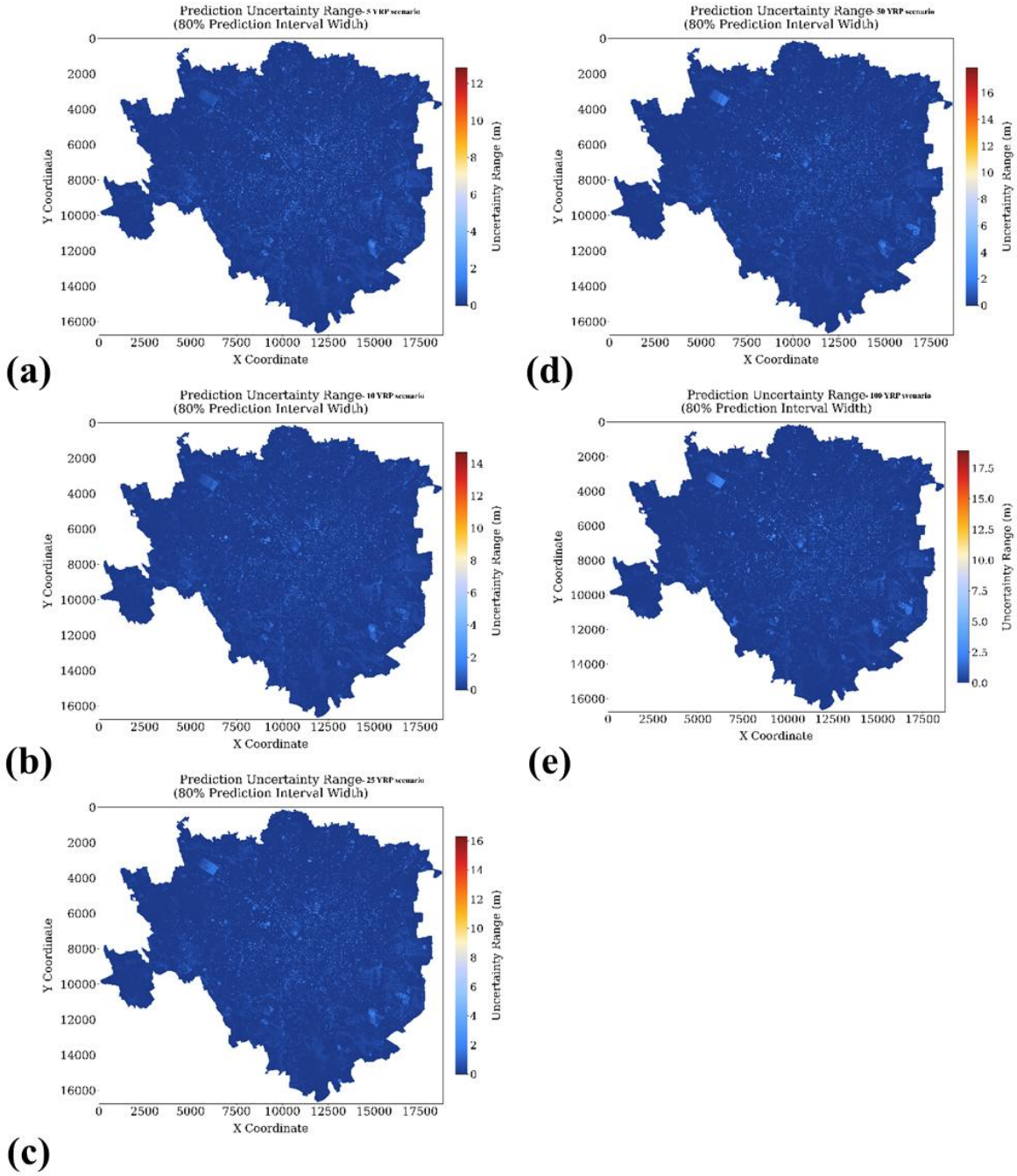


Figure 8. Spatial distribution of prediction uncertainty (10th–90th percentile range) for RF-predicted flood water depth in Milan across multiple return period scenarios: (a) 5-year, (b) 10-year, (c) 25-year, (d) 50-year, and (e) 100-year precipitation events.

Higher uncertainty values are concentrated in areas affected by deeper and more heterogeneous inundation, including topographic depressions, flow convergence zones, and transitional urban areas. Conversely, shallow-flooded or well-drained zones exhibit consistently low uncertainty, indicating stable and well-constrained model behaviour in regions

dominated by clear topographic control. Although uncertainty magnitudes increase locally under higher return period events, the overall spatial structure of uncertainty remains comparable across precipitation scenarios.

To evaluate whether these uncertainty patterns translate into systematic differences in predictive accuracy, the relationship between prediction confidence and model errors is further examined by analysing error distributions stratified by confidence class (**Figure 9**).

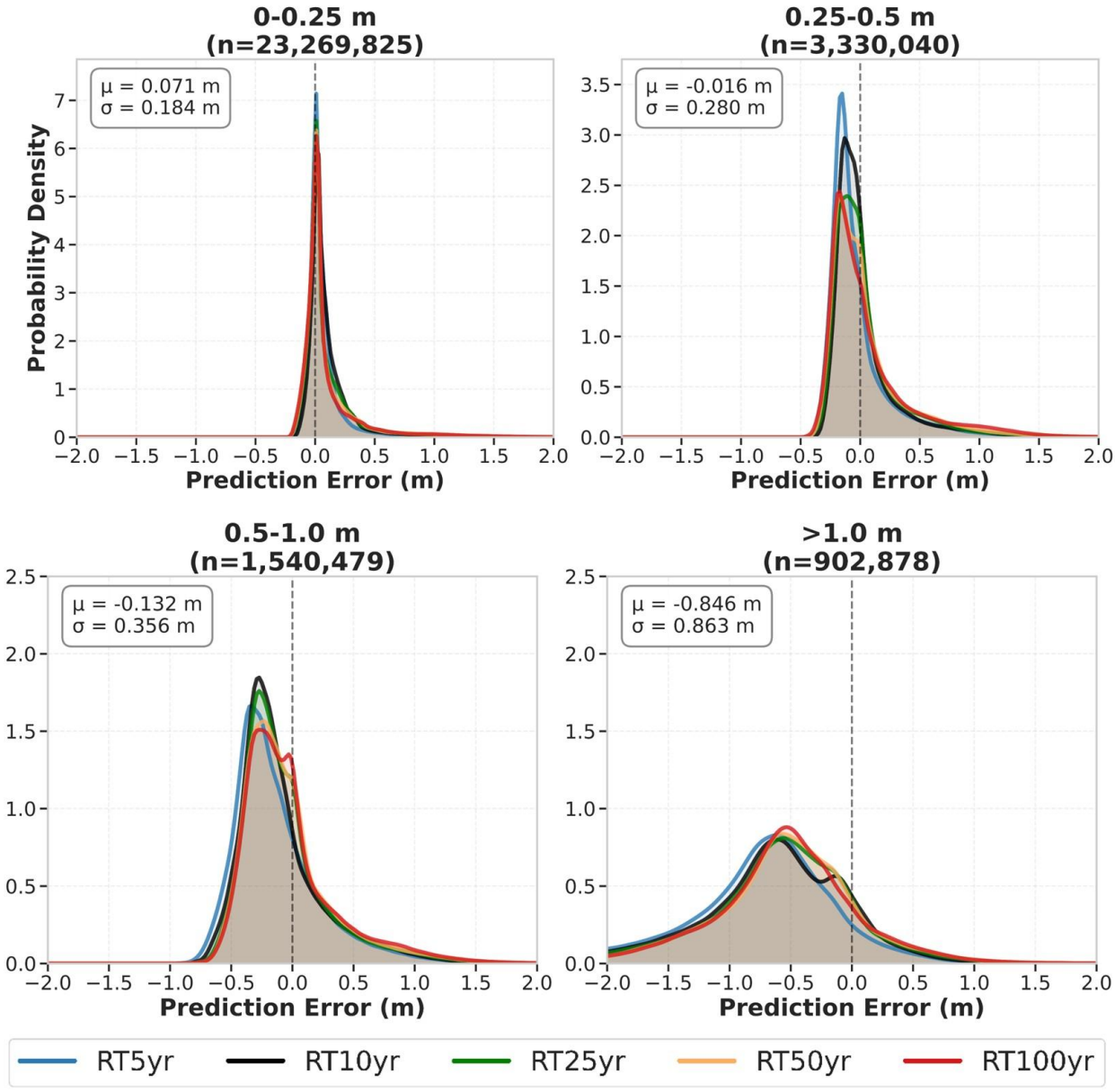


Figure 9. Probability density distributions of prediction errors stratified by confidence class for RF-predicted flood water depth in Milan (Italy) under different precipitation return period scenarios: 5-year, 10-year, 25-year, 50-year, and 100-year events. The distributions are shown for Very High, High, Medium, and Low confidence classes. Dashed vertical lines indicate zero error.

Across all return periods, high- and very-high-confidence predictions are associated with lower error magnitudes and narrower dispersion ranges. In contrast, medium- and low-confidence classes exhibit broader error distributions and greater variability. This consistent relationship confirms that the derived confidence classes provide a meaningful indicator of local prediction reliability. They are closely aligned with both uncertainty patterns and error behaviour. To complement the spatially explicit confidence patterns shown in **Figure 7**, **Table 3** provides a domain-wide quantitative summary of prediction confidence class distributions across all precipitation scenarios.

Table 3. Distribution of prediction confidence classes across precipitation scenarios for RF-predicted flood water depth in Milan.

Confidence level	5-YRP	10-YRP	25-YRP	50-YRP	100-YRP
Very High	27.5%	21.7%	29.3%	29.1%	30.1%
High	33.6%	33.8%	22.1%	22.1%	20.3%
Medium	20.4%	25.4%	26.4%	24.2%	22.5%
Low	18.5%	19.1%	22.2%	24.6%	27.0%

High and very high confidence predictions collectively dominate the study area for all return periods, although their relative contributions vary with increasing event severity. In particular, the proportion of low-confidence areas increases progressively from lower to higher return periods, reflecting the amplification of spatial heterogeneity and localised inundation under more extreme rainfall conditions. These aggregated trends are consistent with the spatial confidence distributions in **Figure 7**, the uncertainty patterns in **Figure 8**, and the confidence–error relationships in **Figure 9**. The results indicate that the model captures the primary topography-driven structure of urban flooding while maintaining robust predictive capability across a wide range of rainfall severities. This analysis establishes a clear reference baseline for subsequent investigations of model transferability to other urban areas and sensitivity to variations in terrain data resolution.

3.3. Model interpretability and feature contribution results

To quantify and compare the relative contribution of each predictor to the RF flood water depth model, **Table 4** summarises the global feature importance and ranking derived from the SHAP and permutation importance analyses.

Table 4. Global feature importance and ranking for the Random Forest flood water depth model based on SHAP values and permutation importance.

Feature	Depression	TRI	Precipitation	Land use	TPI	Slope (Radians)	TWI	Sum
Global SHAP importance (-) [Eq. 1]	0.24	0.07	0.05	0.02	0.01	0.01	0.01	0.41
Global SHAP importance (%) [Eq. 2]	59.07	11.71	16.28	4.92	2.86	3.07	2.09	100.00
Permutation importance [Eq. 3]	1.90	0.27	0.04	0.23	0.08	0.00	0.03	2.55

Normalised permutation importance [Eq. 4]	0.75	0.11	0.02	0.09	0.03	0.00	0.01	1.00
SHAP rank	1	3	2	4	6	5	7	-
Permutation rank	1	2	5	3	4	7	6	-
Average rank	1	2.5	3.5	3.5	5	6	6.5	-

The SHAP results indicate that mapped topographic depressions exert the dominant control on predicted flood water depth, accounting for approximately 59.1% of the total model contribution. This highlights the primary role of micro-topographic storage in governing spatial patterns of pluvial flooding. Precipitation intensity represents the second most influential factor (16.3%), reflecting its role in scaling flood magnitude across different return-period scenarios, while the TRI follows (11.7%), indicating that local terrain heterogeneity modulates lateral flow redistribution and localised ponding. Land use exhibits a moderate contribution (4.9%), consistent with differences in imperviousness and surface characteristics. In contrast, slope, TPI, and TWI contribute comparatively less (each below 3.1%).

Permutation importance analysis confirms that topographic depressions are the most performance-critical predictor, accounting for approximately 75% of the total permutation importance. This indicates that model accuracy strongly depends on preserving depression information. TRI (10%) and land use (9%) follow, suggesting that terrain heterogeneity and surface characteristics provide additional, non-redundant predictive information. TPI contributes marginally (3%), while precipitation exhibits a relatively small permutation effect (2%). Slope and TWI show negligible influence on predictive performance once dominant terrain controls are represented.

The apparent difference between SHAP and permutation rankings for precipitation is physically interpretable and does not indicate model inconsistency. SHAP measures how strongly a feature contributes to predicted flood depths, whereas permutation importance quantifies how much predictive skill deteriorates when that feature is disrupted. This behaviour is expected in the presence of correlated and interacting predictors. In this study, precipitation is encoded through discrete return-period scenarios and interacts with dominant terrain predictors such as depressions and TRI. As a result, other variables can partially compensate when precipitation is permuted. This leads to a limited reduction in model performance, even though precipitation substantially shifts predicted flood magnitudes. This behaviour reflects predictor redundancy and interaction effects rather than instability in the model. It is consistent with observations reported in recent explainable flood susceptibility studies (Molnar et al., 2020; Gulshad et al., 2024; Pradhan et al., 2023). To provide a unified interpretation, SHAP and permutation results were combined through an average-ranking approach**.**

This synthesis clearly identifies topographic depressions as a Very High importance predictor, followed by TRI and precipitation as High importance controls, land use as Medium importance, and TPI, slope, and TWI as Low importance predictors. The resulting hierarchy is physically consistent with the established understanding of urban pluvial flooding processes. Numerous studies have demonstrated that micro-topographic storage and depression representation exert dominant control over water accumulation and ponding patterns in urban environments (Cook & Merwade, 2009; Leitão et al., 2009; Schubert & Sanders, 2012). Terrain heterogeneity influences lateral redistribution of runoff (Teng et al., 2017), while precipitation intensity primarily scales flood magnitude across events (Arnbjerg-Nielsen et al., 2013; Cristiano et al., 2017). Land surface characteristics, including imperviousness and urban land use,

typically exert a secondary but non-negligible influence on flood depth distribution (Ochoa-Rodriguez et al., 2015; Zhang et al., 2018). Together, these findings support the robustness, interpretability, and physical plausibility of the selected RF model configuration and predictor set.

3.4. Impact of DEM resolution on flood water depth prediction

To assess the impact of DEM resolution on model performance, the 1 m LiDAR DEM was replaced with a 5 m EU-DEM for the Milan case study. Spatial results are presented for the 25-year return period as a representative intermediate scenario, while the behaviour across all precipitation scenarios (5–100-year return periods) is summarised using aggregated statistical and confidence-based metrics.

3.4.1. **Overall predictive performance under coarser DEM resolution**

Using the 5 m DEM, the RF model maintains stable and physically consistent performance across all precipitation scenarios. Mean Absolute Error (MAE) increases gradually from 0.22 m for the 5-year event to 0.30 m for the 100-year event, while RMSE rises from 0.35 m to 0.46 m. Median errors remain low (≤0.06 m) for all scenarios, indicating that the majority of grid-cell predictions remain close to reference flood depths despite the loss of fine-scale elevation detail. The systematic increase in error magnitude with return period reflects the physically expected response to increasing rainfall severity, which produces deeper and more spatially heterogeneous flooding. Importantly, no abrupt deterioration in model performance is observed. This indicates that flood depth prediction is not critically dependent on sub-metre micro-topography but is instead governed by larger-scale terrain organisation. A concise overview of key performance metrics across all scenarios is provided in **Table 5**.

Table 5: Summary performance metrics of Random Forest flood water depth predictions using the 5 m EU-DEM across precipitation return period scenarios.

Scenario	R^2	MAE (m)	RMSE (m)	Within ±0.10 m (%)	Within ±0.25 m (%)	Within ±0.50 m (%)
5-YRP	0.13	0.22	0.35	37.30	68.20	90.10
10-YRP	0.16	0.24	0.36	34.30	65.20	88.10
25-YRP	0.18	0.27	0.40	32.30	61.30	84.40
50-YRP	0.18	0.29	0.43	31.50	59.70	82.30
100-YRP	0.18	0.30	0.46	30.90	58.20	81.00

The coefficient of determination (R^2) values under the 5 m DEM configuration range between 0.13 and 0.18. While lower than those obtained under 1 m LiDAR resolution, this behaviour is expected given the smoothing of fine-scale elevation variability and the highly skewed distribution of flood depths across the urban domain. In flood-depth grids dominated by shallow inundation and sparse extreme values, variance-based metrics such as R^2 are particularly sensitive to attenuation of extremes and ensemble-averaging effects inherent to Random Forest regression. Consequently, R^2 may underestimate operational predictive skill in this context. Absolute error metrics and tolerance-

based accuracy thresholds, therefore, provide a more physically meaningful evaluation of model performance for high-resolution urban flood applications.

3.4.2. **Spatial flood depth prediction and error characteristics**

Figure 10 illustrates the spatial behaviour of the RF model for the 25-year precipitation scenario using the 5 m DEM. The panels show (**Fig. 10a**) predicted flood water depth and (**Fig. 10b**) prediction error, defined as the difference between predicted and baseline flood depths used for model evaluation. Higher error magnitudes are predominantly associated with deep topographic depressions and flow convergence zones, whereas shallow-flooded and well-drained areas exhibit smaller errors.

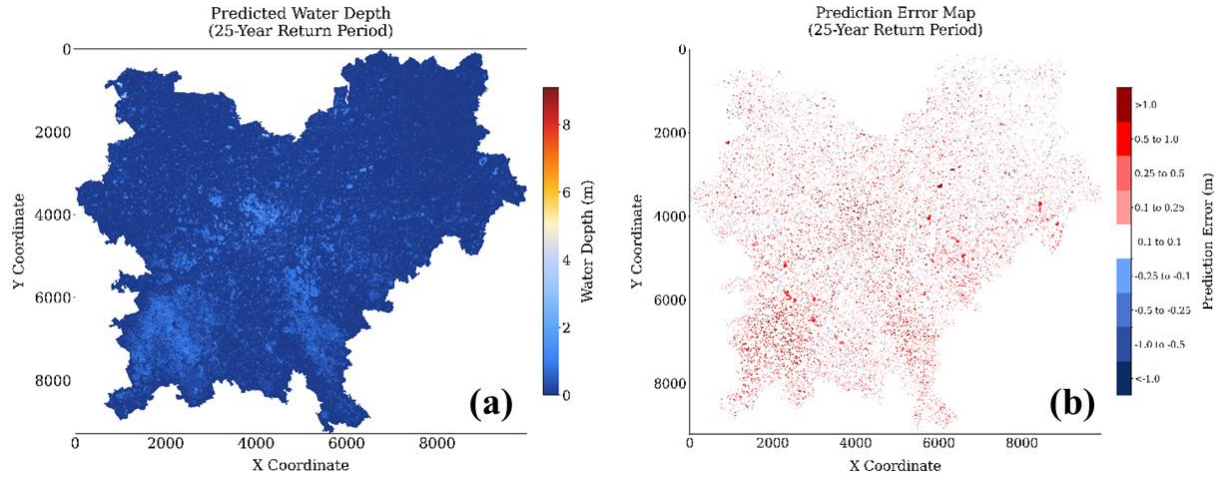


Figure 10. Spatial patterns of Random Forest flood water depth prediction using the 5 m EU-DEM for the 25-year precipitation scenario in Milan (Italy): (a) predicted flood water depth, and (b) prediction error.

3.4.3. **Tolerance-based accuracy, confidence behaviour, and implications for transferability**

Tolerance-based accuracy metrics indicate that predictive skill remains high at impact-relevant thresholds across all precipitation scenarios. Even for the 100-year event, more than 58% and 81% of grid cells fall within ±0.25 m and ±0.50 m of baseline flood depths, respectively (see **Table 5**). These results confirm that, despite increasing hydraulic complexity under extreme rainfall, the model maintains practically relevant accuracy levels when applied using a reduced DEM resolution. Beyond tolerance-based accuracy, confidence diagnostics further demonstrate the robustness of the Random Forest framework across precipitation scenarios. As shown in **Figure 11**, high- and very-high-confidence predictions constitute a substantial proportion of the urban domain for all return periods, while the proportion of low-confidence areas increases gradually with event severity. This trend reflects the growing spatial extent of hydraulically complex flood patterns under extreme rainfall rather than resolution-induced instability.

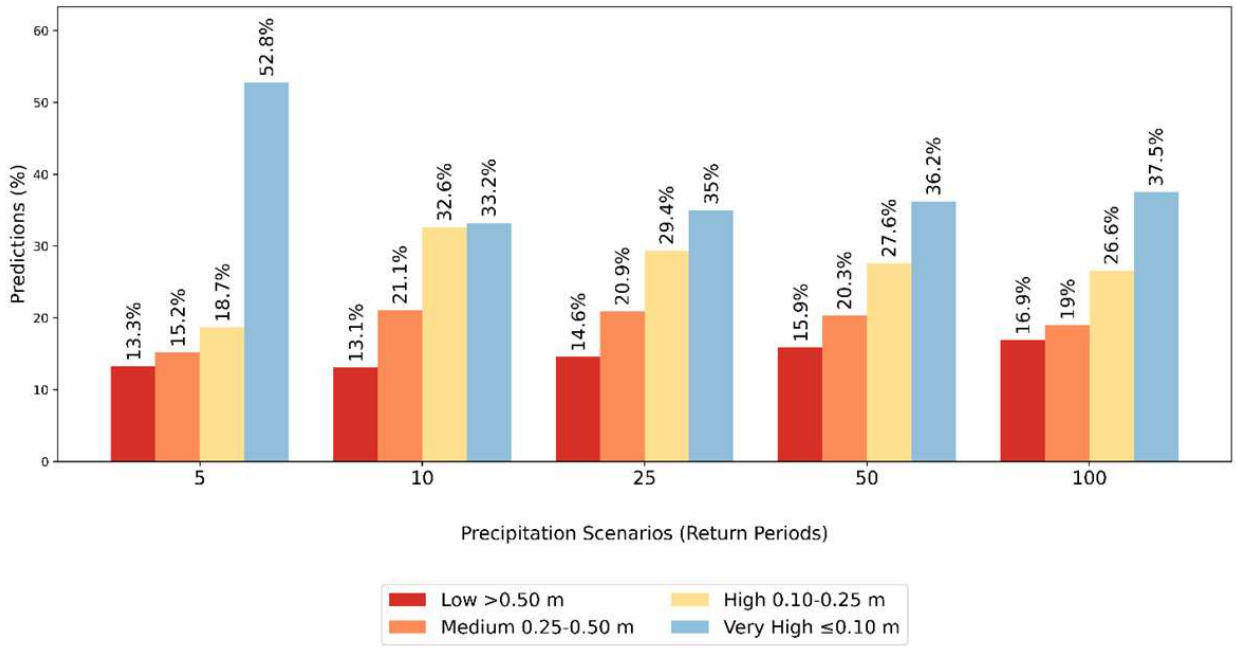


Figure 11. Distribution of prediction confidence classes (low, medium, high, and very high) for RF flood water depth predictions using the 5 m EU-DEM across precipitation return period scenarios (5–100 years). Increasing rainfall severity is associated with a gradual expansion of low-confidence areas, while high- and very-high-confidence predictions remain dominant across all scenarios.

Compared with the 1 m LiDAR-based results presented in section 3.1, the use of a 5 m DEM leads to a moderate and physically consistent increase in prediction error. Across precipitation scenarios, tolerance-based accuracy at the impact-relevant ±0.25 m threshold decreases by approximately 16–18 percentage points relative to the 1 m LiDAR configuration (e.g., from ~76–86% to ~58–68%; Tables 4 and 6). In contrast, accuracy within ±0.50 m remains comparatively stable, decreasing by approximately 4–7 percentage points across return periods (from ~88–94% to ~81–90%; Tables 4 and 6), indicating that most predictions remain within operationally relevant tolerance levels despite DEM coarsening. The largest relative reductions occur under the strict ±0.10 m threshold (approximately 23–27 percentage points), reflecting the increased sensitivity of fine-scale depth variations to DEM resolution.

This occurs without altering the dominant spatial organisation of flood depth patterns or confidence behaviour. This stability highlights the primary role of relative topographic structure, local depressions, and terrain heterogeneity in controlling urban pluvial flooding. From an operational perspective, these findings indicate that reliable, spatially explicit flood water depth predictions can be achieved using medium-resolution DEMs commonly available at national and continental scales. The proposed DEM-centric machine learning framework is therefore well suited for regional screening, rapid urban flood assessment, and climate adaptation planning in data-scarce environments.

3.5. **Model transferability: application to Cologne using 1 m LiDAR predictors**

As described in Section 2.6, the Cologne application represents strict out-of-domain transfer. The Random Forest model trained in Milan was applied directly to Cologne using the identical predictor set derived at 1 m LiDAR resolution, without retraining, recalibration, or hyperparameter modification. All ensemble settings remained unchanged. The results therefore reflect the model's inherent generalization capability rather than location-specific adaptation.

3.5.1. **Scenario-wise predictive performance and error characteristics**

Transfer performance was evaluated under five precipitation return period scenarios ranging from 5 to 100 years, using the same statistical and tolerance-based performance metrics adopted for the Milan case study. **Table 6** summarises the predictive performance of the RF model under transfer conditions.

Table 6: Scenario-wise performance metrics of Random Forest flood water depth predictions for Cologne using 1 m LiDAR-derived predictors across precipitation return period scenarios.

Scenario	R^2	MAE (m)	RMSE (m)	Within ±0.10 m (%)	Within ±0.25 m (%)	Within ±0.50 m (%)
5-YRP	0.09	0.28	0.42	35.10	58.60	81.00
10-YRP	0.10	0.28	0.42	36.50	61.10	80.40
25-YRP	0.10	0.25	0.41	42.70	67.70	83.30
50-YRP	0.11	0.31	0.47	35.80	59.30	78.50
100-YRP	0.11	0.31	0.47	35.90	59.40	78.60

Across all precipitation scenarios, mean absolute error (MAE) values range between 0.25 m and 0.31 m, while root mean square error (RMSE) values vary between 0.41 m and 0.47 m, indicating stable model behaviour across increasing rainfall severity. The lowest error levels are observed for the 25-year return period, which also exhibits the highest tolerance-based agreement. Tolerance-based accuracy metrics further demonstrate robust performance. Between 58.6% and 67.7% of grid cells fall within ±0.25 m of the reference flood depths, while 78.5% to 83.3% fall within ±0.50 m across all scenarios.

Compared with the Milan 1 m LiDAR configuration (**Table 1**), mean absolute error increases under spatial transfer by approximately 0.03–0.06 m across precipitation scenarios, while RMSE increases by approximately 0.01–0.04 m. This moderate degradation reflects the expected performance reduction associated with a strict out-of-domain application. Under spatial transfer, the coefficient of determination (R^2) ranges between 0.09 and 0.11. Although modest, this behaviour is consistent with an out-of-domain application without retraining or recalibration. In addition, the strongly skewed flood-depth distribution and dominance of shallow inundation amplify the sensitivity of variance-based metrics to extreme-value attenuation and distributional imbalance. As a result, R^2 remains comparatively low despite stable MAE, RMSE, and tolerance-based performance. For operational flood hazard assessment, tolerance thresholds and absolute error metrics provide a more relevant measure of predictive reliability under transfer conditions.

Although agreement decreases modestly for the most extreme precipitation events, the majority of predictions remain within impact-relevant tolerance thresholds, confirming that the RF model preserves practical predictive skill when transferred to an independent urban setting. Overall, these results indicate that the DEM-centric RF framework maintains consistent and operationally relevant performance under spatial transfer, despite differences in urban morphology and drainage configuration between Milan and Cologne. The stability of both absolute error metrics and tolerance-based accuracy across scenarios provides a strong quantitative basis for further spatial and confidence-based analyses presented in the following sections.

3.5.2. Spatial behaviour under a representative scenario

To illustrate spatial transferability without expanding the manuscript with five sets of maps, the 25-year return period is used as a representative intermediate scenario because it combines (i) the highest tolerance agreement (±0.25 m and ±0.50 m) among Cologne scenarios and (ii) substantial depth variability relative to frequent events (5–10 years), without being dominated by the broadest/deepest inundation associated with the 50–100-year scenarios (**Table 5**). **Figure 12** illustrates the spatial behaviour of the Random Forest model for the 25-year precipitation scenario in Cologne using 1 m LiDAR-derived predictors.

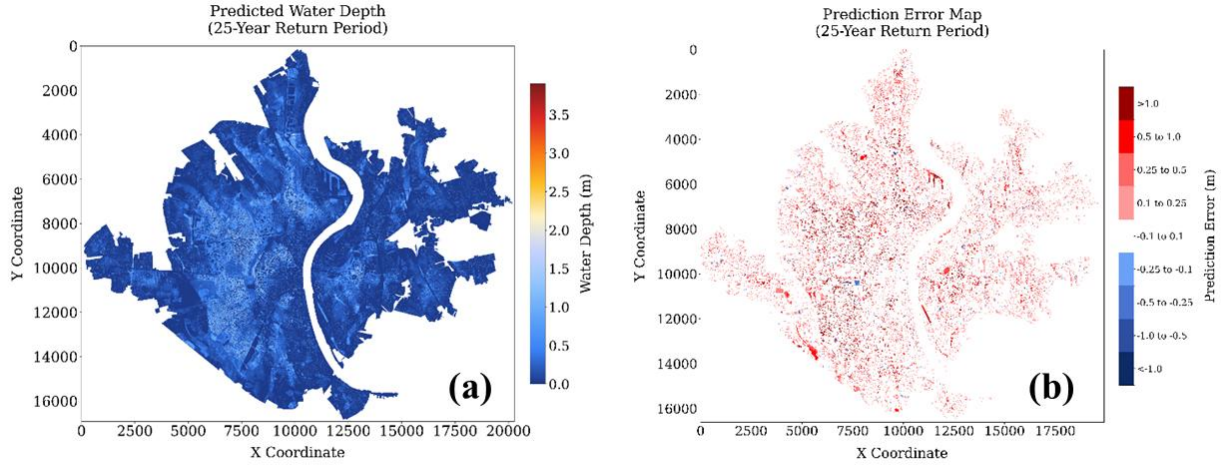


Figure 12. Spatial distribution of RF–predicted flood water depth for Cologne (Germany) using 1 m LiDAR-derived predictors under the 25-year precipitation return period scenario: (a) predicted flood water depth and (b) prediction error.

The predicted flood water depth map (**Fig. 12a**) reproduces the dominant spatial organisation of deeper flood zones and surrounding shallower areas across the urban domain. Areas of elevated predicted water depth are spatially coherent and correspond to locations expected to experience enhanced surface water accumulation. The corresponding prediction error map (**Fig. 12b**) reveals structured and localised discrepancies, with larger errors concentrated in limited portions of the domain. This spatial organisation indicates that model performance in Cologne is governed by a combination of transferable topographic controls and location-specific flood complexity, while the majority of the urban area exhibits stable and physically consistent predictions.

3.5.3. **Confidence behaviour across scenarios**

Confidence diagnostics for Cologne indicate that high-confidence predictions occupy a substantial proportion of the mapped urban domain across all precipitation return periods. Their relative distribution, however, varies with event severity. The Very High confidence class accounts for 46.0–46.9% of valid grid cells for the 5–25-year scenarios. It decreases to 38.6% for the 50–100-year scenarios, while the High confidence class increases to 23.6% for the most extreme events. Low confidence areas remain non-negligible across all scenarios (22.7–29.9% of valid points), indicating that uncertainty-aware interpretation remains necessary when applying the model outside the Milan training domain (**Figure 13**).

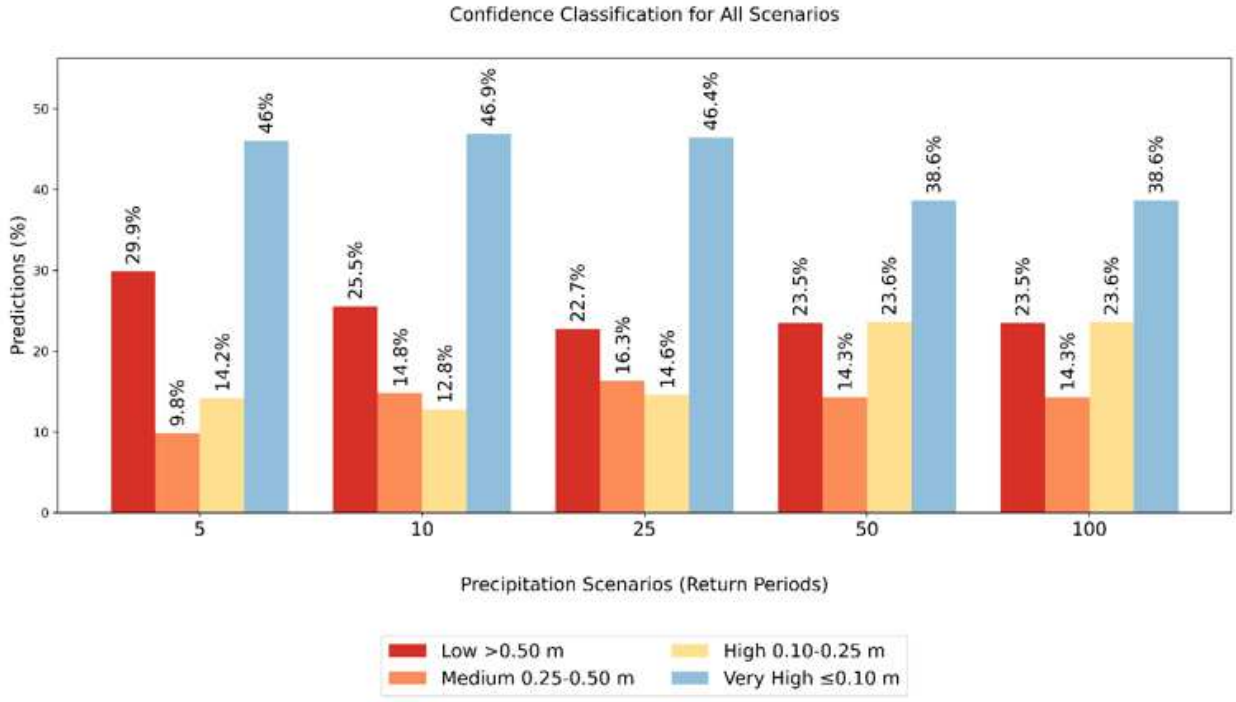


Figure 13: Distribution of prediction confidence classes (low, medium, high, and very high) for Random Forest flood water depth predictions in Cologne (Germany) using 1 m LiDAR-derived predictors across precipitation return period scenarios (5–100 years).

Confidence-stratified error metrics show that confidence classes are strongly discriminative under transfer conditions. For all precipitation scenarios, predictions classified as Very High confidence exhibit near-complete tolerance compliance, with approximately 99.9–100% of grid cells within ±0.25 m and ±0.50 m of baseline flood depths. Predictions in the Low confidence class exhibit substantially weaker tolerance performance (±0.25 m compliance of 18.2–26.7%) and higher MAE and RMSE values. This consistent separation across confidence classes demonstrates that the confidence framework remains operationally meaningful in an independent urban setting, enabling users to prioritise locations where predictions are most reliable.

4. Conclusion

This study demonstrates that a DEM-centric Random Forest framework can provide reliable and transferable predictions of urban pluvial flood water depth across multiple precipitation return periods and spatial contexts. Results from Milan and Cologne show that predictive performance remains stable under reduced DEM resolution and independent urban transfer, with mean absolute errors typically below 0.30 m and a high proportion of grid cells consistently falling within impact-relevant tolerance thresholds (±0.25 m and ±0.50 m). Spatial analyses confirm that prediction errors are not randomly distributed. They are instead concentrated in hydraulically complex zones characterised by deep depressions and flow convergence, while the majority of urban areas exhibit shallow flooding and stable, well-constrained predictions. These patterns persist across DEM resolutions and case studies, highlighting the dominant role of relative terrain structure and depression-scale storage in controlling urban pluvial flooding. Confidence diagnostics further reinforce the robustness of the modelling framework. High- and very-high-confidence predictions account for a substantial fraction of the urban domain across all precipitation scenarios. They consistently achieve near-complete tolerance compliance. The systematic separation between confidence classes demonstrates that the confidence framework remains operationally meaningful under spatial transfer. It enables users to identify areas where predictions are most reliable and where caution is warranted. Limitations include the reliance on scenario-based reference flood depths and the absence of explicit drainage network representation, which may influence performance in highly engineered urban systems. The focus on surface-topographic controls reflects the intended application to large-scale screening contexts, where detailed drainage network data are often unavailable or inconsistent across cities. Overall, the findings indicate that reliable, spatially explicit flood water depth predictions can be achieved using medium-resolution DEMs commonly available at national and continental scales. This significantly broadens the applicability of data-driven flood modelling beyond LiDAR-rich regions. The proposed framework is therefore well suited for large-scale urban flood screening, rapid hazard assessment, and climate adaptation planning, particularly in data-scarce environments where traditional high-resolution hydraulic modelling is infeasible. Future work will extend the proposed framework by replacing constant design-storm precipitation with spatially distributed, grid-cell–level precipitation fields derived from Copernicus datasets, enabling explicit representation of rainfall heterogeneity and further improving the realism of urban pluvial flood depth predictions.

References

Apel H, Thieken AH, Merz B, Blöschl G (2004). Flood risk assessment and associated uncertainty. *Natural Hazards and Earth System Sciences* 4:295–308. https://doi.org/10.5194/nhess-4-295-2004

Apel H, Merz B, Thieken AH (2009). Quantification of uncertainties in flood risk assessments. *International Journal of River Basin Management* 7:1–14. https://doi.org/10.1080/15715124.2008.9635344

Arnbjerg-Nielsen, K., Willems, P., Olsson, J., et al. (2013). Impacts of climate change on rainfall extremes and urban drainage systems: A review. *Water Science and Technology*, 68(1), 16–28. https://doi.org/10.2166/wst.2013.251

Bates PD, Pappenberger F, Romanowicz RJ (2014). Uncertainty in flood inundation modelling. In: *Applied Uncertainty Analysis for Flood Risk Management*. World Scientific. https://doi.org/10.1142/9781848162716_0010

Bates, P. D., Horritt, M. S., & Fewtrell, T. J. (2010). A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling. *Journal of Hydrology*, 387(1–2), 33–45. https://doi.org/10.1016/j.jhydrol.2010.03.027

Belgiu M, Drăguț L (2016). Random forest in remote sensing: a review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing* 114:24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bentivoglio R, Isufi E, Jonkman SN, Taormina R (2022). Deep learning methods for flood mapping: a review. *Hydrology and Earth System Sciences* 26:4345–4378. https://doi.org/10.5194/hess-26-4345-2022

Beven KJ, Kirkby MJ (1979). A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin* 24(1):43–69. https://doi.org/10.1080/02626667909491834

Breiman L (2001). Random forests. *Machine Learning* 45:5–32. https://doi.org/10.1023/A:1010933404324

Bowes BD, Wang C, Ercan MB, Culver TB, Beling PA, Goodall JL (2022). Reinforcement learning-based real-time control of coastal urban stormwater systems to mitigate flooding and improve water quality. *Environmental Science: Water Research & Technology* 8:1980–1995. https://doi.org/10.1039/D1EW00582K

Büchele B, Kreibich H, Kron A, Thieken A, Ihringer J, Oberle P, Merz B, Nestmann F (2006). Flood-risk mapping: contributions towards an enhanced assessment of extreme events and associated risks. *Natural Hazards and Earth System Sciences* 6:485–503. https://doi.org/10.5194/nhess-6-485-2006

Chitwatkulsiri D, Miyamoto H (2023). Real-time urban flood forecasting systems for Southeast Asia: a review of present modelling and its future prospects. *Water* 15(1):178. https://doi.org/10.3390/w15010178

Choubin, B., Jaafari, A., Henareh, J., Karimi, O., & Sajedi Hosseini, F. (2025). Explainable artificial intelligence (XAI) for interpreting predictive models and key variables in flood susceptibility. *Results in Engineering*, 27, 105976. https://doi.org/10.1016/j.rineng.2025.105976

Cook, A., & Merwade, V. (2009). Effect of topographic data, geometric configuration and modeling approach on flood inundation mapping. *Journal of Hydrology*, 377(1–2), 131–142. https://doi.org/10.1016/j.jhydrol.2009.08.015

Copernicus Land Monitoring Service (2018). CORINE Land Cover (CLC) 2018. European Environment Agency. https://land.copernicus.eu

Costabile P, Costanzo C, De Lorenzo G, Macchione F (2019). Is local flood hazard assessment in urban areas significantly influenced by the physical complexity of the hydrodynamic inundation model? *Journal of Hydrology* 578:124231. https://doi.org/10.1016/j.jhydrol.2019.124231

Cristiano, E., ten Veldhuis, M.-C., & van de Giesen, N. (2017). Spatial and temporal variability of rainfall and their effects on hydrological response in urban areas – A review. *Hydrology and Earth System Sciences*, 21, 3859–3878. https://doi.org/10.5194/hess-21-3859-2017

Criminisi A, Shotton J, Konukoglu E (2012). Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision* 7(2–3):81–227. https://doi.org/10.1561/0600000035

DHI (2017). *MIKE URBAN: Integrated urban water management software*. DHI. https://www.dhigroup.com/software/mike-urban

Do Lago CAF, Giacomoni MH, Bentivoglio R, Taormina R, Gomes MN, Mendiondo EM (2023). Generalizing rapid flood predictions to unseen urban catchments with conditional generative adversarial networks. *Journal of Hydrology* 621:129276. https://doi.org/10.1016/j.jhydrol.2023.129276

781 Donat MG, Lowry AL, Alexander LV, O'Gorman PA, Maher N (2016). More extreme precipitation in the world's
782 dry and wet regions. *Nature Climate Change* 6:508–513. https://doi.org/10.1038/nclimate2941

783 Dottori F, Szewczyk W, Ciscar JC, et al. (2018). Increased human and economic losses from river flooding with
784 anthropogenic warming. *Nature Climate Change*, **8**, 781–786. https://doi.org/10.1038/s41558-018-0257-z

785 El Baida M, Boushaba F, Chourak M, Hosni M (2024). Real-time urban flood depth mapping: convolutional neural
786 networks for pluvial and fluvial flood emulation. *Water Resources Management* 38(12):4763–4782.
787 https://doi.org/10.1007/s11269-024-03886-w

788 Essenfelder AH, Bagli S, Mysiak J, Pal JS, Mercogliano P, Reder A, Rianna G, Mazzoli P, Broccoli D, Luzzi V
789 (2022). Probabilistic assessment of pluvial flood risk across 20 European cities: a demonstrator of the Copernicus
790 Disaster Risk Reduction Service for pluvial flood risk in urban areas. *Water Economics and Policy* 8(3):2240007.
791 https://doi.org/10.1142/S2382624X22400070

792 Fewtrell, T. J., Duncan, A., Sampson, C. C., Neal, J. C., & Bates, P. D. (2011). Benchmarking urban flood models of
793 varying complexity and scale using high resolution terrestrial LiDAR data. *Physics and Chemistry of the Earth, Parts*
794 *A/B/C*, 36(7–8), 281–291. https://doi.org/10.1016/j.pce.2010.12.011

795 Fisher A, Rudin C, Dominici F (2019). All models are wrong, but many are useful: learning a variable's importance
796 by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research* 20:1–81.
797 https://doi.org/10.48550/arXiv.1801.01489

798 Genuer R, Poggi JM, Tuleau-Malot C (2010). Variable selection using random forests. *Pattern Recognition Letters*
799 31:2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

800 Gulshad, K., Yaseen, A., & Szydłowski, M. (2024). From data to decision: Interpretable machine learning for
801 predicting flood susceptibility in Gdańsk, Poland. *Remote Sensing*, 16(20), 3902.
802 https://doi.org/10.3390/rs16203902

803 Hastie T, Tibshirani R, Friedman J, Friedman JH (2009). *The elements of statistical learning.* Springer Series in
804 Statistics. Springer, New York. https://doi.org/10.1007/978-0-387-84858-7

805 Hunter NM, Horritt MS, Bates PD, Wilson MD, Werner MGF (2005). An adaptive time step solution for raster-based
806 storage cell modelling of floodplain inundation. *Advances in Water Resources* 28(9):975–991.
807 https://doi.org/10.1016/j.advwatres.2005.03.007

808 Islam T, Zeleke EB, Afroz M, Melesse AM (2025). A systematic review of urban flood susceptibility mapping: remote
809 sensing, machine learning, and other modeling approaches. *Remote Sensing* 17(3):524.
810 https://doi.org/10.3390/rs17030524

811 IPCC (2023). Climate Change 2023: Synthesis Report. *Contribution of Working Groups I, II and III to the Sixth*
812 *Assessment Report*. IPCC, Geneva, Switzerland. https://www.ipcc.ch/report/ar6/syr/

813 James G, Witten D, Hastie T, Tibshirani R (2021). *An introduction to statistical learning.* Springer US.
814 https://doi.org/10.1007/978-1-0716-1418-1

815 Kuhn M, Johnson K (2013). *Applied predictive modeling.* Springer-Verlag New York. https://doi.org/10.1007/978-1-
816 4614-6849-3

817 Leitão JP, Boonya-Aroonnet S, Prodanović D, Maksimović Č (2009). The influence of digital elevation model
818 resolution on overland flow networks for modelling urban pluvial flooding. *Water Science & Technology*
819 60(12):3137–3149. https://doi.org/10.2166/wst.2009.754

820 Liaw A, Wiener M (2002). Classification and regression by Random Forest. *R News* 2(3):18–22. https://journal.r-
821 project.org/articles/RN-2002-022/

822 Lin L, Wu Z, Liang Q (2019). Urban flood susceptibility analysis using a GIS-based multi-criteria analysis framework.
823 *Natural Hazards* 97:455–475. https://doi.org/10.1007/s11069-019-03615-2

Lindsay JB, Creed IF (2005). Removal of artifact depressions from digital elevation models: towards a minimum impact approach. *Hydrological Processes* 19:3113–3126. https://doi.org/10.1002/hyp.5835

Liu B, Li Y, Ma M et al (2025). A comprehensive review of machine learning approaches for flood depth estimation. *International Journal of Disaster Risk Science* 16:433–445. https://doi.org/10.1007/s13753-025-00639-0

Lundberg SM, Lee SI (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30. https://doi.org/10.48550/arXiv.1705.07874

Mercogliano, P., Rianna, G., Reder, A., Raffa, M., Padulano, R., Essenfelder, A., Mazzoli, P., and Bagli, S. (2021). Flood risk indicators for European cities from 1989 to 2018. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.9d3db0eb

Merz B, Kreibich H, Schwarze R, Thieken A (2010). Assessment of economic flood damage. *Natural Hazards and Earth System Sciences* 10:1697–1724. https://doi.org/10.5194/nhess-10-1697-2010

Merz B, Aerts J, Arnbjerg-Nielsen K, et al. (2014). Floods and climate: emerging perspectives for flood risk assessment and management. *Natural Hazards and Earth System Sciences* 14:1921–1942. https://doi.org/10.5194/nhess-14-1921-2014

Molnar, C. (2022). Interpretable Machine Learning: A guide for making black box models explainable. Online book. Available at: https://christophm.github.io/interpretable-ml-book/ (Accessed online).

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning – A brief history, state-of-the-art and challenges. Online resource. Available at: https://christophm.github.io/interpretable-ml-book/33 (Accessed online).

Moore ID, Grayson RB, Ladson AR (1991). Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrological Processes* 5(1):3–30. https://doi.org/10.1002/hyp.3360050103

Montanari, A. (2007). What do we mean by uncertainty in hydrology and hydrological modelling? *Hydrological Processes*, 21(6), 841–846. https://doi.org/10.1002/hyp.6623

Mosavi A, Ozturk P, Chau K-W (2018). Flood prediction using machine learning models: a literature review. *Water* 10(11):1536. https://doi.org/10.3390/w10111536

Mudashiru RB, Sabtu N, Abustan I (2021). Quantitative and semi-quantitative methods in flood hazard and susceptibility mapping: a review. *Arabian Journal of Geosciences* 14:941. https://doi.org/10.1007/s12517-021-07263-4

Nearing GS, Kratzert F, Sampson AK, Pelissier CS, Klotz D, Frame JM, et al. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research* 57(3):e2020WR028091. https://doi.org/10.1029/2020WR028091

Neal J, Schumann G, Bates P (2012). A subgrid channel model for simulating river hydraulics and floodplain inundation over large and data-sparse areas. *Water Resources Research* 48:W11506. https://doi.org/10.1029/2012WR012514

Norallahi M, Kaboli HS (2021). Urban flood hazard mapping using machine learning models: GARP, RF, MaxEnt and NB. *Natural Hazards* 106:119–137. https://doi.org/10.1007/s11069-020-04453-3

Ochoa-Rodriguez, S., Wang, L., Willems, P., et al. (2015). Impact of spatial and temporal resolution of rainfall inputs on urban hydrodynamic modelling outputs: A multi-catchment investigation. *Journal of Hydrology*, 531, 389–407. https://doi.org/10.1016/j.jhydrol.2015.05.035

Pradhan, B., Lee, S., Dikshit, A., & Kim, H. (2023). Spatial flood susceptibility mapping using an explainable artificial intelligence (XAI) model. Geoscience Frontiers, 14, 101625. https://doi.org/10.1016/j.gsf.2023.101625

Probst P, Wright MN, Boulesteix A-L (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery* 9:e1301. https://doi.org/10.1002/widm.1301

QGIS Development Team (2024). QGIS Geographic Information System. Open Source Geospatial Foundation. https://qgis.org

Reder A, Raffa M, Padulano R, Rianna G, Mercogliano P (2022). Characterizing extreme values of precipitation at very high resolution: An experiment over twenty European cities. *Weather and Climate Extremes*, 35, 100415. https://doi.org/10.1016/j.wace.2022.100407

Rentschler J, Salhab M, Jafino BA (2022). Flood exposure and poverty in 188 countries. *Nature Communications* 13:3527. https://doi.org/10.1038/s41467-022-30727-4

Riley SJ, DeGloria SD, Elliot R (1999). A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences* 5(1–4):23–27.

Sanders, B. F. (2007). Evaluation of on-line DEMs for flood inundation modeling. *Advances in Water Resources*, 30(8), 1831–1843. https://doi.org/10.1016/j.advwatres.2007.02.005

Samela C, Albano R, Sole A, Manfreda S (2018). A GIS tool for cost-effective delineation of flood-prone areas. *Computers, Environment and Urban Systems* 70:43–52. https://doi.org/10.1016/j.compenvurbsys.2018.01.013

Seleem O, Ayzel G, Bronstert A, Heistermann M (2023). Transferability of data-driven models to predict urban pluvial flood water depth in Berlin, Germany. *Natural Hazards and Earth System Sciences* 23:809–822. https://doi.org/10.5194/nhess-23-809-2023

Schumann GJ-P, Bates PD, Apel H, Aronica GT (2018). Global flood hazard mapping, modeling, and forecasting: challenges and perspectives. In: *Flood Risk Science and Management*. Wiley-Blackwell. https://doi.org/10.1002/9781119217886.ch14

Schubert, J. E., & Sanders, B. F. (2012). Building treatments for urban flood inundation models and implications for predictive skill and modeling efficiency. *Advances in Water Resources*, 41, 49–64. https://doi.org/10.1016/j.advwatres.2012.02.012

Shen, C. (2018). A transdisciplinary review of deep learning research for water resources. *Water Resources Research*, 54(11), 8558–8593. https://doi.org/10.1029/2018WR022643

Shirzadi, A., Salvati, A., Hajizadeh Tahan, M., Shahabi, H., Jafari Nodoushan, E., Ramezani, M., Hashim, M., & Clague, J. J. (2020). Urban flood susceptibility mapping using deep and machine learning algorithms as a management tool: A case study of Sanandaj City, Iran. *Science of the Total Environment*, 701, 134813. https://doi.org/10.1016/j.scitotenv.2019.134813

Tehrany MS, Pradhan B, Jebur MN (2015). Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. *Stochastic Environmental Research and Risk Assessment* 29:1149–1165. https://doi.org/10.1007/s00477-015-1021-9

Teng J, Jakeman AJ, Vaze J, Croke BFW, Dutta D, Kim S (2017). Flood inundation modelling: a review of methods, recent advances and uncertainty analysis. *Environmental Modelling & Software* 90:201–216. https://doi.org/10.1016/j.envsoft.2017.01.006

U.S. Environmental Protection Agency (EPA) (2017). *Storm Water Management Model (SWMM) User's Manual, Version 5.1*. U.S. Environmental Protection Agency. https://www.epa.gov/water-research/storm-water-management-model-swmm

U.S. Army Corps of Engineers (2016). *Hydrologic Modeling System HEC-HMS, User's Manual*. U.S. Army Corps of Engineers, Hydrologic Engineering Center. https://www.hec.usace.army.mil/software/hec-hms/

U.S. Army Corps of Engineers (2016). *Hydrologic Engineering Center's River Analysis System (HEC-RAS) Version 5.0, User's Manual*. U.S. Army Corps of Engineers, Hydrologic Engineering Center. https://www.hec.usace.army.mil/software/hec-ras/

Wang Y, Hong H, Chen W, et al. (2019). A hybrid GIS multi-criteria decision-making method for flood susceptibility mapping at Shangyou, China. *Remote Sensing* 11(1):62. https://doi.org/10.3390/rs11010062

Weiss A (2001). Topographic position and landforms analysis. Poster presented at the ESRI User Conference, San Diego, CA.

Westra S, Fowler HJ, Evans JP, et al. (2014). Future changes to the intensity and frequency of short-duration extreme rainfall. *Reviews of Geophysics* 52:522–555. https://doi.org/10.1002/2014RG000464

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences* (3rd ed.). Academic Press, Amsterdam. ISBN: 978-0-12-385022-5. Publisher page (Elsevier): https://www.sciencedirect.com/book/9780123850225/statistical-methods-in-the-atmospheric-sciences.

Zhang J, Guo W, Chang SW, Nguyen DD, Ngo HH (2025). Data-driven innovations in flood hazard assessment with machine learning. *Earth Environmental Sustainability* 1(1):21–41. https://doi.org/10.53941/eesus.2025.100003

Zhang J, Zhang X, Tan X, Yuan X (2022). Extraction of urban built-up area based on deep learning and multi-sources data fusion: the application of an emerging technology in urban planning. *Land* 11:1212. https://doi.org/10.3390/land11081212

Zhang, W., Villarini, G., Vecchi, G. A., & Smith, J. A. (2018). *Urbanization exacerbated the rainfall and flooding caused by hurricane Harvey in Houston.* Nature, 563, 384–388. DOI: 10.1038/s41586-018-0676-z

Yaseen ZM (2021). An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: review, challenges and solutions. *Chemosphere* 277:130126. https://doi.org/10.1016/j.chemosphere.2021.130126