Politecnico di Torino

Master's Thesis in Hydraulic Civil Engineering

Master Degree, 30 credits

Comparative Analysis of Machine Learning Surrogates for Accurate and Real-Time Flood Prediction

A case study for the Po River Basin

Author
Costanza Angelotti

Supervisor
L. Brandimarte
I. Butera
S. Li

Academic Year 2024-2025

Civil Engineering and Urban Management, DD - Politecnico di Torino - CSAMH, Hydraulic Engineering (AF283X), 30 credits

Date: March 31, 2026

Supervisors: L. Bradimarte, I. Butera, S. Li

Royal Institute of technology - KTH

Italian title: Analisi Comparativa di diverse Architetture di Machine Learning per la Previsione Accurata e in Tempo Reale dei Fenomeni di Piena
Italian subtitle: Sviluppo di un Modello Surrogato ad HEC-RAS per una Sezione del Bacino del Fiume Po: Valutazione della Robustezza, dell'Estrapolazione e Quantificazione dell'Incertezza

Acknowledgment

I would like to express my deepest gratitude to my supervisor at Politecnico, Professor Ilaria Butera. Thank you for your invaluable support as I navigated my career path and for guiding me through my very first steps when approaching this thesis. I am especially grateful that you chose to believe in me and support my journey, even when my initial plan for this project was completely different at the beginning.

I am also profoundly grateful to my supervisor at KTH, Professor Luigia Brandimarte, for giving me the opportunity to work with her on this research. This experience has fundamentally changed how I approach hydraulic problems and has the potential to shape the rest of my career. Working with Luigia has taught me an immense amount, particularly the crucial importance of asking the right questions and relentlessly searching for the "why" behind every phenomenon.

Last, but certainly not least, I want to extend my heartfelt thanks to Shicheng Li. Over the past six months, he has patiently helped me navigate every single doubt and question that popped into my mind. He was incredibly kind, open, available, and an absolute pleasure to work with. He truly taught me how inspiring and effortless it can be to work on something you genuinely like and believe in.

Finally, my sincere thanks go to Politecnico di Torino and the Royal Institute of Technology (KTH). Being able to experience both of these distinct and enriching academic worlds has been a profound privilege, providing an environment where my skills and knowledge could continuously expand and flourish.

Abstract

The escalating threat of climate-driven flooding (Core Writing Team et al., 2023) creates an urgent imperative for disaster preparedness systems that are both rapid and precise. Traditional hydrodynamic models, such as HEC-RAS, provide the necessary physical fidelity for mapping inundation but are computationally prohibitive for real-time Early Warning Systems (EWS). Conversely, standard library interpolation methods often fail to capture the non-linear complexities of modern flood dynamics, particularly when accounting for multiple hydrological drivers.

This Master's thesis addresses this critical operational gap by developing and rigorously evaluating Machine Learning (ML) surrogate models capable of predicting flood extents instantly without sacrificing hydraulic accuracy. The study focuses on a high-risk 98 km reach of the Po River (Cremona-Borgoforte), utilizing high-fidelity HEC-RAS simulations to generate a synthetic training dataset. To identify the optimal architecture, five distinct regression models were implemented: Multiple Linear Regression, Support Vector Regression, Random Forest, Extreme Gradient Boosting (XGBoost), and Gaussian Process Regression (GPR) . These models were subjected to a hierarchical evaluation, testing their performance not only on interpolation within the training range but also on their "Extrapolation Robustness" against extreme, unobserved flood events (450-year and 500-year return periods).

The comparative analysis reveals a critical trade-off between local precision and global robustness. While tree-based ensembles (Random Forest and XGBoost) demonstrated superior accuracy in the interpolation regime, they exhibited a structural "saturation" bias during extreme events, systematically underestimating flood depths for discharges exceeding the training domain. In contrast, Gaussian Process Regression emerged as the robustest architecture. It maintained physical consistency during extrapolation and provided essential uncertainty quantification, offering a spatial map of model confidence. Consequently, this research establishes GPR as the viable candidate for operational EWS, enabling civil protection agencies to generate real-time, risk-informed inundation maps.

Keywords: Hydraulic Modeling, Machine Learning Surrogates, Real-time Flood Forecasting, Gaussian Process Regression, Extrapolation Robustness, Uncertainty Quantification, Po River Basin, Early Warning Systems, Flood Inundation Mapping.

Main SDGs: SDG 13 - Climate Action; SDG 11 - Sustainable Cities and Communities; SDG 9 - Industry, Innovation and Infrastructure.

Swedish Abstract

Det ökande hotet från klimatdrivna översvämningar (Core Writing Team et al., 2023) skapar ett akut behov av katastrofberedskapssystem som är både snabba och precisa. Traditionella hydrodynamiska modeller, såsom HEC-RAS, erbjuder den nödvändiga fysikaliska noggrannheten för att kartlägga översvämningar men är för beräkningsmässigt ineffektiva i realtid för tidiga varningssystem (Early Warning Systems, EWS). Samtidigt misslyckas ofta standardmetoder för interpolering av biblioteksdata med att fånga de icke-linjära komplexiteterna i modern översvämningsdynamik, särskilt när hänsyn tas till flera hydrologiska drivkrafter.

Detta examensarbete åtgärdar denna kritiska operationella klyfta genom att utveckla och rigoröst utvärdera surrogatmodeller baserade på maskininlärning (ML), kapabla att förutsäga översvämningsutbredning omedelbart utan att förlora hydraulisk noggrannhet. Studien fokuserar på en högrisksträcka på 98 km av Po-floden (Cremona-Borgoforte) och utnyttjar högprecisionssimuleringar i HEC-RAS för att generera ett syntetiskt träningsdataset. För att identifiera den optimala arkitekturen implementerades fem olika regressionsmodeller: Multipel linjär regression, Support Vector Regression, Random Forest, Extreme Gradient Boosting (XGBoost) och Gaussisk processregression (GPR). Dessa modeller genomgick en hierarkisk utvärdering där deras prestanda testades inte bara vid interpolering inom träningsintervallet utan även på deras "extrapoleringsrobusthet" mot extrema, icke-observerade översvämningshändelser (med återkomsttider på 450 och 500 år).

Den jämförande analysen avslöjar en kritisk avvägning mellan lokal precision och global robusthet. Medan trädbaserade ensembler (Random Forest och XGBoost) uppvisade överlägsen noggrannhet i interpolationsregimen, uppvisade de en strukturell "mättnadsbias" vid extrema händelser och underskattade systematiskt översvämningsdjup för flöden som översteg träningsdomänen. I kontrast framstod gaussisk processregression (Gaussian Process Regression) som den mest robusta arkitekturen. Den bibehöll fysikalisk konsistens även vid extrapolering och gav väsentlig osäkerhetskvantifiering genom att erbjuda en spatial karta över modellens konfidens. Följaktligen etablerar denna forskning GPR som den lämpliga kandidaten för operationella EWS, vilket möjliggör för civilskyddsmyndigheter att generera riskinformerade översvämningskartor i realtid.

Italian Abstract

La crescente minaccia di alluvioni causate dai cambiamenti climatici (Core Writing Team et al., 2023) crea un imperativo urgente per i sistemi di preparazione ai disastri che devono adeguarisi alla rapidità e alla precisione ora a richieste. I modelli idrodinamici tradizionali, come quelli prodotti in HEC-RAS, forniscono la fedeltà fisica necessaria per la mappatura delle inondazioni, ma sono computazionalmente proibitivi per i sistemi di allerta preventiva (Early Warning Systems, EWS) in tempo reale a causa dell'elevato tempo computazionale richiesto. Al contrario, i metodi standard di interpolazione da libreria spesso non riescono a catturare le complessità non lineari delle moderne dinamiche alluvionali, in particolare quando si considerano molteplici fattori idrologici.

La presente tesi di laurea affronta questa critica lacuna operativa sviluppando e valutando modelli surrogati di Machine Learning (ML) in grado di prevedere l'estensione delle inondazioni istantaneamente senza sacrificare l'accuratezza idraulica. Lo studio si concentra su un tratto ad alto rischio di 98 km del fiume Po (Cremona-Borgoforte), utilizzando simulazioni HEC-RAS ad alta fedeltà per generare un set di dati sintetici di addestramento. Per identificare l'architettura ML ottimale, sono stati implementati cinque distinti modelli di regressione: Regressione Lineare Multipla (MLP), Support Vector Regression (SVR), Random Forest, Extreme Gradient Boosting (XGBoost) e Regressione basata su Processi Gaussiani (GPR). Questi modelli sono stati sottoposti a una valutazione gerarchica, testando le loro prestazioni non solo sull'interpolazione all'interno del range di addestramento, ma anche sulla loro "robustezza all'estrapolazione" contro eventi di piena estremi e non osservati (tempi di ritorno di 450 e 500 anni).

L'analisi comparativa di questi modelli rivela un compromesso critico tra precisione locale e robustezza globale. Mentre gli algoritmi basati su modelli ad "albero" (Random Forest e XGBoost) hanno dimostrato un'accuratezza superiore nel regime di interpolazione, hanno anche mostrato un "bias di saturazione" strutturale durante gli eventi estremi, sottostimando sistematicamente le profondità di inondazione per portate eccedenti al dominio di addestramento. Al contrario, la Regressione basata su Processi Gaussiani è emersa come l'architettura più robusta. Essa ha mantenuto la coerenza fisica durante l'estrapolazione e ha fornito una quantificazione essenziale dell'incertezza, offrendo una mappa spaziale della confidenza del modello. Di conseguenza, questa ricerca stabilisce il GPR come il candidato praticabile per EWS operativi, possibilmente consentendo alla protezione civile di generare mappe di inondazione.

Contents

List of Figures

List of Tables

1 Introduction

The escalating threat of flooding (Core Writing Team et al., 2023) represents one of the most immediate and costly consequences of global climate change, fundamentally challenging human safety, economic stability, and infrastructure resilience worldwide. The urgency of this issue is evidenced by the Intergovernmental Panel on Climate Change (IPCC), which confirms that the frequency and intensity of heavy precipitation events have observably increased since the 1950s, a trend that is unequivocally linked to human influence through greenhouse gas emissions (Core Writing Team et al., 2023). This intensification of the hydrological cycle translates directly into a higher incidence of rain-generated local flooding, demanding a radical transformation in disaster preparedness and mitigation strategies.

The observed impacts of climate-driven floods on human systems are systemic and widespread. Urban areas, which concentrate populations and critical infrastructure, are particularly vulnerable. Extreme flood events compromise essential services such as transportation, water, and sanitation systems, leading to cascading failures and significant economic losses (Core Writing Team et al., 2023). These damages are projected to multiply as global temperatures rise: direct flood damages are estimated to be 1.4 to 2 times higher at $2°C$ and up to 3.9 times higher at $3°C$ of warming compared to the $1.5°C$ level, assuming no effective adaptation (Core Writing Team et al., 2023). Furthermore, the adverse effects are often concentrated in vulnerable communities, exacerbating existing social inequalities and complicating recovery efforts (Core Writing Team et al., 2023). To mitigate these severe economic and social costs, the IPCC stresses the importance of non-structural measures like early warning systems (EWS), recognizing that timely and accurate predictions are vital components of climate resilience (Core Writing Team et al., 2023). However, achieving the required level of accuracy and speed in real-time flood forecasting is a formidable challenge.

Accurate flood prediction necessitates the rapid and precise modeling of complex, dynamic hydraulic and hydrological processes. Historically, this task has relied on physics-based, high-resolution models, which simulate water movement based on governing equations (i.e. HEC-RAS). While robust in theory, these traditional high-resolution flood models are notoriously too slow for real-time prediction, a critical limitation when an early warning system requires lead times of hours, not days (Wang et al., 2025).

To speed up predictions, current industry practice relies on interpolating results from a saved database of flood maps (Wang et al., 2025). This library-interpolation approach, though faster, faces significant obstacles when dealing with the increasing complexity of contemporary flood events. As highlighted by Wang et al. (2025), a simple extension of these traditional interpolation methods becomes "overly complex" when tasked with accounting for multiple flood drivers (such as the interaction of river flows, tributary inputs, and tidal influences) and various flood attributes (such as the shape and timing of the hydrographs) (Wang et al., 2025). Earlier review

work in the field has already emphasized the ongoing challenges in flood inundation modelling, particularly regarding methods, recent advances, and uncertainty analysis (Teng et al., 2017). The need for rapid and efficient models that can capture the complex relationship between the terrain and the inundation wave, remains a core research focus (Teng et al., 2019).

The gap between the escalating flood risk, the need for rapid EWS, and the computational limitations of traditional methods creates an imperative for the adoption of advanced computational techniques. This research identifies Machine Learning (ML) as the critical solution to bridge this gap. By leveraging ML, we can hypothesize to achieve the necessary speed for real-time applications without sacrificing the predictive fidelity required for effective disaster preparedness.

ML methods excel at identifying complex, non-linear relationships within vast datasets, a critical advantage over traditional models when attempting to rapidly emulate dynamic flood behaviour. The integration of ML techniques provides a pathway to achieve rapid prediction of flood inundation by interpolation between flood library maps for real-time applications (Wang et al., 2022). Consequently, the primary benefit and purpose of this study is to demonstrate how ML surrogates can instantly predict inundation extents, replacing computationally heavy simulations to support operational decision-making

Furthermore, ML is not intended merely as a replacement for physics-based models but as a powerful enhancer. Research is increasingly focused on coupling traditional hydraulic models, such as HEC-RAS, with Artificial Intelligence (AI) to enhance the assessment of river morphodynamics under changing flow regimes (Qureshi et al., 2025). This integrated approach capitalizes on the physical realism of hydraulic models while leveraging the pattern recognition and computational efficiency of AI, thereby improving the overall robustness of disaster preparedness systems (Qureshi et al., 2025).

To validate this approach, the research is applied to a strategic case study: the Po River, specifically the 98 km reach between Cremona and Borgoforte. This area is classified as a "High Hydraulic Risk" zone, characterized by a complex embankment system and a low water-surface gradient that makes it particularly vulnerable to the intensification of extreme events.

1.1 Research question

In the context of escalating flood risk and the recognized need for rapid, high-fidelity forecasting, this thesis aims to bridge the gap between complex hydrodynamic modeling and operational disaster preparedness. Traditionally, real-time warning systems have relied on library interpolation methods over pre-computed flood maps, an approach that rapidly becomes overly complex when accounting for multiple hydrological drivers and attributes (Wang et al., 2025). To surpass this computational and methodological limitation, this research proposes the development

of data-driven surrogate models. This involves training these algorithms using the outputs of a high-fidelity hydrodynamic model to effectively emulate the complex, non-linear physical relationships governing inundation. Specifically, this study evaluates both traditional parametric statistical models (which fit data to a predefined equation) and true Machine Learning (ML) approaches, defined here as non-parametric models with flexible functional forms that learn their structure directly from the data. By leveraging these flexible models, the goal is to achieve rapid predictions potentially suitable for real-time Early Warning Systems (EWS) (Qureshi et al., 2025).

The central Research Question guiding this investigation is:

How effectively can Machine Learning algorithms accurately and robustly predict Water Surface Elevations (WSE) as hydraulic parameter to generate flood inundation maps for the chosen case study, and how do their comparative performance metrics justify the adoption of a specific model for real-time prediction?

To address this question, the research will pursue the following specific objectives:

- Model Comparison: Systematically evaluate and compare two distinct modeling paradigms: a fixed-form parametric statistical model (Multiple Linear Regression) utilized to establish a baseline and quantify system non-linearity, against true Machine Learning approaches characterized by flexible functional forms (Support Vector Regression, Random Forest, XGBoost, and Gaussian Process Regression) that learn the underlying hydraulic structure directly from the data.
- Performance Metrics: Quantify the comparative performance using industry-standard metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2) and Mean Bias Error (MBE).
- Feasibility Assessment: Demonstrate the computational efficiency of the optimal ML framework to validate its suitability for deployment in operational flood forecasting by emergency agencies.

The proposed surrogate models are designed to directly predict the Water Surface Elevation $H(Q, x, y)$ as a function of discharge and spatial coordinates. Subsequently, flood inundation maps, including both flood depth and boundary extents, are derived by intersecting the predicted $H(Q, x, y)$ surface with the local Digital Terrain Model (DTM).

The thesis is structured to systematically address these challenges. Following this introduction, Chapter 2 - Methodology details the generation of high-fidelity "ground truth" data using HEC-RAS and the implementation of five distinct ML architectures. Chapter 3 - Case Study provides an in-depth analysis of the Po River's hydrological and topographical characteristics

used for calibration and input data. Finally, Chapter 4 - Results and Discussion presents a comparative analysis of the models' accuracy and efficiency, leading to the Conclusions in Chapter 5 regarding the feasibility of these tools for real-time warning systems.

2 Methodology

The preparation and pre-processing of input data constitute a critical phase in the development of effective and robust machine learning models. The primary objective of this research is to evaluate the performance of five distinct machine learning architectures and conduct a comparative analysis of their results. To achieve this, a systematic, step-by-step methodology, illustrated in the flow chart in Fig. 2, was implemented. The workflow is anchored by an initial data generation phase using the HEC-RAS software. This step is pivotal, as it allows for the elaboration of a robust input dataset derived from hydraulic simulations. These simulations provide the foundational "ground truth" upon which a larger volume of synthetic data is subsequently generated for model training. Consequently, the research methodology adopted in this study is bifurcated into two main phases:

- Hydraulic Modeling and Data Generation: The processing of input data leveraging the HEC-RAS environment.
- Machine Learning Implementation: The subsequent training, validation, and evaluation of the predictive models. These phases are unpacked in detail in the following subsections.

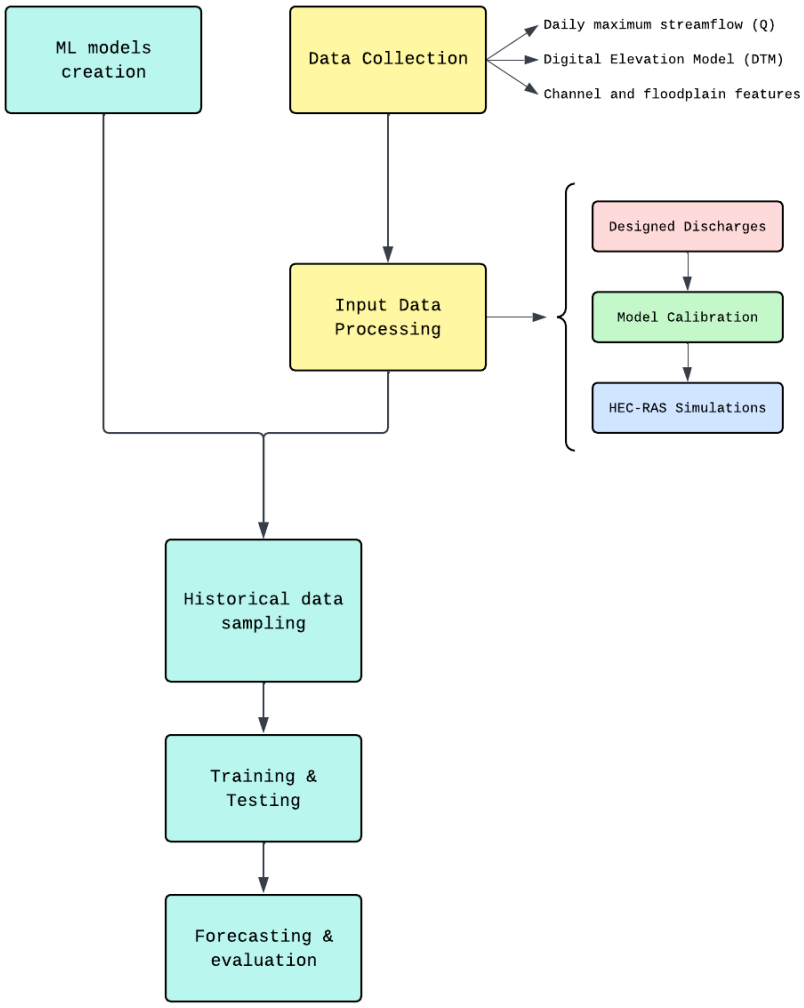


Figure 2: General Process Flow Chart

2.1 HEC-RAS Hydraulic Models

The foundational phase of this research involves the application of the Hydrologic Engineering Center's River Analysis System (HEC-RAS), a sophisticated software suite developed by the U.S. Army Corps of Engineers designed for computational fluid dynamics. HEC-RAS is an integrated system that simulates the hydraulics of water flow through natural rivers and floodplains by solving the governing equations for water movement. This model serves as the "ground truth" for the Po River's hydraulic behavior, providing the water surface elevation (WSE) and other useful outputs necessary to train and validate the subsequent Machine Learning architectures.

HEC-RAS allows for several types of river analysis, primarily focusing on the interaction between model parameters and water profiles:

- Steady Flow Water Surface Profiles: This component calculates water surface profiles for gradually varied flow, making it a standard tool for flood insurance studies and evaluating levee performance.
- 1D Hydrodynamic Modeling: The software simulates flow along a single dimension (the channel centerline), which is highly effective for large river reaches where flow is predominantly longitudinal.
- RAS Mapper: This specialized geospatial interface within HEC-RAS is used to combine geometric data with simulation results to generate detailed flood inundation, depth, and velocity maps

To calculate the water surface profiles for steady flow, HEC-RAS solves the one-dimensional energy equation using the Standard Step Method. The software computes the energy balance from one cross-section to the next, moving upstream for subcritical flow. The governing equation is defined as follows (Brunner, 2021):

$$Z_2 + Y_2 + \frac{\alpha_2 V_2^2}{2g} = Z_1 + Y_1 + \frac{\alpha_1 V_1^2}{2g} + h_e \tag{1}$$

where:

- Z_1, Z_2 are the elevations of the main channel inverts;
- Y_1, Y_2 represent the depths of the water at cross-sections 1 and 2;
- V_1, V_2 are the average velocities (total discharge / total flow area);
- α_1, α_2 are the velocity weighting coefficients;
- g is the gravitational acceleration;
- h_e is the energy head loss.

The term h_e is critical for this study, as it represents the energy loss due to friction, contraction and expansion. HEC-RAS calculates friction loss (h_f) using Manning's equation:

$$h_f = L\bar{S}_f = L\left(\frac{nQ}{AR^{2/3}}\right)^2 \tag{2}$$

where L is the discharge-weighted reach length, $\bar{S}_f$ is the representative friction slope between two sections, and n is the Manning's roughness coefficient. Since n is the primary parameter governing energy loss and, consequently, the resulting water surface elevation, its accurate estimation via calibration is essential for ensuring the reliability of the hydraulic model.

The geometry of the HEC-RAS model was not developed using RAS Mapper, although this tool can be very useful for extracting river cross-sections directly from a Digital Terrain Models (DEMs) by importing data from QGIS. Instead, this study employed a pre-defined geometric dataset consisting of 68 cross-sections along the 98 km reach of the Po River between Cremona and Borgoforte, representing a simplified and optimized version of the 88-section geometry used in the previous study by Bradimarte and Di Baldassarre (2012) of the same reach. Cross-sections are assigned stationing numbers that increase in the upstream direction; the most downstream point of the system is designated as Station 0.

The model environment was configured to simulate gradually varied steady flow across the study area, so it is necessary to set the boundary conditions and the Manning's Roughness coefficient:

- Boundary conditions: the upstream boundary at Cremona is defined by the estimated discharge (Q); the downstream boundary at Borgoforte utilizes the Normal Depth method, which relies on a calculated mean bed slope (S_0) of approximately 0.12
- Manning's Roughness: friction parameters were assigned to the main channel and floodplains based on land use characteristics and standard literature ranges.

While the boundary condition data are relatively straightforward to determine, discharges (Q) are derived from flow measurements at the Cremona gauging station, and the channel slope is calculated geometrically from the DEM, there is no precise estimation of the Manning's roughness coefficient. For this reason, a calibration process was required to identify the most appropriate values for both the main channel and the floodplain. As more detail explained in paragraph 3.2.3, the Manning's roughness coefficients were iteratively adjusted within physically plausible ranges between 0.02 and 0.06 for the main channel, and between 0.06 and 0.14 for the floodplains. The optimization aimed to minimize the Mean Absolute Error (MAE) between simulated water surface elevations and observed historical high-water marks.

The MAE quantifies the average magnitude of the errors in a set of predictions without considering their direction. Indeed, the MAE measures the average vertical distance between simulated water surface profiles and observed high-water marks expressing it in the same units as the input data (meters) and providing a physically intuitive assessment of model accuracy. Unlike the Root Mean Square Error (RMSE), which squares differences and emphasizes outliers, MAE weights all errors equally. This is beneficial in river reaches where local measurement errors in high-water marks may be present but should not disproportionately bias the entire Manning's n selection.

Upon the completion of the calibration phase and the identification of the optimal Manning's roughness coefficients, the upstream boundary condition was systematically modified to generate 20 distinct steady-flow simulations, covering a broad range of discharge values as detailed in

Section 3.2.2. The primary output required for the subsequent predictive modeling was the geospatial definition of the flood extent, represented in HEC-RAS by an inundation bounding polygon. Unlike the dense point cloud used to define the channel bathymetry, this polygon is constructed by identifying the specific intersection points where the computed water surface elevation meets the terrain profile at each cross-section (Brunner, 2021). These "wet/dry" limits are then interpolated longitudinally between cross-sections, guided by the river centerline and bank lines, to form a continuous perimeter that delineates the maximum flood extent while excluding the internal bathymetric points (U.S. Army Corps of Engineers, 2023). This derived boundary data was exported as GIS-formatted text files, preserving the spatial integrity of the model in the UTM-WGS84 coordinate system. A data parsing procedure was then implemented to extract the coordinate pairs (Northing and Easting) and their corresponding water surface elevations from these files.

The processed data were consolidated into a structured Excel database, organized by simulation sheets to ensure consistency. Within this dataset, every entry is characterized by four primary attributes: Discharge (Flow), Northing (N), Easting (E), and Water Surface Elevation (H). This comprehensive tabular matrix constitutes the foundational dataset used to train and validate the Machine Learning architectures.

In addition to the tabular datasets, the simulations yielded geospatial outputs in the form of flood inundation maps. These were generated within the HEC-RAS environment by superimposing the computed Water Surface Elevation (WSE) polygons onto the Digital Elevation Model (DEM) of the study area.

2.2 Machine Learning Models

Following the generation of the high-fidelity hydraulic dataset, the research methodology progresses to the second phase: the development of data-driven surrogate models. The primary objective of this stage is to establish a computational framework capable of predicting water surface elevations (H) with high accuracy while significantly reducing the computational burden associated with physical hydrodynamic simulations.

To achieve this objective, this study selected five distinct machine learning architectures. This selection was designed to cover the principal families of regression algorithms available in the literature, starting from a conceptually simpler one to the more complex, enabling a comparative analysis of their strengths and limitations in the context of hydraulic modeling. The chosen models range from simple deterministic baselines to advanced probabilistic and ensemble frameworks:

1. Multiple Linear Regression (MLR): Included as a baseline to benchmark the performance of complex models. It serves to quantify the degree of non-linearity in the hydraulic system; if the complex models do not significantly outperform MLR, the use of high computational

resources is not justified.

2. Support Vector Regression (SVR): Selected for its proven capability to handle non-linear boundaries in high-dimensional spaces through kernel mapping (Smola and Schölkopf, 2004), offering a robust deterministic alternative to linear methods.

3. Random Forest (RF): A "Bagging" (Bootstrap Aggregating) ensemble method. It was chosen for its robustness against overfitting and its ability to reduce variance by averaging multiple uncorrelated decision trees (Breiman, 2001).

4. eXtreme Gradient Boosting (XGBoost): A "Boosting" ensemble method. Chosen as the current state-of-the-art for tabular data, it builds trees sequentially to correct previous errors (reducing bias) and includes advanced regularization techniques to improve generalization (Chen and Guestrin, 2016).

5. Gaussian Process Regression (GPR): A non-parametric, Bayesian approach chosen for its unique ability to provide probabilistic predictions (uncertainty quantification). Specifically, this study utilizes the Matérn kernel, which allows for fine-grained control over the smoothness of the resulting function, making it particularly suitable for modeling physical surfaces like water profiles (Rasmussen and Williams, 2006).

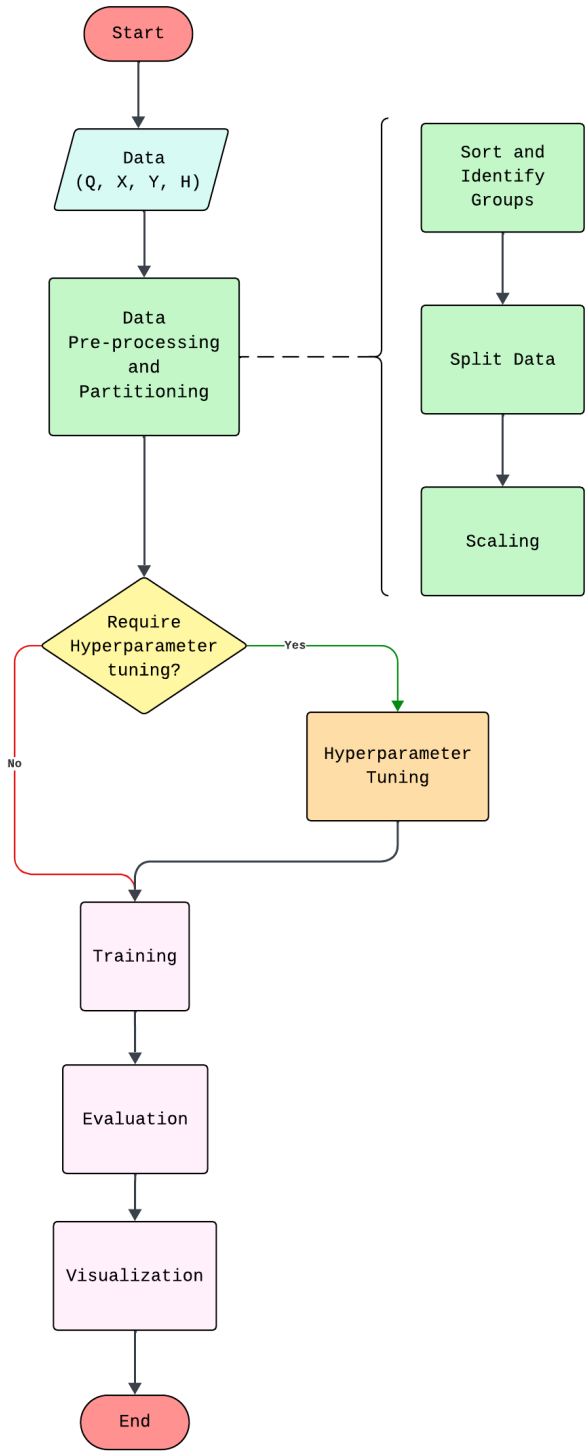


Figure 3: Workflow of Machine Learning Modeling for flood forecasting

The implementation was developed using the Python programming language, utilizing the Pandas library for data manipulation and Scikit-Learn and XGBoost for predictive modeling. To ensure a rigorous and fair comparison between the five selected architectures, a standardized computational workflow was adopted. As illustrated in the process logic shown in Fig.3, the code structure can be resumed in 4 main processes:

1. **Data Pre-processing and Partitioning:** the raw tabular data extracted from the HEC-RAS simulations were first ingested and cleaned. To evaluate the models' generalization capabilities under different hydrological conditions, the dataset was stratified into three distinct subsets:

   - Training Set: Comprising 15 steady-flow simulations, used to learn the mapping function $f(Q, X, Y) \rightarrow H$.
   - Testing Set (Interpolation): Containing three specific return periods (10, 100, and 300 years) to assess performance on unseen data within the training range.
   - Extrapolation Set: Composed specifically of the two highest return periods (450 and 500 years). This selection was driven by two factors: first, the hydraulic necessity to stress-test the models' ability to capture the non-linear physical transition from in-bank to overbank flow (Chow, 1959); and second, to rigorously evaluate the known limitations of tree-based algorithms (e.g., Random Forest, XGBoost) in extrapolating beyond the boundaries of the training domain (Hastie et al., 2009).

2. **Feature Scaling:** Given the different magnitudes of the input variables (e.g., Discharge in m^3/s vs. UTM Coordinates in m), a standardization procedure (Z-score normalization) was applied to the training features to prevent numerical bias and ensure faster convergence during optimization.

3. **Model Optimization and Training:** At this stage, the workflow diverges based on the model architecture. For the parametric baseline (Multiple Linear Regression), the model was fitted directly to the training data to derive the regression coefficients. Conversely, for the non-parametric and ensemble methods (SVR, Random Forest, XGBoost, GP), a distinct Hyperparameter Tuning phase was interposed. This process utilized Grid Search with Cross-Validation to identify the optimal configuration of internal parameters (e.g., tree depth, learning rate) before the final training. During the training phase, the primary optimization objective for the machine learning architectures (with the exception of SVR, which utilizes Structural Risk Minimization) was to minimize the Mean Squared Error (MSE) loss function over the N training data points. The training criterion is explicitly defined as:

$$L = \min \frac{1}{N} \sum_{i=1}^{N} (H_{pred,i} - H_{HEC-RAS,i})^2 \tag{3}$$

By squaring the residuals, the optimization procedure heavily penalizes large local errors, ensuring the algorithms prioritize capturing the peak water surface elevations critical for flood mapping.

4. **Evaluation and Visualization:** Finally, the trained models were evaluated across all three data subsets using statistical metrics (RMSE, R^2, MAE) and geo-spatial error mapping.

2.2.1 Linear Regression Model

This study evaluates the capability of statistical and data-driven models to predict $H(Q, x, y)$ as simulated by HEC-RAS. To establish a baseline and evaluate the non-linearity of the system, a traditional parametric statistical model, called Multiple Linear Regression, is employed to determine parameter values for a fixed equation. This is compared against true Machine Learning approaches, specifically non-parametric regressions with flexible functional forms, that learn the model structure directly from the data. These flexible models include Random Forest, XGBoost, and Gaussian Process Regression.

In the context of this study, MLR is employed to establish a direct functional mapping between the primary hydrological and topographical drivers, discharge (Q) and coordinates (X, Y) and the target variable, the Water Surface Elevation (H). To achieve these objectives, the model operates based on the assumption of a linear relationship between the independent input features and the dependent output variable. The specific regression function utilized in this study can be expressed mathematically as:

$$H = \beta_0 + \beta_1 Q + \beta_2 X + \beta_3 Y + \epsilon \tag{4}$$

where:

- H represents the predicted Water Surface Elevation (the target variable)
- β_0 is the y-intercept (bias term).
- $\beta_1, \beta_2, \beta_3$ are the regression coefficients representing the weight and influence of discharge, the coordinate X and the coordinate Y, respectively.
- ϵ denotes the residual error.

A critical theoretical distinction within this methodology is the classification of MLR as a parametric model, as opposed to the non-parametric and ensemble approaches (e.g., Gaussian Process Regression, Random Forest) utilized in the sections below.

In the parametric framework used by the MLR model, it is assumed a priori that the form of the function $f(X)$ is linear. The "learning" process is limited to estimating a fixed, finite set of parameters (the coefficients β) that define this pre-determined shape (Hastie et al., 2009).

While computationally efficient, this imposes a high "inductive bias": if the true physical relationship is far from linear, the model will inherently under-perform regardless of how much data is provided. Conversely, non-parametric models, such as the Gaussian Process Regression (GPR) used later in this study, do not assume a global functional form defined by a finite set of parameters. Instead, the complexity of a non-parametric model grows with the sample size, allowing the data to determine the shape of the function rather than forcing it into a linear mold. As noted by Rasmussen and Williams (2006), non-parametric methods effectively "let the data speak", which in this study is particularly relevant given the strong spatial variability of the floodplain geometry, by assuming that similar inputs yield similar outputs, rather than assuming the outputs follow a specific equation. In this study, MLR minimizes the sum of squared residuals (Ordinary Least Squares) to fit its fixed parameters. The resulting performance gap between this rigid parametric structure and the flexible non-parametric alternatives is used, in this study, as a quantitative measure of the hydraulic non-linearity in the Po River system.

In the domain of hydraulic engineering and flood forecasting, Multiple Linear Regression (MLR) has historically served as a standard statistical tool due to its interpretability and computational efficiency (Garen, 1992; Mosavi et al., 2018). Because it directly maps the linear relationship between meteorological inputs and hydrological responses without the heavy computational burden of physically-based models, MLR is frequently utilized as the baseline benchmark against which the performance of more complex, non-linear machine learning architectures is evaluated (Abrahart and See, 2000; Latt and Wittenberg, 2014).

While accurate flood prediction typically necessitates physics-based models that solve complex governing equations, such as the one-dimensional energy equation utilized by HEC-RAS, these methods are often computationally expensive and time consuming for real-time prediction.

In facts, MLR has been taken in consideration as a rapid alternative, capable of generating predictions almost instantaneously once trained. However, as noted in literature it is frequently relegated to the role of a baseline to validate the necessity of advanced Machine Learning (ML) techniques.

For this reason, MLR was selected as the inaugural model of the study. Its role is not predictive excellence, but methodological comparison: it acts as a parametric baseline to evaluate the complexity of the dataset, providing a reference point for error.

If advanced ensemble models (like Random Forest or XGBoost) or probabilistic models (Gaussian Processes) do not yield significantly superior accuracy compared to this simple linear baseline, the deployment of high-computational resources is not justified. Also, the performance gap between the MLR and non-linear models serves as a quantitative proxy for the degree of non-linearity present in the Po River's hydraulic behavior.

Despite its utility as a benchmark, the MLR architecture faces significant limitations when applied to the complex physics such flood inundation mapping.

- Hydraulic processes are inherently non-linear. While the HEC-RAS software is designed to navigate these complexities using robust, physics-based governing equations, approximating such dynamic behavior with a simple Linear Regression model (Equation 4) presents immediate theoretical limitations. The most critical disconnect arises at the physical threshold between in-bank and overbank flow. When a river breaches its banks, the relationship between discharge and water level undergoes an abrupt, non-linear shift. A linear model, rigidly constrained to a constant slope, simply lacks the mathematical flexibility to adapt to these distinct hydraulic regimes, rendering it incapable of capturing the complex reality of the flood wave (Chow, 1959).

- Although linear models can mathematically project values indefinitely, they assume that the linear trend observed in the training data remains valid for extreme, unseen events (e.g., 500-year return period). In physical reality, extreme discharges often trigger complex interactions with the floodplain that diverge from the linear trend.

2.2.2 Support Vector Regression Model

Support Vector Regression (SVR) represents the second deterministic architecture employed in this study, selected to address the non-linear complexities of the hydraulic system that the linear baseline (MLR) cannot resolve. SVR is an adaptation of the well-known Support Vector Machine (SVM) algorithm. While standard SVM is a classification algorithm designed to predict discrete classes which are distinct, non-overlapping categories, SVR is modified to predict continuous variables (Smola and Schölkopf, 2004).

In statistical learning, the distinction between these variable types dictates the modeling approach:

- Discrete Classes: These refer to qualitative outputs where the target variable belongs to a finite set of categories (e.g., Class A or Class B). The goal is to find a decision boundary that best separates these classes (Hastie et al., 2009).

- Continuous Variables: These refer to quantitative outputs that can assume any numerical value within a range (e.g., a water surface elevation of 13.45 meters). The goal is to estimate a function that fits the data points with minimal error.

In this study, the target variable (H) is continuous; therefore, the regression variant (SVR) is required rather than the classification variant.

The theoretical framework for SVR used in this study is derived from the foundational work of Smola and Schölkopf (2004). Unlike the Ordinary Least Squares (OLS) method used in MLR, which seeks to minimize the sum of squared errors, SVR operates on the principle of Structural Risk Minimization (SRM). Its objective is to find a function $f(X)$ that deviates from the actual target values y_i by a value no greater than ϵ for each training point x_i, while simultaneously keeping the function as "flat" as possible to ensure generalization.

The governing minimization problem is defined as:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{5}$$

Subject to the constraints:

$$\begin{cases} y_i - (w^T\phi(x_i) + b) \leq \epsilon + \xi_i \\ (w^T\phi(x_i) + b) - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

where:

- $||w||^2$ represents the complexity (flatness) of the model. Minimizing this term prevents overfitting (Smola and Schölkopf, 2004).
- C is the regularization hyperparameter that governs the trade-off between the model's complexity and the amount of deviation allowed (bias-variance tradeoff) (Smola and Schölkopf, 2004).
- ξ_i, ξ_i^* are slack variables that measure the deviation of training samples that fall outside the ϵ-insensitive zone.
- ϵ defines a "tube" around the regression function within which errors are ignored. This characteristic makes SVR robust to noise, as small residuals do not affect the model parameters.

To capture non-linear hydraulic behaviors (such as the abrupt transition from channel to floodplain), this study employs the Radial Basis Function (RBF) kernel. As described Smola and Schölkopf (2004), the kernel trick maps the input vectors into a high-dimensional feature space where non-linear relationships become linear. This capability allows SVR to model complex stage-discharge relationships that MLR fails to capture.

SVR is widely applied in hydrology for rainfall-runoff modeling and flood stage forecasting because of its robustness to noise and ability to generalize well on smaller datasets compared to deep neural networks (Teng et al., 2017). In this study, SVR serves as a bridge between the simple linear baseline and the complex ensemble methods. It is particularly useful for testing whether a deterministic non-linear mapping is sufficient to reduce the prediction error (RMSE) observed in the linear model.

The choice to implement SVR was driven by its theoretical strength in handling non-linear boundaries (Smola and Schölkopf, 2004). However, the model faces specific limitations:

- Computational Cost: The training complexity scales quadratically or cubically with the number of samples (N) (Smola and Schölkopf, 2004), making it computationally expensive

for the large datasets generated by high-resolution HEC-RAS simulations.

- Extrapolation: SVR with an RBF kernel typically struggles with extrapolation (predicting beyond the training range), as the kernel values decay to zero for inputs far from the support vectors.

2.2.3 Random Forest Model

Random Forest (RF) represents the first ensemble learning architecture implemented in this study, marking a fundamental methodological shift from the single deterministic models (MLR and SVR) to aggregate-based approaches. While MLR and SVR rely on constructing a single, global predictive model, either through a fixed parametric equation or a kernel-based mapping in high-dimensional space, Random Forest operates on the principle of recursive partitioning.

It divides the feature space into smaller, localized regions rather than attempting to fit a single function to the entire dataset. To achieve this, RF employs the "Bagging" (Bootstrap Aggregating) technique, a family of algorithms that generates multiple independent models (in this case, decision trees) and aggregates their predictions. This aggregation is critical for addressing the primary weakness of single decision trees: high variance. By averaging the outputs of diverse trees, the ensemble reduces the model's sensitivity to noise and prevents the overfitting often observed in individual, complex models (Hastie et al., 2009).

The core mechanic of the algorithm is visualized in Fig. 4, which depicts the flow of a single input sample through the ensemble permitting the reduction of variance without increasing bias.

The diagram highlights three critical components of the inference process:

- Parallel Tree Construction: the input data is fed simultaneously into multiple independent estimators (labeled Decision Tree 1, 2, and 3). Each tree possesses a unique structure because it was trained on a different bootstrap sample of the hydraulic history. While a single decision tree is highly sensitive to small fluctuations in the training data, averaging the results of many trees stabilizes the prediction. The algorithm operates through two distinct randomization processes at the beginning of the process, described by Breiman (2001):

  1. Bootstrap Sampling: for each tree b in the forest (where $b = 1$ to B), a random subset of the training data is sampled with replacement.
  2. Feature Randomness: At each node split within a tree, the algorithm considers only a random subset of the input features (e.g., Discharge, X, Y) rather than the full set.

  This de-correlates the trees, ensuring they capture different aspects of the hydraulic system.

- Divergent Decision Paths: the dashed lines in the figure trace the specific trajectory of a single input sample (e.g., a specific discharge event) as it navigates through the trees.

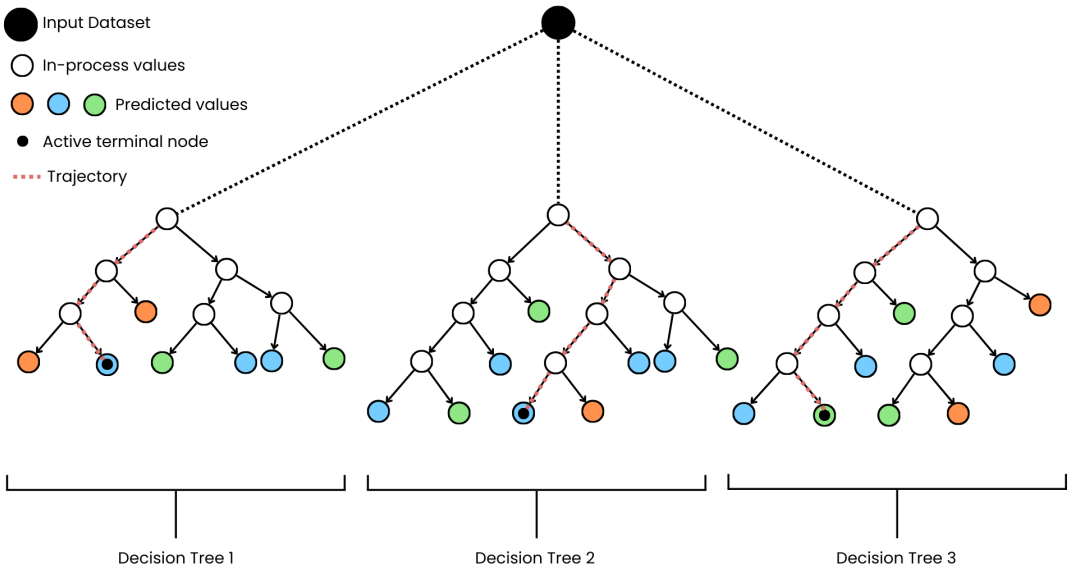


Figure 4: Random Forest core mechanic diagram. The schematic illustrates the parallel construction of decision trees, divergent decision trajectories (dashed lines) for a given input, and the final aggregation of active terminal nodes. Concept adapted from IBM Cloud Education IBM Cloud Education (2020b).

- The Black Dots: the black dots located inside specific circles at the bottom of the trees mark the active terminal node (leaf) reached by the input data for that specific tree. For instance, in Tree 1, the input follows the left-most path to a blue node, whereas in Tree 2, the same input follows a central path to a distinct node.
- Leaf Node Colors (Blue, Green, Orange): The colored circles at the bottom represent the predicted values stored in the leaves. The variation in color, Blue, Green, and Orange, visually demonstrates the diversity of the ensemble: different trees may predict slightly different Water Surface Elevations for the same input based on their unique training. For example, an Orange node might represent a higher water level prediction, while a Blue node represents a lower one.
- Aggregation :The final step is represented by the single node which is the Result. In this regression study, the model calculates the average of the values in the "black dot" leaves through the generalized prediction function for a Random Forest Regressor:

$$\hat{H}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{6}$$

where:

- $\hat{H}(x)$ is the predicted Water Surface Elevation.
- B is the total number of trees (estimators) in the forest.
- $T_b(x)$ is the output of the b-th decision tree grown on a bootstrap sample.

In the context of this study, the hyperparameters, specifically the number of trees ($n_{estimators}$) and the maximum depth of the trees (max_{depth}), were optimized using the Grid Search strategy to balance computational efficiency with predictive accuracy.

In the domain of flood forecasting, Random Forest is widely employed for its ability to generate high-resolution flood maps without the heavy computational cost of hydrodynamic simulations. It is particularly valued for its interpretability and its ability to handle "mixed" data types. Unlike MLR, which assumes a constant rate of rise, the tree-based structure naturally partitions the domain into "in-channel" and "overbank" regions.

Random Forest was selected as the third model to strictly address the Bias-Variance Tradeoff. While SVR introduced non-linearity, it can be sensitive to hyperparameter settings. Random Forest was chosen because:

- Stability: by averaging multiple trees, it reduces the risk of overfitting. As shown in the figure, even if Tree 3 makes an outlier prediction (e.g., the Orange node), the averaging with Tree 1 and Tree 2 (Blue and Green) smooths out the error.
- Non-Linearity: its hierarchical structure allows it to model complex, discontinuous hydraulic surfaces that smooth kernel methods might miss.

Despite its robustness, Random Forest possesses a critical mathematical limitation relevant to this study: inability to extrapolate. As noted by Hastie et al. (2009), decision tree-based models predict values by averaging the target variables in the leaf nodes of the training data. Consequently, the model cannot predict a water level higher than the maximum level observed in the training set. This is the primary driver for the creation of the specific Extrapolation Set (RP 450 and 500 years). In fact, it is hypothesized that RF will fail to accurately predict the peak water levels for these extreme return periods if the associated discharges exceed the training domain.

2.2.4 Extreme Gradient Boosting Model

Extreme Gradient Boosting (XGBoost) serves as the fourth predictive architecture implemented in this study. While the previously discussed Random Forest model employs "Bagging" (building independent, complex trees in parallel and averaging them to reduce variance), XGBoost utilizes the "Boosting" framework. In this approach, an ensemble of shallow weak learners (regression

trees) is built sequentially, with each new tree explicitly designed to correct the prediction errors made by the cumulative combination of previous trees (IBM Cloud Education, 2020a).

The fundamental mathematical principle driving XGBoost is Gradient Boosting, first formalized by Friedman (2001). The model does not try to predict the target variable H (Water Surface Elevation) directly in one step. Instead, it optimizes a fixed objective function through an iterative additive process (NHI, 2024).

As illustrated in Figure X below, the process begins with a base prediction (often the mean of the target). Subsequent trees do not predict the actual water level; they predict the residuals (the error term) of the preceding model.

Mathematically, if $\hat{H}_i^{(t)}$ is the prediction for the i-th instance at iteration t, the algorithm updates the prediction by adding a new function $f_t(x_i)$:

$$\hat{H}_i^{(t)} = \hat{H}_i^{(t-1)} + \eta \cdot f_t(x_i) \tag{7}$$

where:

- $\hat{H}_i^{(t-1)}$ is the cumulative prediction from the previous step (XGBoost Developers, 2024).
- $f_t(x_i)$ is the new regression tree learned at step t. Crucially, this tree is trained to minimize the loss function with respect to the current predictions, effectively pointing in the direction of the negative gradient of the loss (NHI, 2024).
- η (eta) is the Learning Rate (defined as `learning_rate` in the Python code, e.g., 0.1). It scales the contribution of each new tree. This "shrinkage" technique slows down the learning process, preventing the model from adapting too quickly to noise and significantly improving generalization (Hastie et al., 2009).

A primary rationale for moving from Random Forest to XGBoost in this study is to explore the other side of the bias-variance decomposition.

- Random Forest uses deep, fully grown trees. Individually, these trees have low bias (they fit their training subset well) but high variance (they are unstable and overfit). By averaging un-correlated trees, Bagging reduces the variance component of the error.
- XGBoost typically uses shallow trees (weak learners, controlled by the `max_depth` hyperparameter in the code). Individually, these trees have low variance but high bias (they underfit the complex hydraulic reality). By sequentially focusing on the "hardest" remaining errors, Boosting progressively reduces the bias component of the error (NVIDIA Corporation (2024)).

In the context of flood modeling, if the HEC-RAS calibration data contains significant noise or outliers, RF's variance reduction might be superior. However, if the data is clean and the

underlying hydraulic relationships are highly complex and subtle, XGBoost's capability to aggressively reduce bias often yields greater predictive accuracy.

In the field of hydro-informatics, XGBoost has rapidly become a preferred tool for real-time flood forecasting due to its winning combination of high accuracy on structured (tabular) data and computational efficiency. Recent benchmarking studies (e.g.,Ni et al. (2020)) have demonstrated its superiority over traditional methods like SVM and even standard neural networks in predicting streamflow and peak water levels, particularly in scenarios requiring complex nonlinear mapping.

However, despite its theoretical sophistication, XGBoost shares the fundamental topological limitation of all tree-based methods utilized in this study: Inability to Extrapolate. Like Random Forest, the final prediction is a weighted sum of values located in the terminal leaves of the trees. The model cannot predict a water surface elevation higher than the maximum value observed during training. Consequently, for the Extrapolation Set (RP 450 and 500), it is hypothesized that XGBoost predictions will "saturate" at the highest training value, failing to capture the true physical extent of these extreme, unseen events.

2.2.5 Gaussian Process Regression Model

Gaussian Process Regression (GPR) represents the fifth and most mathematically sophisticated architecture employed in this study, marking a decisive transition from the deterministic ensemble methods (Random Forest and XGBoost) to probabilistic Bayesian modeling. While the Random Forest model minimizes variance through averaging, it ultimately produces a single "best guess" point estimate. In contrast, GPR treats the flood prediction problem as a stochastic process, providing not only the predicted water surface elevation (H) but also a rigorous quantification of the predictive uncertainty (σ) for every point in the domain (Rasmussen and Williams, 2006).

Fundamentally, a Gaussian Process (GP) is defined as a collection of random variables, any finite number of which have a joint Gaussian distribution. In the context of this study, we assume that the function $f(x)$, which maps the hydraulic inputs (Discharge Q, Coordinates X, Y) to the output (H), is a sample from a GP distribution:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{8}$$

where:

- $m(x)$ is the mean function (assumed to be zero after standardizing the data).
- $k(x, x')$ is the Covariance Function (Kernel), which defines the similarity between input points.

The learning process in GPR is framed as Bayesian Inference, a statistical method that differs from traditional frequentist approaches by updating the probability of a hypothesis as more evidence becomes available (Rasmussen and Williams, 2006).

In facts, the model begins with a prior distribution over all possible functions. As the model "observes" the HEC-RAS training data (15 simulations), it conditions this prior to produce a posterior distribution. For a new test input x_*, the predictive distribution is Gaussian with mean $\bar{f}_*$ and variance $\mathbb{V}[f_*]$ given by:

$$\bar{f}_* = k_*^T (K + \sigma_n^2 I)^{-1} y \tag{9}$$

$$\mathbb{V}[f_*] = k(x_*, x_*) - k_*^T (K + \sigma_n^2 I)^{-1} k_* \tag{10}$$

In the provided Python implementation (`train_gpr_model_test3`), this probabilistic nature is explicitly leveraged. The function call `model.predict(..., return_std=True)` retrieves the posterior standard deviation (σ), which serves as the metric for model confidence.

The Choice of Kernel

A critical methodological decision in GPR is the selection of the kernel function, as it encodes the model's assumptions about the smoothness and differentiability of the underlying physical system. While the Radial Basis Function (RBF) (or Squared Exponential) is the default kernel in many machine learning applications, preliminary testing indicated it was unsatisfactory for this hydraulic case study. The RBF kernel is infinitely differentiable, implying that the modeled surface is perfectly smooth. However, hydraulic surfaces interact with complex topography, levees, and turbulent flow regimes, creating physical irregularities that are not perfectly smooth (Rasmussen and Williams, 2006).

To address this, the code implements the Matérn Kernel with the specific parameter $\nu = 1.5$. As established in the geo-statistical literature by Stein (1999), the Matérn class of kernels provides a generalization of the RBF. The parameter ν controls the smoothness of the resulting function. The value $\nu = 1.5$ chose corresponds to a function that is only once differentiable. This allows the Gaussian Process to model the "roughness" and sharper physical transitions typical of flood propagation over terrain, avoiding the unrealistic "oversmoothing" artifacts produced by the RBF kernel (Rasmussen and Williams, 2006).

In the domain of flood forecasting, the transition to GPR is driven by the need for Uncertainty Quantification. For an operational Early Warning System (EWS), a single predicted water level is insufficient if the confidence in that prediction varies significantly across the domain. GPR provides a spatial map of uncertainty (σ). This represents the model's epistemic uncertainty in emulating the hydraulic baseline, not the total physical uncertainty of the flood event. A high σ value near a critical embankment signals to decision-makers that the model is extrapolating or lacks sufficient data in that region, necessitating a more conservative safety margin (Wang

et al., 2025).

GPR was selected as the fifth model to overcome the inherent structural limitations of the previously tested tree-based architectures:

- Probabilistic Output: neither Random Forest nor XGBoost can provide a mathematically rigorous confidence interval for its predictions. GPR naturally outputs a probability distribution, essential for risk-based decision making (Rasmussen and Williams, 2006).
- Smoothness and Continuity: tree-based models, being a collection of decision trees, produce predictions that are effectively piece-wise constant (step functions). GPR produces a continuous, smooth predictive surface that is physically more consistent with the behavior of fluid dynamics (Rasmussen and Williams, 2006).

However there are some limitations, for instance the primary barrier to the widespread adoption of GPR is its computational complexity. The training process requires the inversion of the covariance matrix K, an operation with a time complexity of $O(N^3)$, where N is the number of training points (Rasmussen and Williams, 2006). Also, while the Random Forest model could easily ingest tens of thousands of data points, the GPR model's memory and processing requirements grow cubically. This mathematical constraint necessitated the specific data processing strategy observed in the code (`split_data_by_groups`), where the training dataset was rigorously limited to 15 representative simulations to ensure the matrix inversion remained computationally feasible.

2.2.6 Hyperparameter Optimization Strategy

A critical distinction in machine learning design is the difference between *model parameters* and *hyperparameters*. Model parameters, such as the regression coefficients (β) in linear models or the weights in a neural network, are internal variables learned automatically from the data during the training phase. In contrast, hyperparameters are external configuration variables that control the learning process itself (e.g., the learning rate, the maximum depth of a decision tree, or the number of estimators in an ensemble). These cannot be estimated directly from the training data and must be defined *a priori* (Goodfellow et al., 2016). The selection of these hyperparameters is pivotal because they directly govern the model's complexity and its position on the Bias-Variance Tradeoff curve.

- Underfitting (High Bias): If the hyperparameters constrain the model too strictly (e.g., a tree depth limited to 2), the model fails to capture the underlying physical trends of the hydraulic data.
- Overfitting (High Variance): If the model is too complex (e.g., an unconstrained tree depth), it effectively "memorizes" the training noise rather than learning the general hydraulic laws, leading to poor performance on the Extrapolation set.

To address this, a grid search optimization strategy was implemented for all non-linear models. This method performs an exhaustive search over a manually specified subset of the hyperparameter space. For every combination of parameters in the grid, the model's performance is evaluated using k-fold Cross-Validation (with $k = 3$ or $k = 5$).In k-fold Cross-Validation, the training set is randomly partitioned into k equal subsamples (folds). The model is trained on $k-1$ folds and validated on the remaining fold. This process is repeated k times, with each fold serving as the validation set exactly once. The results are then averaged to produce a single performance metric (e.g., Mean Squared Error). This technique ensures that the selected hyperparameters are robust and not biased by a specific random split of the training data (Kohavi et al., 1995). The specific hyperparameters tuned for each architecture, a brief description of them and the values choose for the tuning are reported in the table below.

Model	Parameter	Description	Tested Values
SVR	C	Regularization parameter controlling the training error penalty.	1, 10, 100
	Epsilon	Width of the tolerance margin (tube) where no error penalty is applied.	0.01, 0.1
	gamma	Kernel coefficient defining the influence of a single training point.	function “scale”: automatically calculates the optimal numerical value for a specific dataset
Random Forest	n_estimators	The total number of decision trees built in the forest ensemble. More trees increase stability but cost more computation.	50, 100, 200
	max_depth	The maximum number of levels (splits) allowed in each individual tree. It controls model complexity and prevents overfitting.	10, 20
GPR	kernel	The mathematical function defining spatial correlation. A Matern kernel with nu=1.5 is standard for modeling physical/hydraulic surfaces as it allows for realistic roughness.	nu=1.5 (fixed)
	Kernel Bounds	The search space provided to the internal optimizer to find the perfect variance (C) and distance correlation (length_scale).	C: 10^(-3), 10^(3) length_scale: 10^(-2), 10^(2)
XGBoost	n_estimators	The number of boosting rounds (the number of sequential trees built).	100, 200
	max_depth	The maximum depth of each individual tree. Lower values restrict complexity to prevent overfitting on specific coordinates.	6, 10
	learning_rate	The step-size shrinkage. It scales down the contribution of each new tree, forcing the model to learn slowly and robustly rather than memorizing the data instantly.	0.05, 0.1
	subsample	The fraction of the training data randomly sampled to grow each tree. Using 0.8 (80%) introduces randomness to prevent overfitting.	0.8 (fixed)

Table 1: Summary Table of the Parameters tested for each of the Model chosen

2.3 Evaluation metrics

To comprehensively evaluate the predictive performance and physical plausibility of the implemented surrogate models, a dual-faceted validation framework was adopted. This approach combines quantitative statistical metrics, to assess global accuracy and overall error distribution, with a spatially explicit comparative analysis, ensuring that the models reliably replicate localized flood dynamics without masking critical spatial errors.

2.3.1 Statistical Measures

To rigorously assess the predictive performance of the five implemented models, a suite of four distinct statistical metrics was employed. These indicators were selected not just to measure global accuracy, but to diagnose specific model behaviors, such as sensitivity to outliers and systematic bias. The mathematical formulations and the rationale for their selection are detailed below.

- **Root Mean Square Error** (RMSE) served as the primary criterion for model selection and optimization in this study. It quantifies the square root of the average squared differences between the predicted water levels ($\{y_{pred}$) and the actual values simulated by HEC-RAS (y_{true}).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{true} - y_{pred})^2} \tag{11}$$

  Where N represents the total number of evaluation points. As argued by Chai and Draxler (2014), RMSE is particularly appropriate for hydrological modeling because the squaring mechanism disproportionately penalizes large errors. In the context of flood risk, this property is desirable: a model that misses a peak water level by 1 meter (a large error) is significantly more dangerous than a model that is consistently off by a few centimeters, and RMSE reflects this severity more effectively than linear metrics.

- **Mean Absolute Error** (MAE) provides a linear assessment of the average error magnitude, treating all deviations equally regardless of their direction.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_{true} - y_{pred}| \tag{12}$$

  While RMSE is sensitive to outliers, MAE provides a more stable view of the "typical" error one might expect from the model. According to Willmott and Matsuura (2005), comparing MAE and RMSE allows for a diagnostic interpretation of the error distribution: if the RMSE is significantly larger than the MAE, it indicates the presence of large, sporadic outliers in the prediction, whereas similar values suggest a consistent error spread.

- **Coefficient of Determination** (R^2) describes the proportion of variance in the observed

water levels that is explained by the model's inputs (Discharge and Coordinates).

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_{true} - y_{pred})^2}{\sum_{i=1}^{N}(y_{true} - \bar{y})^2} \tag{13}$$

Where $\bar{y}$ denotes the mean of the observed data. An R^2 of 1.0 implies a perfect fit. However, as noted by Legates and McCabe Jr (1999), relying solely on R^2 in hydrological goodness-of-fit tests can be misleading, as it is overly sensitive to extreme values (like flood peaks) and insensitive to additive bias. Therefore, in this study, R^2 is treated as a secondary metric to confirm general correlation rather than absolute accuracy.

- **Mean Bias Error** (MBE) was implemented (via the `calculate_mbe function`) to detect systematic drift in the predictions.

$$MBE = \frac{1}{N}\sum_{i=1}^{N}(y_{pred} - y_{true}) \tag{14}$$

Unlike the previous metrics which use absolute values or squares, MBE preserves the sign of the error. A positive value signals a tendency to overestimate water levels, while a negative value indicates underestimation.

2.3.2 Spatial Comparative Analysis

In this study, to effectively evaluate the spatial accuracy of ML models model against the physically-based HEC-RAS benchmarks, a specific visualization workflow was developed. Rather than relying solely on global error metrics such as the ones described above, it was decided to implement a spatially explicit comparison to identify local patterns of overestimation or underestimation.

The process adopted for generating the flood maps relies on the Triangulated Irregular Network (TIN) approach, specifically utilizing Delaunay Triangulation. This method was selected to preserve the exact nodal values of the hydraulic mesh without introducing smoothing artifacts.

The hydraulic data provided by HEC-RAS consists of unstructured scatter points (x, y) with associated flow depths (H). To visualize this as a continuous field, the study utilized Delaunay Triangulation, a fundamental geometric algorithm that connects a set of discrete points into a mesh of non-overlapping triangles (Environmental Systems Research Institute (ESRI), 2023). Mathematically, this method maximizes the minimum angle of all the angles of the triangles in the triangulation, thereby avoiding "sliver" triangles that can distort spatial representation (Lee and Schachter, 1980).

By constructing a TIN, the visualization performs linear barycentric interpolation within each triangle. This means that for any point p inside a triangle with vertices A, B, C, the value

h_p is calculated as a weighted average of the vertices:

$$h_p = \lambda_1 h_A + \lambda_2 h_B + \lambda_3 h_C \tag{15}$$

where λ_i are the barycentric coordinates. This approach was deemed most effective for this study because it respects the original hydraulic calculations of the 2D mesh, ensuring that the visualized surface is a direct linear representation of the node values (Peucker et al., 1978).

A significant challenge in visualizing irregular river channels using Delaunay Triangulation is the "Convex Hull" property. By definition, standard Delaunay algorithms connect the outermost points of the dataset, which often results in spurious triangles, artificial connections that span across non-flooded areas (e.g., connecting two distant riverbanks across a dry meander). To rectify this, a geometric constraint was implemented in the code. In this study, a heuristic edge-pruning algorithm was applied. This calculates the Euclidean distance d_{ij} for every edge of the generated triangles (Edelsbrunner et al., 1983):

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{16}$$

A dynamic threshold was established based on the local statistical properties of the mesh. Specifically, a cutoff limit was defined as:

$$L_{max} = \mu_L \times 2.5 \tag{17}$$

where μ_L is the mean edge length of the local triangulation. Any triangle containing an edge exceeding L_{max} was masked (removed) from the visualization. This procedure ensured that the flood maps accurately respected the complex, non-convex boundaries of the river channel without artificially "flooding" dry zones.

The visual comparison of two similar gradient maps, one representing the HEC-RAS WSE values for a given discharge and the second one representing the prediction of the model for the same given discharge, is often subject to perceptual limitations especially if the errors are minimal. To provide a rigorous validation, this study introduced a Spatial Difference Map (or Residual Map) as the third component of the comparison panel. To generate the visual Spatial Difference Map (and not as a training optimization objective), the local residual error ϵ at every coordinate (x, y) is computed as:

$$\epsilon(x, y) = H_{ML}(x, y) - H_{Hec-ras}(x, y) \tag{18}$$

A diverging colormap (seismic) was applied to this residual field. This visualization technique allows for the immediate identification of:

- Overestimation (Positive Residuals): Areas where the Machine Learning model predicts higher water levels than the physical model.

- Underestimation (Negative Residuals): Areas where the model fails to capture the peak depth.

By centering the color map at zero (white), the applied method ensures that the visual emphasis is placed solely on significant deviations, facilitating a rapid diagnostic of the models' spatial performance (Willmott, 1981). This comparative analysis between the models predictions and the HEC-RAS benchmark was deliberately restricted to the Test and Extrapolation datasets (Training Set was excluded). This decision is grounded in the fundamental principle of machine learning validation known as Generalization Capability. In facts, the Training Set consists of the data the model has already "seen" and used to adjust its internal parameters. If we were to plot the comparisons for the Training discharges, we would likely see near-perfect results. Whereas, the test set represents the model's ability to interpolate, so by focusing on these discharges, the study demonstrates that the models has learned, or not, the underlying physical laws of the river system rather than just memorizing coordinates (Hastie et al., 2009).

The Extrapolation Set is the most rigorous test of the model's robustness. In flood risk management, the most critical events are often extreme, low-probability floods that may exceed historical records or available simulations. By isolating and analyzing these high-magnitude events, this study assesses whether the mathematical functions learned by the models remain stable outside their training domain (Hastie et al., 2009).

2.4 Post-Processing and Flood Inundation Mapping

While the primary output of the machine learning models consists of discrete water surface elevation (H) predictions at specific coordinate points, effective flood risk assessment requires the spatial visualization of these results over the continuous topography.

To achieve this, a dedicated post-processing algorithm was developed in Python to transform the tabular prediction datasets into inundation maps. This procedure integrates the model outputs with the original LiDAR-derived Digital Terrain Model (DTM) to calculate spatially distributed water depths.

The mapping workflow, implemented using the `Rasterio` (Gillies et al., 2013) and `SciPy` (Virtanen et al., 2020) libraries, proceeds through three distinct analytical stages:

1. Spatial Interpolation: The input dataset consists of sparse coordinate points (x, y) with associated water levels (H). To generate a continuous water surface from these discrete points, the study employed a Linear N-Dimensional Interpolation method, presented previously in paragraph 2.3.2. This technique first constructs a Delaunay triangulation of the input points and then performs barycentric linear interpolation within each triangle (Barber et al., 1996). This approach ensures that the resulting Water Surface Elevation (WSE) grid remains mathematically consistent with the predicted values while smoothly transitioning across the domain.

2. Depth Calculation and Masking: The interpolated WSE grid was aligned with the high-resolution LiDAR DTM (2-meter resolution). The flood depth (D) for each pixel (i, j) was calculated using the subtraction method:

$$D_{i,j} = WSE_{i,j} - Z_{DTM_{i,j}} \tag{19}$$

where Z_{DTM} represents the ground elevation. To define the inundation extent accurately, a binary mask was applied to exclude non-flooded areas (where $D \leq 0$) and regions with missing terrain data.

3. Geospatial Visualization: To provide context for the flood extents, the resulting depth maps were superimposed onto satellite imagery. This was achieved using the `contextily` library (Barber et al., 1996), which retrieves and integrates georeferenced map tiles (Esri World Imagery) directly into the plotting environment. To ensure sufficient clarity for detailed analysis, the satellite tiles were retrieved at zoom levels ranging from 16 to 18. This final step allows for the visual verification of flood limits against recognizable land use features and infrastructure.

By automating this workflow, this study successfully generated georeferenced flood maps for the entire 98 km reach of the Po River, enabling a direct visual comparison between the machine learning predictions and the hydraulic modeling benchmarks.

To rigorously assess the model's generalization capability, a specialized Automated Inundation Mapping Function (`generate_inundation_maps`) was embedded directly within the machine learning inference pipeline. Unlike standard post-processing, this tool operates on the independent Test and Extrapolation datasets immediately after training. It utilizes the trained model to generate fresh water level predictions for specific discharge events (grouped by Q) and automatically renders high-resolution inundation maps within a user-defined focus window (typically a 3000×3000 meter bounding box). This workflow allows for an immediate, visual validation of the model's predictive performance on unseen data, facilitating a direct comparison between the Machine Learning spatial outputs and the expected hydraulic behavior (HEC-RAS simulation) in complex topographic regions.

3 Case Study

The Po River is recognized as both the longest and one of the most geographically and economically significant rivers in Italy. It originates in the Piedmont Region at the Monviso spring, situated at an elevation of 2,020 meters above sea level. It is characterized by a mixed pluvio-nival hydrological regime, maintaining an overall mean annual discharge of approximately $1,500\ m^3/s$ at its closing section, though this masks significant seasonal volatility (Montanari, 2012).

Figure 5: Po River Position and Po River Basin Maps

The annual cycle exhibits two high-flow and two low-flow periods. Up to 90% of the annual surface flow volume is concentrated between September and May (Vezzani, 2014). During the spring (driven by Alpine snow melt) and autumn (driven by intense rainfall) high-flow seasons, daily peak discharges regularly exceed the annual mean, occasionally resulting in extreme flood peaks ranging from $8,000$ to $10,300\ m^3/s$ (Zanchettin et al., 2008). Conversely, the river experiences a secondary low-flow period in winter, when Alpine precipitation is temporarily stored as snow and ice, and a primary, critical low-flow period in summer. Driven by high evaporation and scarce rainfall, summer discharges at Pontelagoscuro, little municipality at the

beginning of the closing section, historically average below 600 m^3/s. However, recent severe droughts (e.g., 2017 and 2022) have depressed absolute daily minimums to dangerously low levels between 168 and 275 m^3/s (Montanari, 2012).

The main channel of the river has a total length of approximately 650 km passing through 4 regions, Piemonte, Lombardia, Emilia Romagna and Veneto, and it is fed by a network of 141 tributaries, as it is shown in Fig 5. This river system underpins approximately 40% of the national agricultural production and significant industrial activity, making the basin an area of paramount national strategic interest (Autorità di Bacino Distrettuale del Fiume Po, 2021a). The historical development of this region has necessitated extensive hydraulic engineering, including the construction of a massive, kilometers-long embankment system, to safeguard urban and agricultural areas from flooding.

The Po River basin is classified as an area of High Hydraulic Risk (Autorità di Bacino Distrettuale del Fiume Po (2021a); Lastoria and et al. (2021)). The perennial threat of flooding is amplified by two key main factors: natural factors, that include significant annual variations of the river's natural regime, with spring floods fed by snow melt and autumn floods driven by intense Mediterranean rainfall, and anthropogenic factors caused by the intensive development within the historical floodplains generating the phenomenon known as the "levee effect". As embankments were heightened to contain larger floods, development was encouraged behind them. Consequently, potential flood depths and associated damage have dramatically increased in the event of an embankment breach (Amadio, 2013). The Flood Risk Management Plan (FRMP/PGRA) dictates the methodology for assessing the magnitude and frequency of these critical events.

Indeed, the traditional hydrological regime of the Po is being fundamentally altered by anthropogenic climate change, leading to a dual challenge of both increased drought and intensified flood hazard (Coppola and Giorgi, 2010). While models show a projected decline in average annual river flow, particularly during summer due to decreased precipitation and earlier snow melt (Autorità di Bacino Distrettuale del Fiume Po (2021a); Cozzi and Giani (2011)), scientific consensus indicates a significant increase in the frequency and intensity of extreme events:

- Intensified Precipitation: climate models predict that the Mediterranean area will experience more frequent and more intense high-magnitude rainfall events (Cozzi and Giani (2011); Amadio (2013)). This shifting pattern means that annual discharge maxima are projected to increase by up to 50% in some regions of the basin, and the frequency of exceeding extreme discharge thresholds may increase by over 500% for some rivers by the end of the century (Pregnolato et al., 2022).
- The Drought-Flood Cycle: The Po is increasingly subject to erratic fluctuations between

prolonged periods of severe drought (such as the unprecedented 2022 event) and sudden, catastrophic flood peaks (Tarolli and et al., 2023). These rapid shifts severely stress the river's ecological balance and drastically increase the risk of hydraulic failure due to quick saturation and desiccation of the earthen embankments.

The Po River's significant economic, social, and ecological value, coupled with the classification of its basin as a high-flood-risk area, makes it a compelling subject for this research. However, the sheer complexity of modeling the entire river course is notable, given the extensive length of the fluvial path and the considerable morphological diversity of its drainage basin (Autorità di Bacino Distrettuale del Fiume Po, 2021a). For this reason, the scope of the study has been delimited to the specific segment of the Po River traversing the area from Cremona to Borgoforte.

3.1 Study Area

The study area, represented in Fig. 6 stretching from Cremona to Borgoforte (Province of Mantova), is situated in a particularly flat critical section sector of the Middle Po valley and it is characterized by a stable main channel with a medium value of 300 m of width (Bradimarte and Di Baldassarre, 2012). The hydraulic safety here depends entirely on the integrity of the main embankment system, whose width varies from 400 m to 4 km, which often protects densely populated or highly productive agricultural zones (Autorità di Bacino Distrettuale del Fiume Po, 2021b).

The area is characterized by extremely low water-surface gradient, approximately 0.17 ‰ (per thousand) (Agenzia Interregionale per il fiume PO (AIPO), 2021), which confirms how crucial is this section for flood events. The ADBPO classifies this reach (approximately from km 463 at Cremona to just upstream of the Oglio confluence) as the transition between a braided and sinuous channel section, that mainly characterized the reach under study as it can be seen in Fig. 6, and the final single-thread sinuous reach that extends to the delta head (Autorità di Bacino Distrettuale del Fiume Po, 2009).

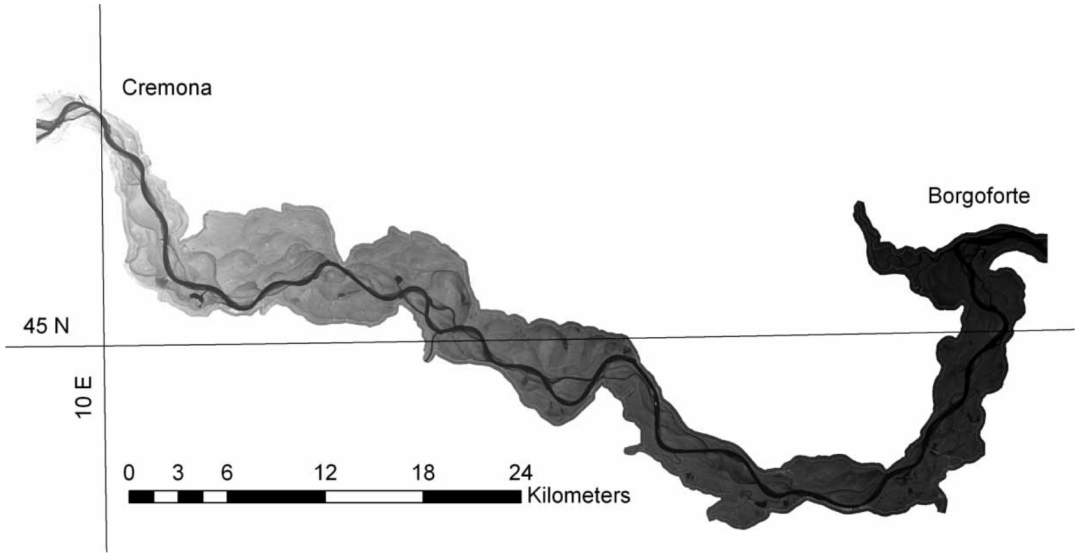


Figure 6: Study Area Cremona-Borgoforte (Bradimarte and Di Baldassarre, 2012)

In Fig. 7 it can be noticed that the area surrounding the Po River in the reach under analysis predominantly falls under the Agricultural class and, more specifically, consists of arable land/non-irrigated arable land (Legend Code 2.1.1) and Permanent Crops. Paved areas fall under the Artificial Surfaces class and are specifically categorized as continuous/discontinuous urban fabric (Legend Code 1.1.2, 1.2.1) (Ministero dell'Ambiente e della Sicurezza Energetica (2025); Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA) (2018)). This percentage is low compared to the vast agricultural area and is concentrated within the municipalities (Cremona, Casalmaggiore, etc.). To obtain the exact percentage of paved ground (e.g., $10 - 20\%$), it is necessary to extract the data from a GIS (Geographic Information System) using the official Corine Land Cover (CLC) datasets (Ministero dell'Ambiente e della Sicurezza Energetica, 2025).

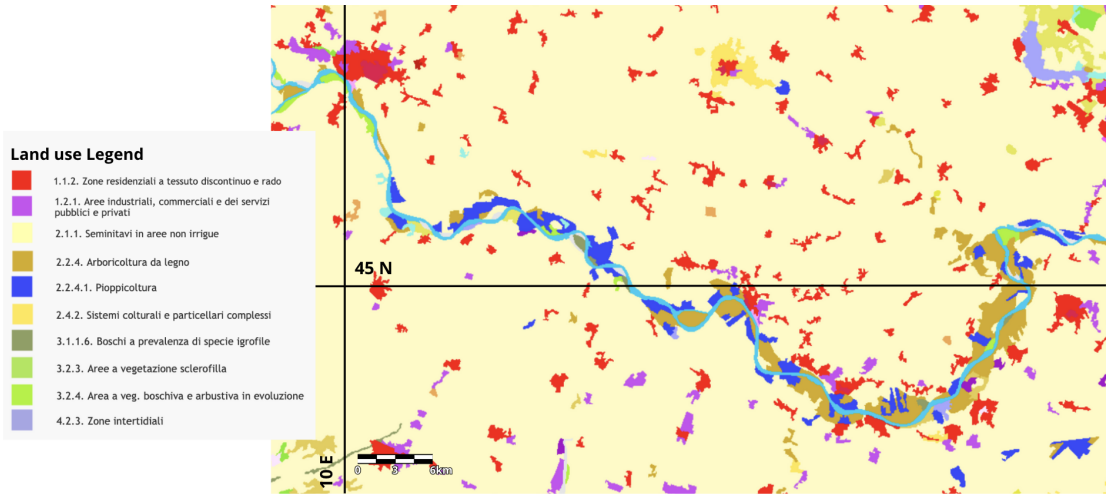


Figure 7: Land Use in the study area (Ministero dell'Ambiente e della Sicurezza Energetica, 2025)

The local Provincial Emergency Plans (e.g.,Provincia di Cremona (2012); Provincia di Mantova (2020) specifically detail the vast extent of inundation that would result from a simulated embankment breach mapping the risk areas in this reach (Fig 8) and highlighting the importance of the few remaining golenal areas (floodplains) for temporary flood storage. The maintenance and hydraulic resistance of these defenses are subject to the heightened stress imposed by the extreme hydrological variability driven by climate change (Pregnolato et al., 2022).

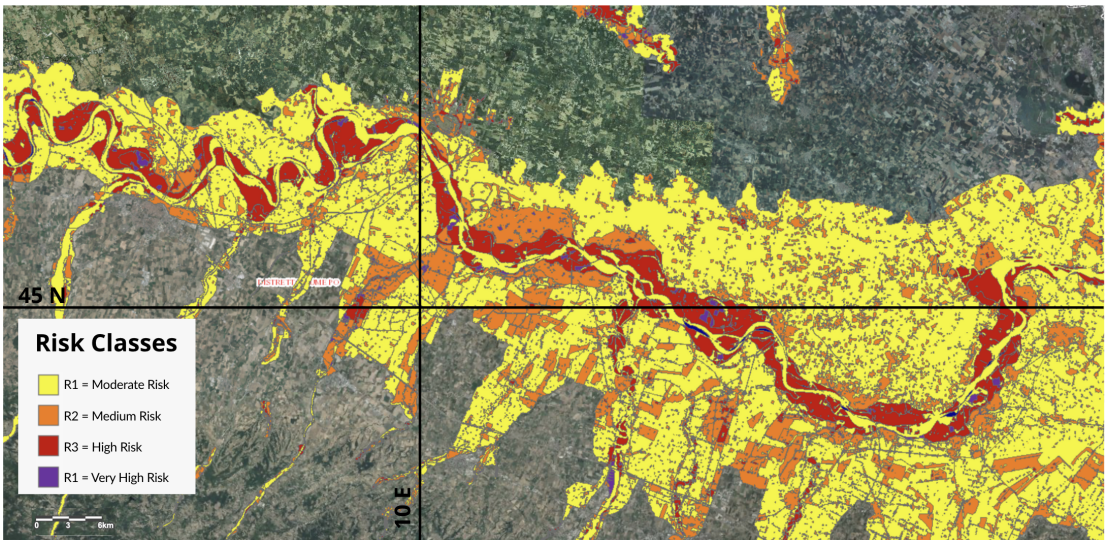


Figure 8: Flooding Risk Map of Cremona-Borgoforte Areas (Ministero dell'Ambiente e della Sicurezza Energetica, 2025)

3.2 Input Data processing

Following a thorough analysis of pivotal studies, such as the research by Wang et al. (2025) and Qureshi et al. (2025) on the Ottawa River, that shaped the research questions of this study, it is necessary to introduce the hydrological and topographical inputs. Summarized in Table 2, these are the starting elements for flood modelling.

	Input Parameter	**Description / Value**	**Source / Notes**
Topographical Data	Digital Elevation Model (DEM)	2m resolution LiDAR image	Brandimarte Documentation
	River Reach Length	98 km (Cremona to Borgoforte)	Brandimarte Paper
	River HEC-RAS Geometry	Point by Point River Reach Description through 68 Sections	Imported from Brandimarte Study document in HEC-RAS Mapper
	Roughness Coefficient (Manning's n) [s*m^(-1/3)]	0.035 (main channel), 0.11 (floodplains)	Calibrated based on historical event
Hydrological Data	Upstream Boundary Condition	Discharge Value (Q [m^3/s])	Variable (20 Discharge Values statistically calculated)
	Downstream Boundary Condition	Normal Depth (Borgoforte) Friction slope: 0.00012	Calculated based on the altimetry of the beginning and end section point
	Simulation Conditions	Steady flow	Base flow conditions

Table 2: Summary Table of the Topographical and Hydrological Input Data

3.2.1 DTM and DTM derived data

A preliminary qualitative analysis of the altimetric data was performed by examining Digital Terrain Models (DTMs) extracted from the Po River Basin Authority portal. High-resolution 2-meter DTMs of the middle-lower portion of the Po River were constructed from data collected in 2005 using two distinct LiDAR (Light Detection and Ranging) laser scanners (3033 Optech ALTM and Toposys Falcon II), flown at altitudes of approximately $1,500$ m (Bradimarte and Di Baldassarre, 2012).

While LiDAR provides highly accurate ground elevation data for the subaerial topography, the complete reconstruction of the terrain elevation required specific measurements for the riverbed (bathymetry). LiDAR systems cannot effectively measure below the water surface due to the physics of light propagation in water: the laser pulses used by standard airborne topographic LiDAR (typically near-infrared) are strongly absorbed and quickly scattered upon hitting the water surface. The change in the refractive index between air and water causes the laser beam to bend (refract), preventing an accurate measurement of the underlying bed

elevation. Furthermore, the signal rapidly attenuates over short distances, rendering the return signal too weak for reliable depth calculation. For this reason, specific complementary measurements were essential. The channel bathymetry for the navigable portion was characterized by a boat survey conducted in the same year using a multi-beam sonar (Kongsberg EM 3000D) (Catsellarin, 2011).

The resultant DTM was, at the end, processed to remove vegetation and validated (Catsellarin (2009), Bradimarte and Di Baldassarre (2012)).

Channel bathymetry was essential for the precise characterization of the reach's cross-sections, which are necessary to establish the geometric data for the 1D model within HEC-RAS. In this study, 68 cross-sections are taken into consideration, corresponding to 68 river stations in the model. Each cross-section is geometrically defined by its total length, the main channel length, the left and right overbank lengths, as well as critical hydraulic parameters like the Manning's coefficient (n) and the expansion/contraction coefficients. A properly spatially referenced Digital Terrain Model (DTM) is fundamental for providing the altimetric inputs to the 1D model. Although HEC-RAS calculates the flood profile in one dimension (along the channel), the model is capable of outputting georeferenced inundation boundaries and associated geographical coordinates. These specific results, detailing the extent and depth of the flooding, will serve as a primary input for the subsequent Machine Learning (ML) algorithm training phase.

As established in Paragraph 2, the downstream boundary condition for the steady-state 1D hydraulic model (HEC-RAS) is defined using the Normal Depth method. This method relies on setting the energy gradient equal to the estimated mean slope of the river reach under analysis. This approach is conventional when detailed stage-discharge relationships or known water surface elevations are unavailable at the downstream control point, such as Borgoforte.

The mean slope (S_0) is methodologically computed from the available topographic data to ensure a robust input for the model, specifically relying on a combination of the Digital Terrain Model (DTM) and the surveyed cross-sections (Bradimarte and Di Baldassarre, 2012). The calculation begins with the Path Length Calculation, where the total length (L) of the flow path is determined by summing the progressive distances (ΔL_i) measured between consecutive cross-sections (River Stations) along the channel centerline, extending from the most upstream section (Cremona) to the most downstream section (Borgoforte). Next, the Elevation Differential (ΔZ) is established by taking the elevation value (Z_{up}) of the lowest point (thalweg or channel invert) at the most upstream cross-section of the area of interest and subtracting the elevation value (Z_{down}) of the corresponding lowest point at the most downstream cross-section. Finally, the Mean Slope Computation is performed, where the mean bed slope (S_0) is mathematically computed as the ratio of the total elevation drop (ΔZ) over the total path length (L):

$$S_0 = \frac{Z_{up} - Z_{down}}{L}$$

This calculated mean slope, which accurately reflects the average bed gradient of the Po River between Cremona and Borgoforte, is then directly inputted into the HEC-RAS model to define the Normal Depth boundary condition. This crucial step allows the model to effectively simulate the uniform flow condition that ultimately dictates the water surface elevation at the downstream limit.

3.2.2 Discharge Data

Discharge data constitute the essential empirical basis for this study. The primary objective of this analysis is to estimate the flood discharge values (Q) associated with specific return periods (T), a critical component for hydraulic design and flood risk assessment. To achieve this, we employ a probabilistic approach, detailed in Paragraph 2, which models the frequency and magnitude of extreme events based on historical records.

The dataset utilized for the flood frequency analysis consists of the annual maximum daily discharges recorded over the period spanning from 2002 to 2016. This sample, comprising $N = 15$ years of observations, was deliberately adopted from the foundational dataset used in Brandimarte's previous study. In statistical hydrology, to achieve robust estimates and reduce uncertainty when extrapolating to rare events with high return periods (e.g., $T = 100$ or $T = 200$ years), engineering practice and international guidelines typically advise the use of historical series extending over at least 30 years (Stedinger and W. Griffis, 2008).

While consciously acknowledging this statistical limitation, the decision to not append more recent years was strictly driven by the need for methodological continuity. The objective of this preliminary phase is to compute design discharges that can be directly validated against the existing literature for this river reach. Introducing new data points would shift the parameters of the probability distribution, fundamentally altering the resulting design flows and breaking the alignment with Brandimarte's established benchmark. Therefore, the analysis proceeds with this exact 15-year window to ensure that the calculated discharges provide a consistent and literature-validated hydrological foundation for the subsequent hydraulic simulations and machine learning training.

The estimation of discharges for different return periods was performed in MATLAB. The analysis utilized the annual maximum daily discharge values from 2002 to 2016 provided by ISPRA Ambiente (Istituto Superiore per la Protezione e la Ricerca Ambientale) which is a public research body in Italy, overseen by the Ministry of the Environment and Energy Security, focusing on the protection and monitoring of the environment across the entire national territory.

which were imported from an Excel spreadsheet. The initial statistical analysis consisted of plotting the time series and the non-exceedance frequency function. This was followed by a distribution analysis, which included calculating the mean, variance, standard deviation, and derived parameters. These statistics were then used to estimate the parameters for the Gumbel distribution using both the method of moments and the method of L-moments.

The Gumbel distribution was the first model tested, as it is commonly used in hydrological studies (e.g., Brandimarte, 2012). A visual comparison of the plotted data against the fitted distribution (Fig. 9) suggested a good initial fit. To quantitatively assess this, formal goodness-of-fit tests were employed. First, the Pearson Chi-Squared test was applied using a 5% significance level. Recognizing that the Chi-Squared test can be less powerful due to the loss of information from binning data into classes, the Anderson-Darling test was also computed as a more robust alternative. Since the Gumbel distribution (with parameters estimated by both methods) passed both tests, it was selected for calculating the quantile function.

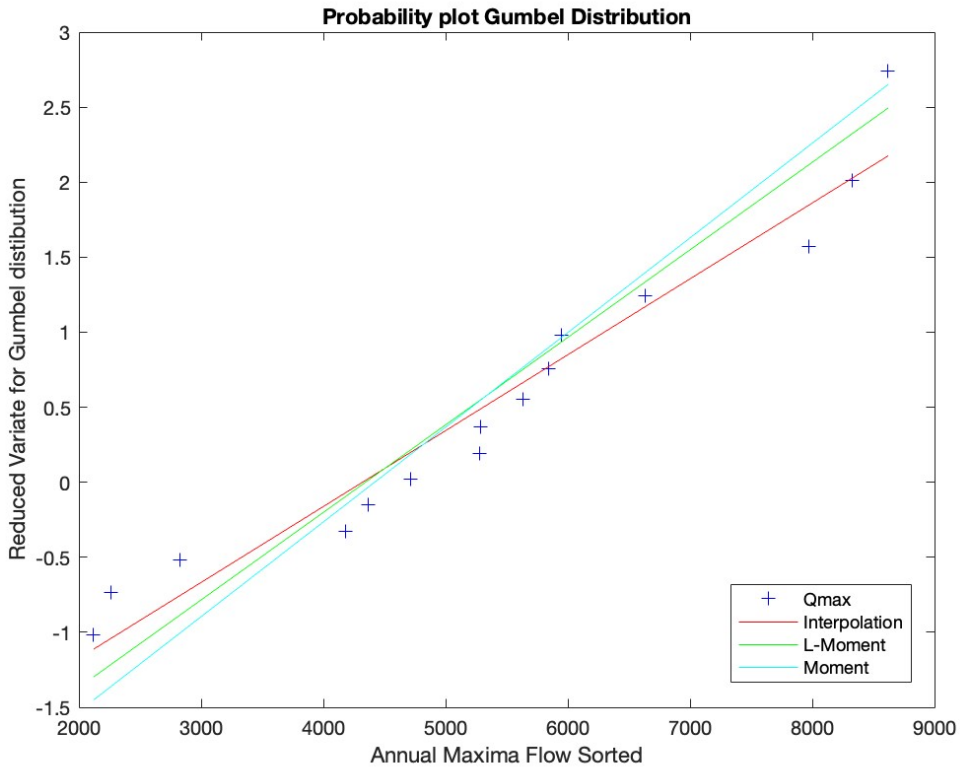


Figure 9: Probability plot for the Gumbel Distribution. The x-axis represents the sorted annual maximum daily discharges (Q) in m^3/s, while the y-axis represents the corresponding Gumbel reduced variate, which is a linear transformation of the non-exceedance probability.

Considering 20 different return periods, the probability associated to them was computed by using the formula:

$$P(x_T) = 1 - \frac{1}{T}$$

Then the design flow was computed by determining the values x_T corresponding to $P(x_T)$ with quantile extraction.

The probabilistic analysis yielded a discharge value for the 200-year return period ($T = 200$) estimated at approximately $Q_{200} \approx 13,400\,\mathrm{m}^3/\mathrm{s}$. This result is deemed acceptable and consistent as it shows agreement with previous studies conducted for the same area. Specifically, the calculated value is closely comparable to the Q_{200} estimate of $13,700\,\mathrm{m}^3/\mathrm{s}$ reported in the study by Brandimarte (2012).The proximity between these two values ($13,400\,\mathrm{m}^3/\mathrm{s}$ versus $13,700\,\mathrm{m}^3/\mathrm{s}$) is particularly significant and enhances the reliability of the estimates. This confidence is further supported by the fact that both studies utilized the same probability distribution (Gumbel distribution) for the frequency analysis, thereby confirming the methodological robustness and the validity of the selected statistical model.

The values of the probable discharges obtained are summarized in Table 3.

Return Period [years]	Q_T $[m^3/s]$
2	4,970
5	6,920
10	8,200
20	9,440
30	10,150
50	11,030
75	11,740
100	12,230
125	12,620
150	12,930
175	13,200
200	13,420
225	13,630
250	13,800
275	14,000
300	14,120
350	14,380
400	14,600
450	14,820
500	15,000

Table 3: Gumbel Distribution Designed Discharge Values

3.2.3 Calibration Data

Thanks to a careful and meticulous analysis of the event's consequences carried out by Autorità di Bacino del Fiume Po, it was possible to estimate the peak discharge of the disastrous October 2000 flood event along the Po River. The estimated peak discharge in the Cremona area was approximately $11,850$ m^3/s (Bradimarte and Di Baldassarre, 2012). In addition to this figure, the event's impact was documented through measurements of the maximum water levels reached in the flooded areas (Autorità di Bacino del Fiume Po, 2004). These measurements enabled the creation of a dataset suitable for use in the model calibration, described in the previous paragraph.

The goal of the model calibration process is to find the values of the Manning's coefficients for the floodplain and the main channel that best fit the observed data, in order to generate the future models based on the most accurate river reach's characteristics obtainable. It's important to mention that there are also the expantion and contraction coefficients. These were not calculated in this study, but they were chosen based on the research of Brandimarte (2012) that not only estimated the two parameters but, also, analyzed how these two values have a minor impact over the water surface elevation results. The process of calibration explained in the previuos paragraph was based on two main ranges of values, according to the Brandimarte (2012)´s study and Chow (1959)'s book:

- Floodplain Manning's coefficient range values: $0.06 - 0.14$, changed with a foot of 0.01
- Main channel Manning´s coefficient range values: $0.02 - 0.06$, changed with a foot of 0.005

The fitting of these values was represented by a contour plot, reported in Fig. 10 , conducted in MATLAB, where the Error (MAE) response surface for the main channel Manning's coefficient value can be observed. The minor value of the MAE for the floodplain is found equal to $0.11\,s \cdot m^{-1/3}$, and for the main channel is equal to $0.035\,s \cdot m^{-1/3}$

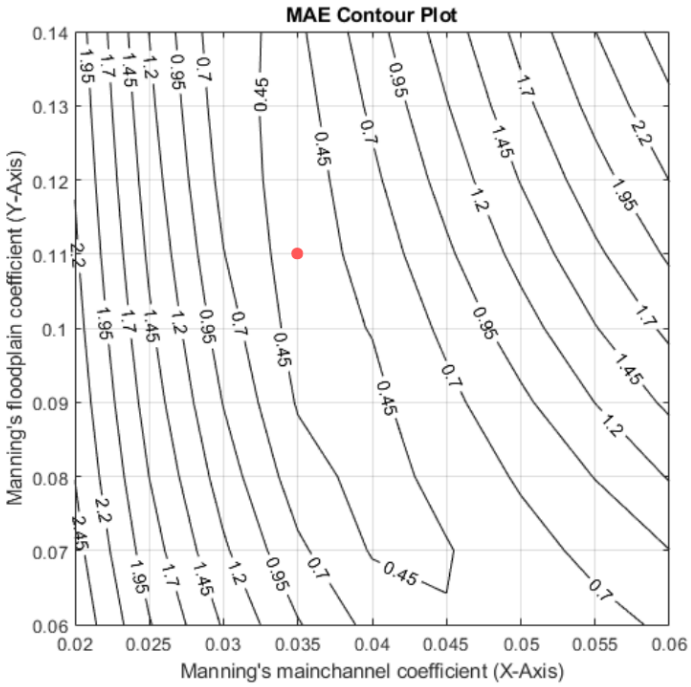


Figure 10: Contour plot of the Mean Absolute Error (MAE) for hydraulic model calibration. The plot illustrates the sensitivity of the model's error to various combinations of Manning's roughness coefficients for the main channel (x-axis) and the floodplain (y-axis). The contour lines represent intervals of constant MAE, facilitating the identification of the optimal parameter set that minimizes simulation error (red point).

Since the geometry used in the Brandimarte (2012)'s study was slightly different from the one the it's used in this study, the results are different. In particular Brandimarte (2012) considered 88 section, whereas in this current study the geometry was simplified with 68 cross section of the same river reach.

With the geometric inputs defined and the Manning's roughness coefficients rigorously calibrated against historical flood marks, the modelling framework for the Po River case study is fully established. These validated hydrological and topographical boundary conditions constitute the physical basis for the HEC-RAS simulations. In the following chapter, this calibrated setup is utilized to generate the high-fidelity 'ground truth' dataset required to train the Machine Learning surrogates, and the resulting predictive performance of the proposed architectures is systematically analyzed.

4 Results and Analysis

4.1 Hydraulic Modeling Results

In this study, the hydraulic baseline was established using the HEC-RAS modeling system. As detailed in the previous sections, the model was rigorously calibrated against a historical flood event to ensure the accuracy of the Water Surface Elevation (WSE) predictions. With the calibration parameters fixed, the focus of the analysis shifts to the geometric representation of the river reach and the generation of the simulations training data required for the machine learning algorithms.

4.1.1 Geometric Representation and Model Setup

The geometric fidelity of the hydraulic model is fundamental to the accuracy of the inundation extent. Figure 11 illustrates the geometric schematization adopted for the 98 km reach of the Po River. The model integrates high-resolution cross-sectional data with the LiDAR-derived digital terrain model, effectively capturing both the main channel bathymetry and the complex topography of the floodplains.

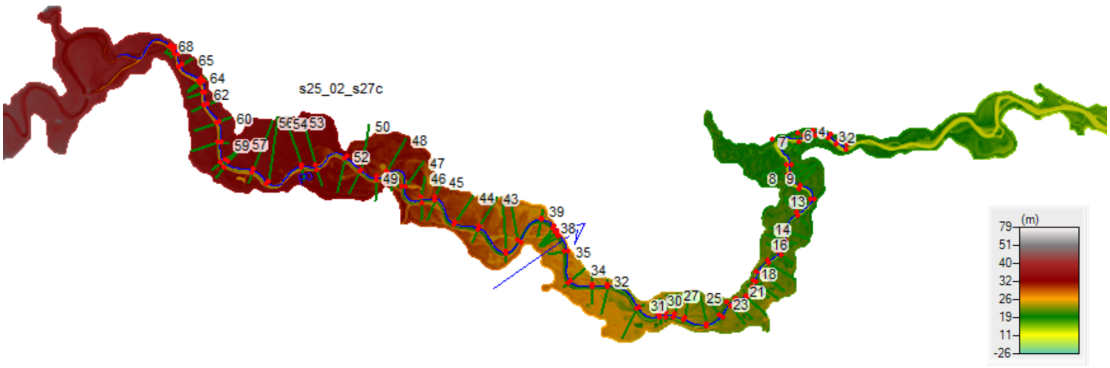


Figure 11: Po River study section Geometry on HEC-RAS

This geometric setup allowed for the simulation of complex flow patterns, particularly in the meander bends and confluence zones, which are critical for accurate flood risk assessment (Brunner, 2016).

4.1.2 Hydraulic Mapping and Output Generation

Following the geometric setup, steady flow simulations were performed for a range of discharge scenarios (Q) to generate comprehensive flood maps. The HEC-RAS Mapper module was uti-

lized to visualize the results, producing spatially distributed Water Surface Elevation (WSE) layers.

Figure 12 presents an example of these outputs, displaying the WSE overlaid directly onto the Digital Elevation Model (DEM). These maps provide a continuous surface of water over the sea level, allowing for the identification of possible inundated areas where the computed WSE could exceed the ground elevation. These hydraulic maps serve as the "ground truth" against which the machine learning predictions are visually and statistically compared.

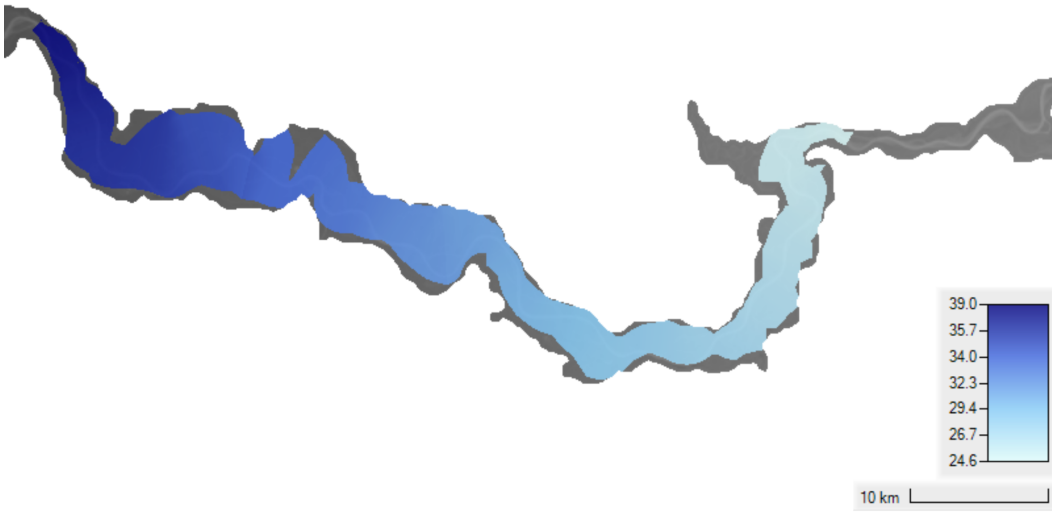


Figure 12: Example of HEC-RAS output showing Water Surface Elevation (WSE) overlaid on the Digital Elevation Model (DEM).

4.1.3 Data Export and Processing

To enable the training of the machine learning models, the spatial results from the hydraulic simulations required structured extraction. It was chosen to export the HEC-RAS results in a standardized GIS format. Specifically, the WSE values at discrete profile points were exported as geospatial text files directly from the HEC-RAS interface.

These raw export files were subsequently processed and consolidated into a master dataset. The data was organized into an Excel file named `Simulation_Results`, containing tabular records of coordinates (X, Y) and their corresponding hydraulic heads (H) for each discharge scenario. This structured dataset formed the foundational input for the training, testing, and validation of the machine learning algorithms described in Section 2 and sumarized in the Table 4 below.

Set	Return Periods (T in years)	Description
Training	2, 5, 20, 30, 50, 75, 125, 150, 175, 200, 225, 250, 275, 350, 400	Training the machine learning algorithms to learn the underlying hydraulic relationships.
Test	10, 100, 300	Evaluating the models interpolation capabilities within the boundaries of the training data.
Extrapolation	450, 500	Assessing model robustness when predicting extreme events outside the training domain.

Table 4: Division of the simulated flood scenarios into Training, Test, and Extrapolation sets based on their Return Periods.

4.2 Machine Learning Results

In this section, the predictive performance of the five implemented Machine Learning architectures is analyzed and compared. The evaluation is conducted hierarchically, first examining the models' ability to learn the training data, then assessing their generalization capability on unseen interpolation data (Testing Set), and finally stress-testing their robustness on extreme, out-of-domain events (Extrapolation Set).

4.2.1 Training Results

The first step in the analysis focuses on the Training Set, which consists of 15 steady-flow simulations used to optimize the model parameters. Analyzing the error distributions and scatter plots for this subset provides insight into the "learning capacity" of each algorithm.

In terms of statistical accuracy, a clear hierarchy emerged among the tested architectures. Quantitative analysis of the performance metrics, specifically the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2), reveals that the tree-based ensemble methods, Random Forest (Fig. 13)and Extreme Gradient Boosting (Fig. 14), achieved the superior results. In this study, both models demonstrated near-perfect goodness-of-fit scores ($R^2 \approx 0.99 - 1.0$) on the training data, accompanied by negligible RMSE values.

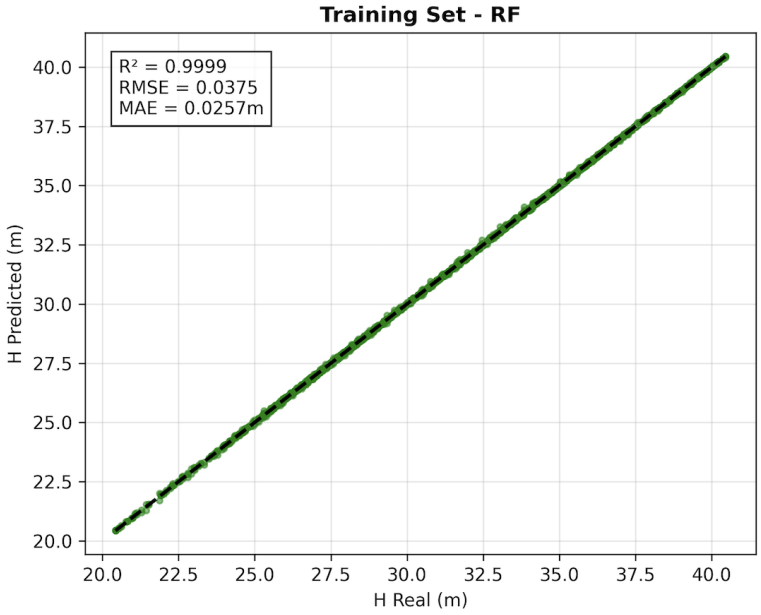


Figure 13: Random Forest Training Set Scatter Plot and Evaluation Metrics results

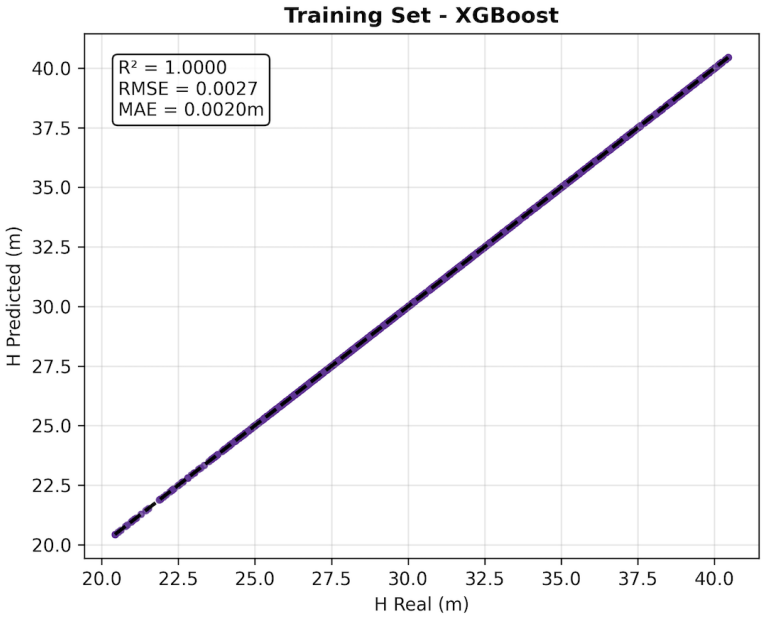


Figure 14: XGBoost Training Set Scatter Plot and Evaluation Metrics results

This statistical dominance is attributable to their non-parametric nature: unlike linear models that are constrained by fixed geometric assumptions, these ensemble methods utilize decision trees to recursively partition the feature space (Hastie et al., 2009). In the context of flood mod-

eling, this allows them to isolate and model highly localized hydraulic features, such as sudden changes in bathymetry or embankment edges, thereby minimizing the residual variance to near-zero levels.

As illustrated in the error distribution histograms (e.g., Figure 15 and 16), a clear distinction emerges between the parametric and non-parametric models. The MLR model displays a broad, bell-shaped error distribution even on the training data Figure 15.

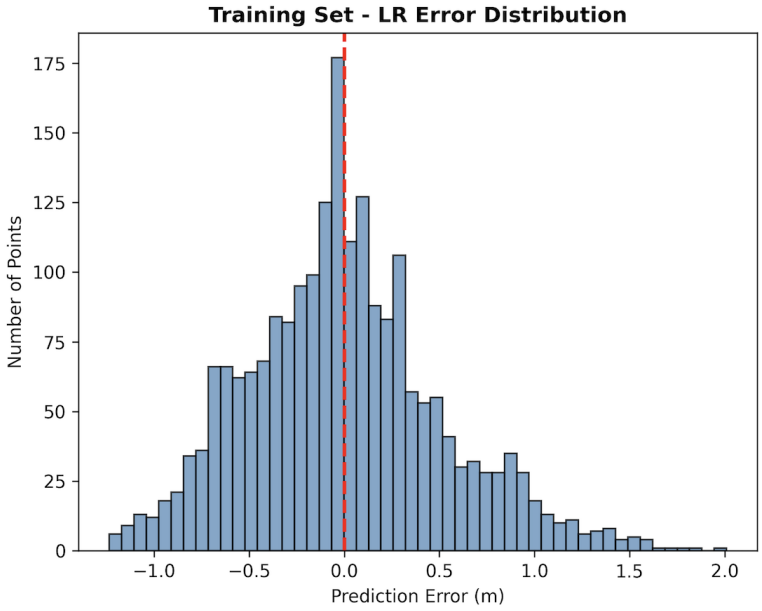


Figure 15: Error Distribution for Multiple Linear Regression model training set

The presence of a broad, bell-shaped error distribution within the Training Set is a textbook indicator of underfitting (or high bias) (Hastie et al., 2009). In machine learning, underfitting occurs when the model architecture is too simple to capture the underlying structure of the data (Bishop, 2006). For the Multiple Linear Regression (MLR) model, this limitation is structural: the algorithm attempts to fit a flat hyperplane through a dataset that is intrinsically non-linear. Hydraulically, the relationship between Discharge (Q) and Water Surface Elevation (H) is governed by complex power laws (e.g., Manning's Equation) and typically exhibits an "S-curve" or distinct breakpoints as water transitions from the main channel to the floodplains. The linear model, constrained by a constant slope (β), cannot bend to accommodate these transitions. Consequently, the model is forced to make a "global compromise": in some regimes (e.g., low flows) it overestimates water levels, whereas in others it underestimates them (e.g., high flows) (Bishop, 2006).

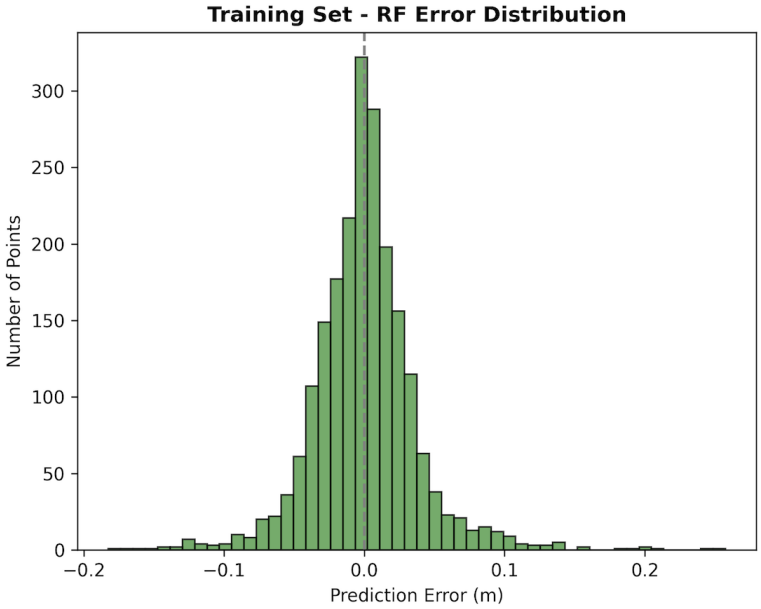


Figure 16: Error Distribution for Random Forest model training set

When these systematic miscalculations are aggregated into a histogram, they form a wide Gaussian-like distribution. Unlike the Random Forest model (Fig. 16), where the error distribution spikes sharply at zero (indicating the model has successfully adapted to local variations), the MLR's bell curve represents a high Residual Sum of Squares (RSS). The width of this bell curve is the standard deviation and it effectively quantifies the "unexplained variance" such as the physical information regarding channel geometry and flow resistance that the linear model failed to capture and instead treated as random noise (Hastie et al., 2009).

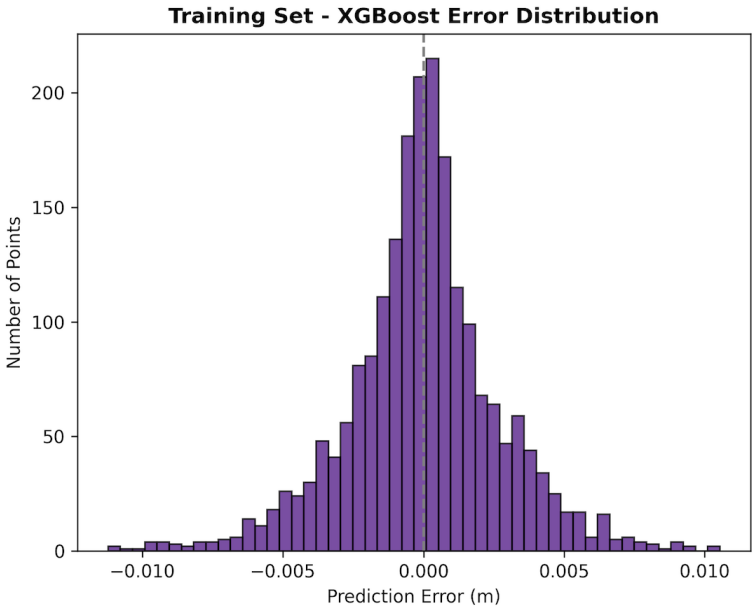


Figure 17: Error Distribution for XGBoost model training set

In contrast, the tree-based ensemble methods, particularly the Random Forest (RF) (Fig. 16) and XGBoost (Fig. 17), exhibit a sharply peaked error distribution centered exactly at zero when evaluated on the Training Set. This behavior is characteristic of low-bias algorithms capable of mapping highly non-linear boundaries (Bishop, 2006).

Specifically, regarding the Random Forest training error histogram in Figure 16, the distribution is characterized by an exceptionally tall and narrow peak. This morphology indicates that the model has effectively "memorized" the training examples, achieving near-perfect reproduction of the provided HEC-RAS water surface elevations. While this demonstrates the algorithm's mathematical power to adapt to the complex, localized topographies of the Po River floodplains, such an extreme fit on the training data inherently raises the concern of "overfitting" (Breiman, 2001). In this scenario, the deep decision trees may be capturing the specific, distinctive artifacts of the 15 training simulations (the numerical "noise" of the grid) rather than learning the general physical laws of fluid dynamics, a risk that must be evaluated in the subsequent testing phase.

4.2.2 Testing Results

Following the assessment of the models learning capacity on the training data, the analysis proceeds to the Testing Set. This dataset comprises three specific return periods ($T = 10, 100, 300$ years) that were deliberately excluded from the training phase. Conceptually, this set evaluates the models' ability to interpolate: the algorithms must predict hydraulic conditions for discharge magnitudes they have never explicitly "seen", but which fall within the range of the training

examples (Hastie et al., 2009). While high performance on the Training Set confirms that a model can reproduce known data, high performance on the Testing Set confirms that it has learned the underlying physical relationship between discharge (Q) and water surface elevation (H) (Hastie et al., 2009).

The quantitative metrics for the Testing Set reveal a natural, yet significant, divergence from the near-perfect scores observed in the training phase. In this study, while the tree-based models (Random Forest and XGBoost) maintained their superiority over the linear baseline, their error metrics increased. For instance, the Root Mean Square Error (RMSE) for the Random Forest model (Figure 18 that include the three discharge simulation outputs) rose from negligible values in training to a measurable error in testing, passing from a value of 3,75 centimeters to a value of almost 41 centimeters. This increase is expected and represents the generalization error, the gap between the model's fit to the training data and its accuracy on new, unseen inputs (Hastie et al., 2009).

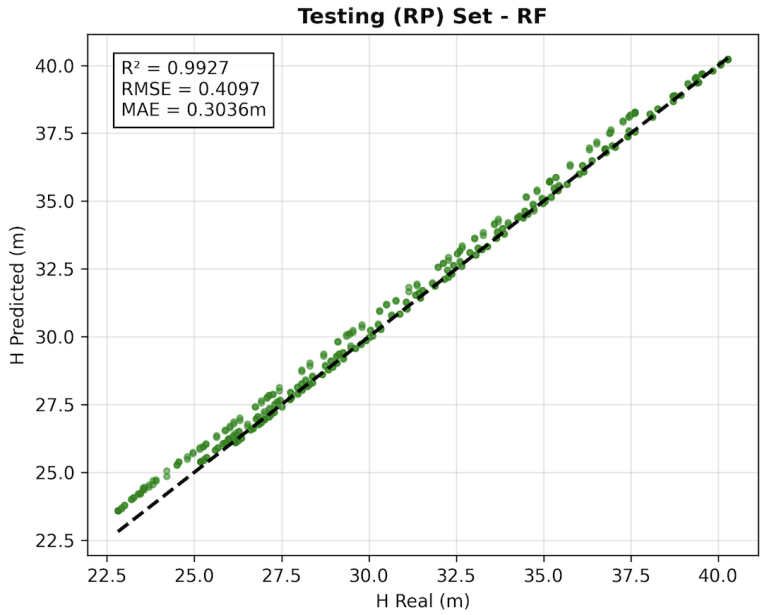


Figure 18: Random Forest Testing Set Scatter Plot and Metrics Results

Despite this increase, the tree-based ensembles continued to outperform the Multiple Linear Regression (MLR) model (Figure 20). The MLR's performance remained consistently poor across both sets, confirming that its high error is due to high bias (underfitting) rather than variance. In contrast, the Gaussian Process Regression model (Figure 19) demonstrated a robust performance stability, with error metrics on the Test Set remaining comparable to those of the Training Set, suggesting a strong capability for smooth interpolation without overfitting (Rasmussen and Williams, 2006).

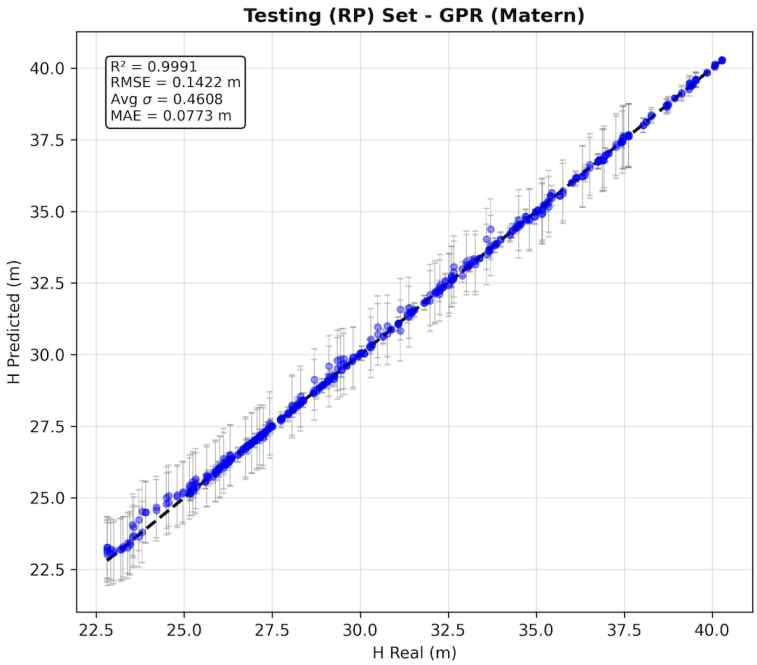


Figure 19: Gaussian Process Regression Testing Set Scatter Plot and Metrics Results

A closer visual inspection of the scatter plots for both the Random Forest (Figure 18) and the Extreme Gradient Boosting (XGBoost) (Figure 21) reveals a distinct morphological characteristic in the error structure. While the coefficient of determination (R^2) remains high, with a value of 0.9886, the cloud of predicted points exhibits a noticeable vertical translation (upward or downward shift) with respect to the perfect 1:1 diagonal line.

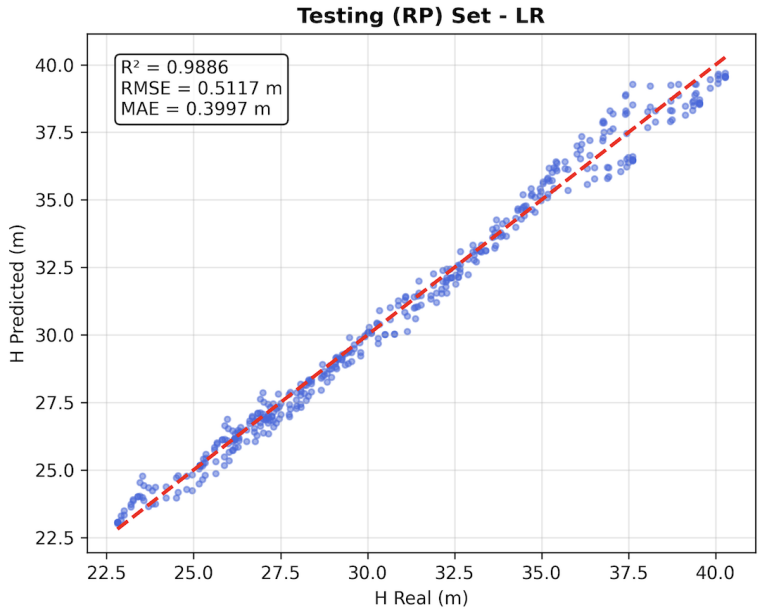


Figure 20: Multiple Linear Regression Testing Set Scatter Plot and Metrics Results

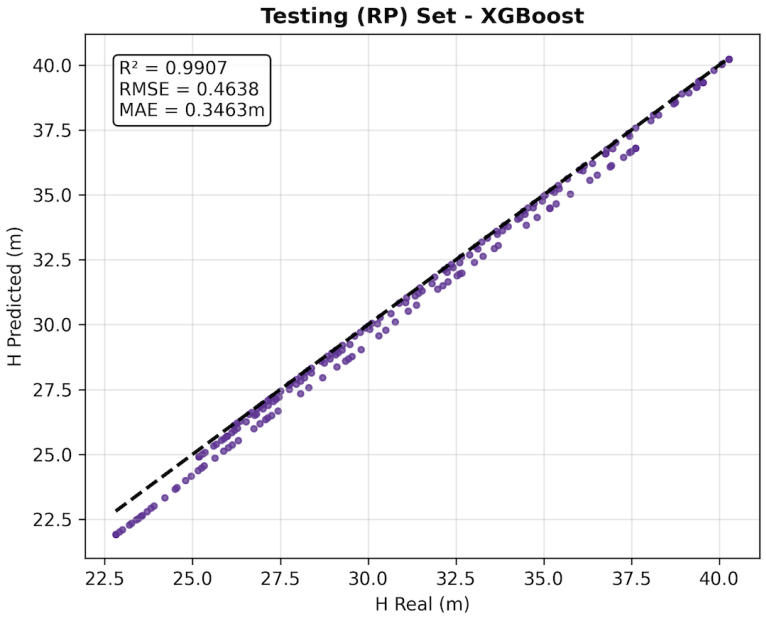


Figure 21: XGBoost Testing Set Scatter Plot and Metrics Results

This geometric translation is indicative of a systematic Bias Error ($MBE \neq 0$), as we can see in Table 5. Unlike variance, which manifests as a random symmetrical scatter around the trend line, this shift implies that both models are consistently overestimating or underestimating the

water surface elevation for specific discharge groups.

This behavior is intrinsic to the mathematical structure of decision tree ensembles. As described by Hastie et al. (2009), tree-based methods approximate continuous functions using piece-wise constant regions (or "steps"). They recursively partition the feature space into hyper-rectangles and assign a constant prediction value to each leaf node. Consequently, while these models capture the spatial variance (the shape of the river) effectively, their "stepped" response to the continuous discharge variable introduces these localized, discharge-specific bias shifts.

The analysis of the error distribution histograms provides further insight into the models' behavior. As observed in the Random Forest and XGBoost error plots (Figure 22 and Figure 23), the histograms display a multi-modal distribution, effectively showing three distinct "bells" or clusters of error.

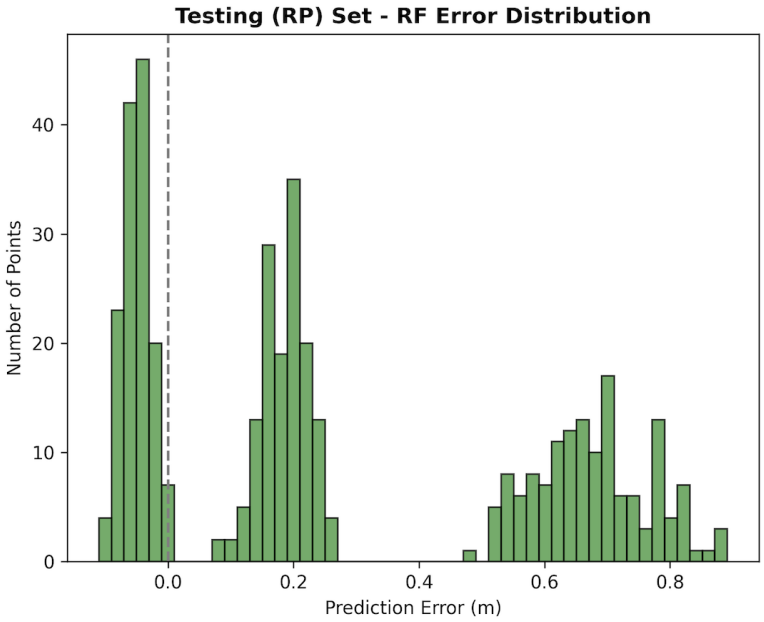


Figure 22: RF Testing Set Error Distribution Histogram

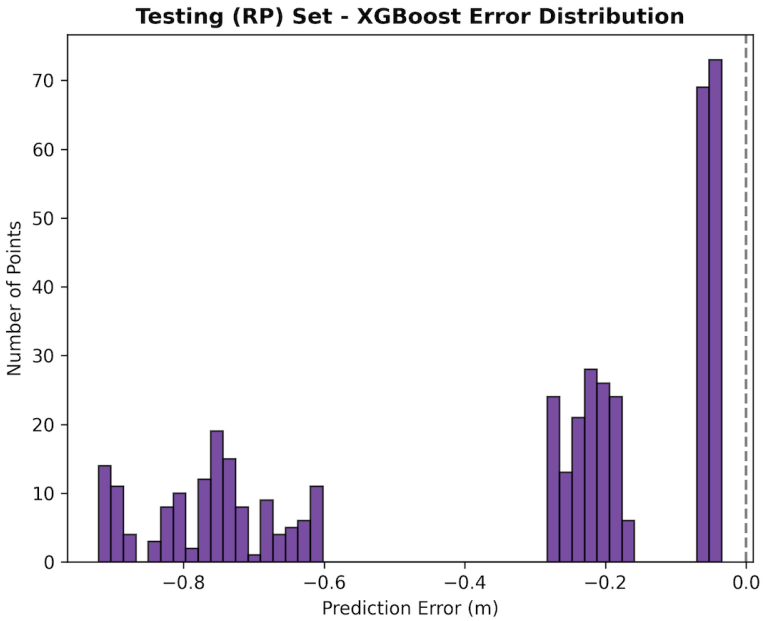


Figure 23: XGBoost Testing Set Error Distribution Histogram

This phenomenon is a direct consequence of the Testing Set's composition, which consists of three discrete discharge events ($T = 10, 100,$ and 300 years). Due to the algorithmic structure of Decision Trees, the models do not produce a continuous error spread across the entire domain. Instead, they exhibit a regime-specific bias:

- Discharge-Dependent Shifts: For each specific discharge scenario (e.g., Q_{10}), both models predict a water surface based on discrete split points learned during training. In Random Forest, this is an average of leaf nodes; in XGBoost, it is a sum of step-function updates. If the resulting value is slightly offset from the true hydraulic value for that specific Q, the entire flood map for that event will carry a systematic "vertical shift" (as seen in the scatter plots) (Hastie et al., 2009).

- Superposition of Distributions: Consequently, the "error bell" for the $T = 10$ event might center at $+0.05$ m, while the $T = 300$ event might center at -0.08 m. When these scenario-specific errors are aggregated into a single histogram, they do not merge into a single zero-mean distribution but rather appear as separate overlapping clusters (Hastie et al., 2009).

This visual evidence reinforces the conclusion that tree-based models, while highly accurate, do not learn a truly continuous function $f(Q)$. Instead, they learn a series of localized approximations, each carrying its own specific bias depending on how the test discharge falls relative to the training split-points (Hastie et al., 2009).

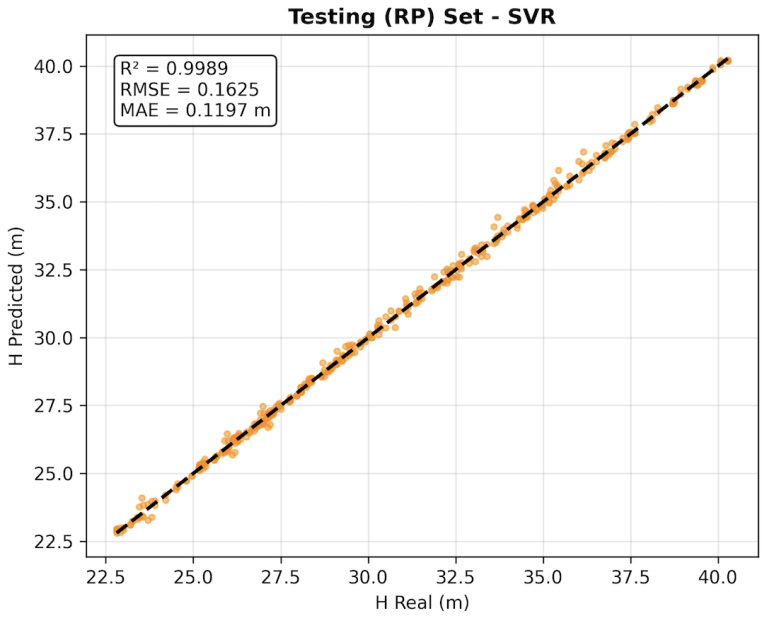


Figure 24: SVR Testing Set Scatter Plot and Metrics Results

Finally, the SVR presents a morphological behavior that contrasts sharply with the tree-based ensembles. As observed in the SVR scatter plot shown in Figure 24, the predictions do not exhibit the discrete "staircase" grouping typical of the Random Forest. Instead, the cloud of points follows a smoother, more continuous distribution along the 1:1 diagonal.

This is corroborated by the error histogram (Figure 25), which does not display the distinct multi-modal "three-bell" shape observed in the tree models. Because SVR relies on the Radial Basis Function (RBF) kernel, it projects the data into a continuous high-dimensional manifold, allowing for a mathematically smooth interpolation between discharge regimes without the abrupt segmentation of decision boundaries (Smola and Schölkopf, 2004).

However, while this smoothness results in a more physically consistent error distribution during interpolation, it hints at a fundamental structural limitation. The RBF kernel depends on the proximity to "support vectors" (training points) to generate predictions. As the input data moves away from these known anchors, the kernel's influence decays, potentially causing the model to revert to the mean (Smola and Schölkopf, 2004). This specific mathematical characteristic serves as the critical pivot to the final phase of this study: the evaluation of the Extrapolation Set, where the models are tested against extreme discharges that lie strictly outside the support domain of the training data.

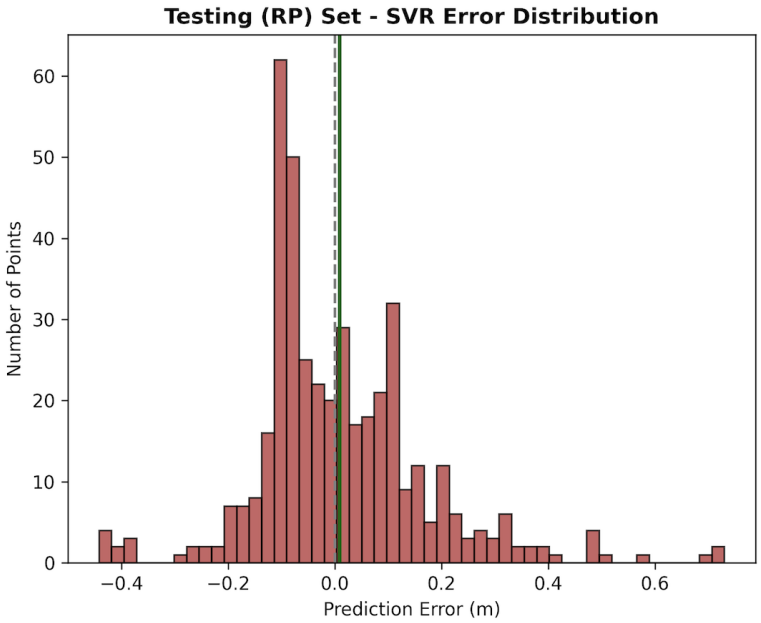


Figure 25: SVR Testing Set Error Distribution Histogram

4.2.3 Extrapolation Results

The final and most rigorous phase of the evaluation focuses on the Extrapolation Set, corresponding to the extreme return periods of $T = 450$ and $T = 500$ years. In this regime, the models are tasked with predicting flood levels for discharge magnitudes ($Q > 14,800\, m^3/s$) that exceed the maximum values observed during the training phase. This test assesses not just accuracy, but the fundamental robustness of the mathematical architectures when pushed beyond their empirical validity domain.

The results for the tree-based ensembles, RF and XGBoost, reveal a critical structural vulnerability. As visualized in the scatter plots (e.g., Figure 26 and Figure 27), the predictions for the Extrapolation Set do not follow the 1:1 diagonal trend line. Instead, they exhibit a distinct "saturation" effect, where the predicted water levels plateau at a constant value regardless of the increasing discharge.

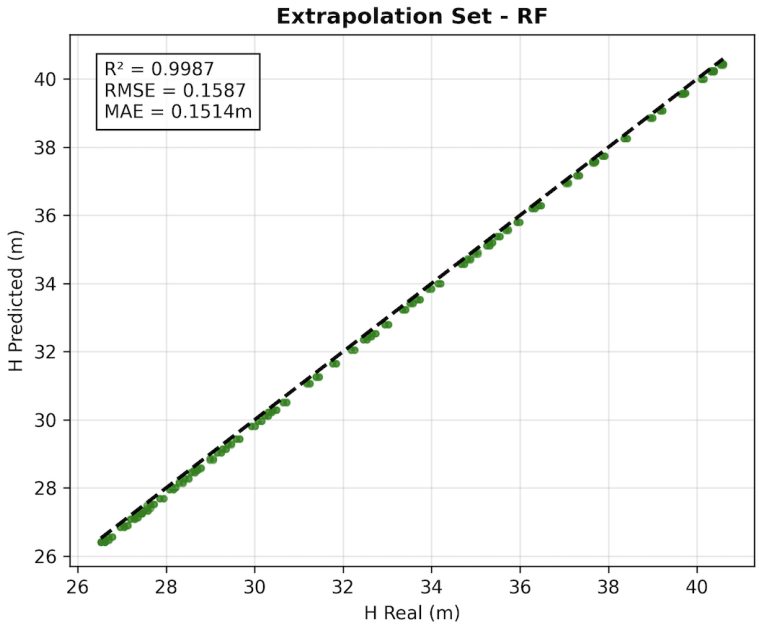


Figure 26: Random Forest Extrapolation Set Scatter Plot

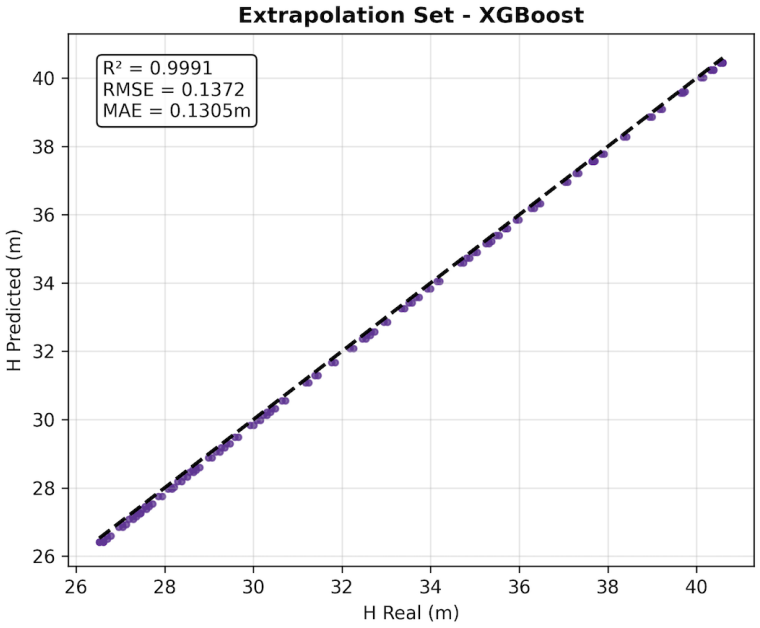


Figure 27: XGBoost Extrapolation Set Scatter Plot

This behavior is mathematically intrinsic to decision tree algorithms. Since trees predict the average value of the leaf node reached by the input, they are incapable of outputting a value higher than the maximum target (H_{max}) present in the training set (Hastie et al., 2009). Because

of this limitation, both models systematically underestimate flood depths during these extreme events. As shown in Figure 28 and in Figure 29, the error distribution histograms confirm this analysis, showing a massive, systematic shift of the extrapolation "bell" towards the negative axis (underestimation). While these models were the superior interpolators, this result confirms they are unsuitable for predicting unprecedented extremes without physical constraints.

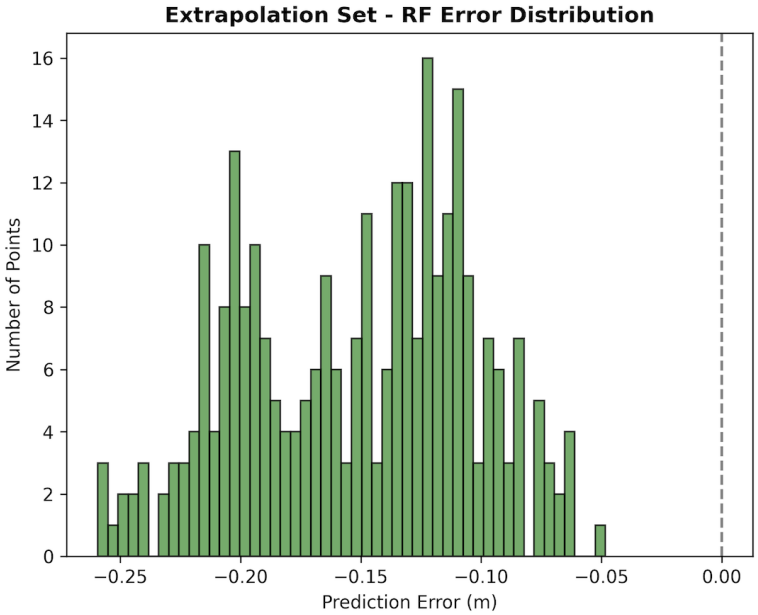


Figure 28: RF Extrapolation Set Error Distribution Histogram

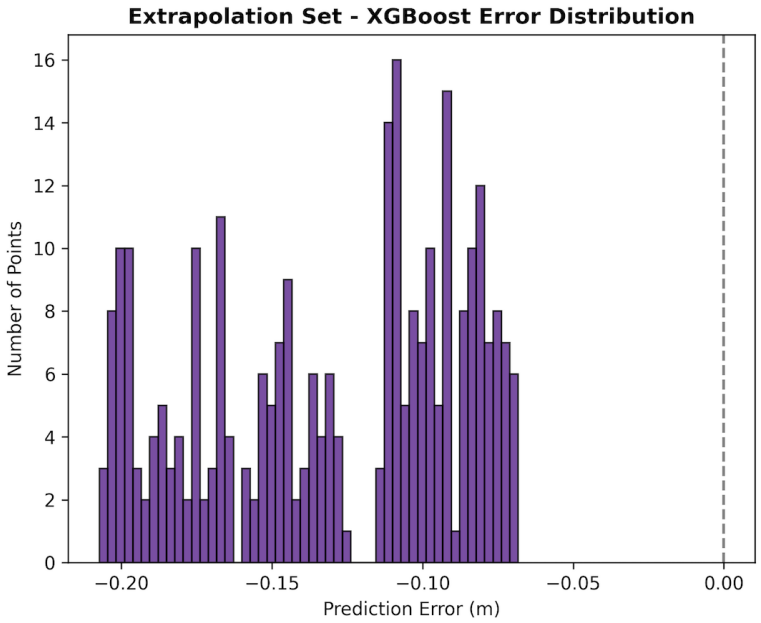


Figure 29: XGBoost Extrapolation Set Error Distribution Histogram

The SVR exhibits a different but equally problematic failure mode. Due to the use of the Radial Basis Function (RBF) kernel, the model's predictions rely on the similarity between the input and the "support vectors" (training points). As the extrapolation inputs move far away from the training feature space, the kernel values decay, causing the prediction to revert towards the mean of the training data (Smola and Schölkopf, 2004). Visually, this results in a scatter plot (Figure 30) that curves downward for the highest discharges, leading to significant underestimation similar to tree-based methods.

The MLR model, conversely, maintains its linear projection indefinitely. While this mathematically allows for extrapolation, the model continues to suffer from the high bias observed in the previous sections. The scatter plot (Figure 31) shows that while it does not "saturate," the large spread of residuals indicates that the linear approximation fails to capture the complex hydraulic interactions that occur during these extreme flows.

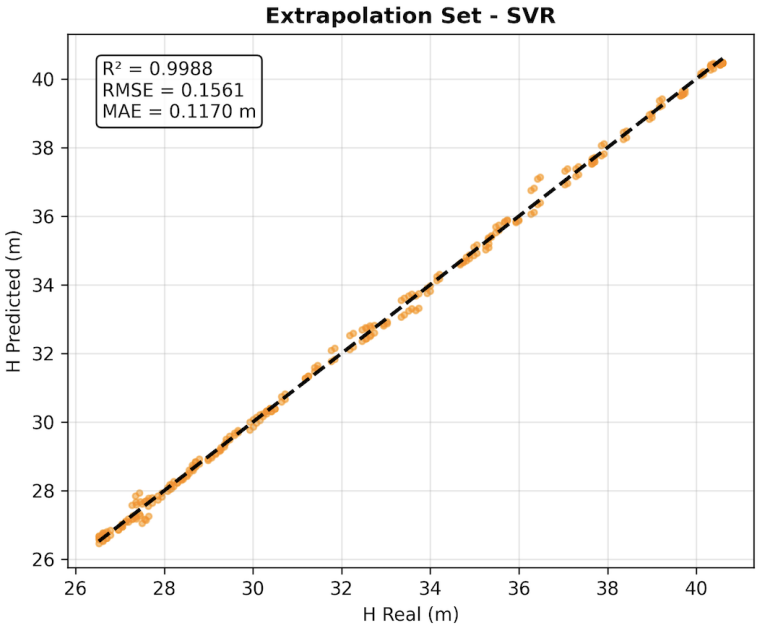


Figure 30: SVR Extrapolation Set Scatter Plot

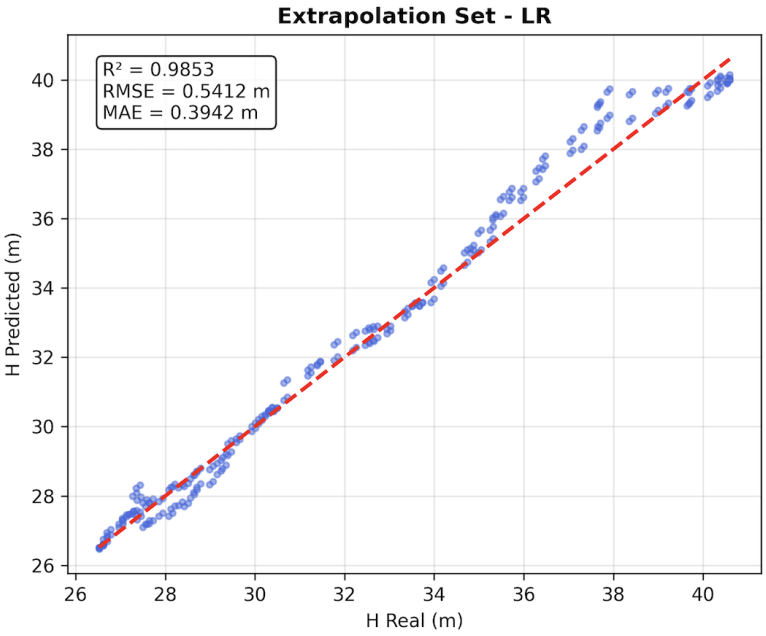


Figure 31: MLR Extrapolation Set Scatter Plot

In this comparative analysis, the Gaussian Process Regression (GPR) emerges as the robustest architecture for the extrapolation regime. While its deterministic accuracy (RMSE) degrades, it offers a decisive advantage: Uncertainty Quantification.

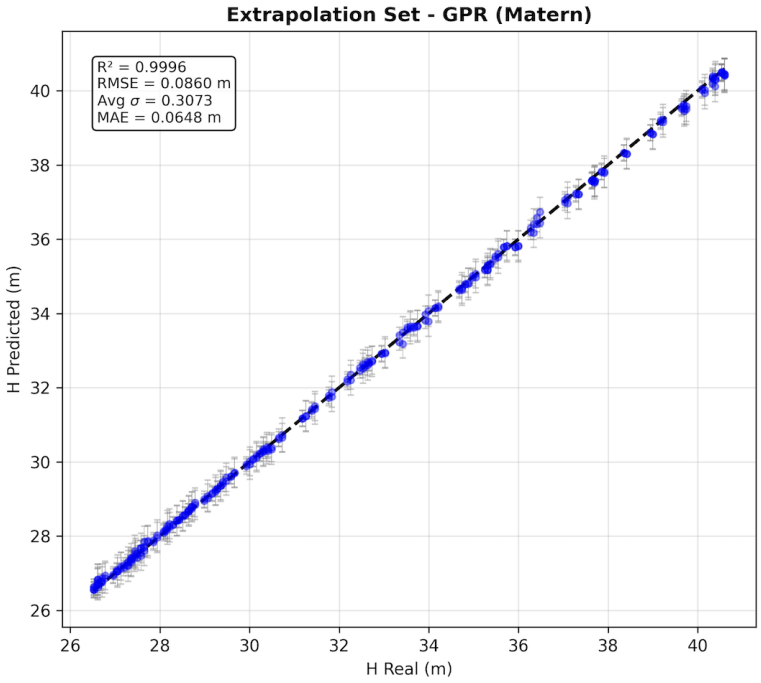


Figure 32: GPR Model Extraction Set Scatter Plot

As illustrated in the GPR scatter plot (Figure 32), the GPR model explicitly signals its lack of confidence. Unlike the Random Forest, which confidently predicts a wrong (saturated) value, the GPR produces a predictive mean accompanied by widening confidence intervals ($\pm 2\sigma$) as the discharge increases (Rasmussen and Williams, 2006).

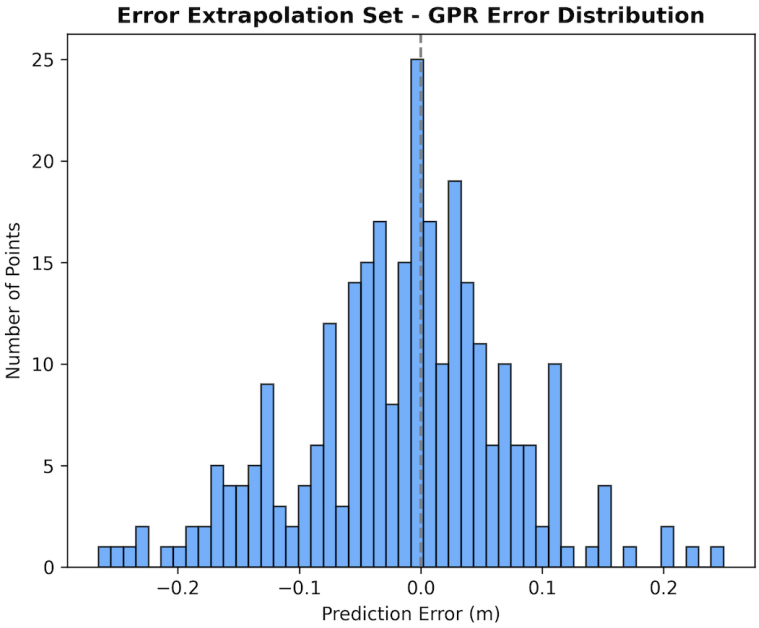


Figure 33: GPR Model Extrapolation Set Error Distribution Histogram

Regarding accuracy, the GPR error distribution (Figure 33) distinguishes itself remarkably from the tree-based models. Instead of the shifted, biased distribution seen in RF model, the GPR maintains a typical Gaussian bell shape centered near zero. This indicates that the model's residuals remain normally distributed and symmetrical even when extrapolating. Mathematically, this stability suggests that the Matérn kernel has successfully captured the underlying smoothness of the hydraulic function, allowing the model to project the stage-discharge trend forward without the artificial "saturation" ceiling that limits decision trees (Rasmussen and Williams, 2006).

In conclusion, the analysis of the extrapolation set reveals a critical trade-off between local precision and global robustness. While the tree-based ensembles demonstrated superior accuracy within the training range, their structural inability to extrapolate renders them unreliable for predicting unprecedented catastrophes. Conversely, the Gaussian Process Regression proved to be the only architecture capable of maintaining physical consistency and providing a necessary safety margin through uncertainty quantification. This distinction identifies the GPR as the most robust candidate for operational scenarios involving extreme events, setting the stage for the broader interpretative analysis in the following chapter.

To synthesize the findings, Table 5 presents a final quantitative comparison of the five surrogate models. By aggregating the global performance metrics (R^2, RMSE, MAE, and MBE) and explicitly separating the Test and Extrapolation datasets, this summary directly contrasts the models' interpolation accuracy with their physical robustness when predicting extreme, out-of-bounds flood events.

Dataset	**Model**	**R^2**	**RMSE (m)**	**MAE (m)**	**MBE (m)**
Test Set	Linear Regression (LR)	0.9886	0.5117	0.3997	-0.0535
	Support Vector Reg. (SVR)	0.9989	0.1625	0.1197	0.0092
	Random Forest (RFR)	0.9927	0.4097	0.3036	0.2706
	XGBoost	0.9907	0.4638	0.3463	-0.3463
	Gaussian Process (GPR)	0.9991	0.1422	0.0773	0.0458
Extrapolation Set	Linear Regression (LR)	0.9853	0.5412	0.3942	0.1588
	Support Vector Reg. (SVR)	0.9988	0.1561	0.1170	-0.0064
	Random Forest (RFR)	0.9987	0.1587	0.1514	-0.1514
	XGBoost	0.9991	0.1372	0.1305	-0.1305
	Gaussian Process (GPR)	0.9996	0.0860	0.0648	-0.0243

Table 5: Direct comparison of statistical performance metrics (R^2, RMSE, MAE, MBE) across the five machine learning architectures. The evaluation is focused exclusively on the models' interpolation capabilities (Test Set) and robustness outside the training domain (Extrapolation Set).

4.3 Flood Maps Comparison

To complement the statistical metrics, a spatial analysis of the predictive errors was conducted by generating difference maps ($Map_{predicted} - Map_{real}$) for specific discharge scenarios. This visualization strategy, as outlined in Section 2.3.1, allows for the identification of localized failure modes that aggregate metrics like RMSE often obscure.

In this section, the Extreme Gradient Boosting (XGBoost) model is selected as the primary case study to represent the behavior of the tree-based architectures. This selection was driven by three key factors:

1. The XGBoost difference maps provide the most distinct visual representation of the "recursive partitioning" mechanism, offering a clear pedagogical example of the decision tree's discretization process.

2. Morphologically, the spatial error distribution of XGBoost serves as a comprehensive proxy for the entire class of tree-based ensembles, mirroring channel saturation and localized discontinuities.

3. Highlighting the discontinuous nature of the XGBoost maps effectively contrasts with the Gaussian Process Regression, sharpening the distinction between the "patchy" behavior of trees and the "fluid" consistency of kernel-based methods.

In the interpolation regime (represented by discharges $Q = 8,200$ $Q = 12,230$ and $14,120$ m^3/s), the difference maps reveal a distinct spatial structure in the error distribution. As shown in Figure 34, the error is not uniformly distributed across the floodplain. Instead, we observe a phenomenon where zones of high error (dark blue patches, indicating underestimation) are immediately adjacent to zones of low error (lighter regions).

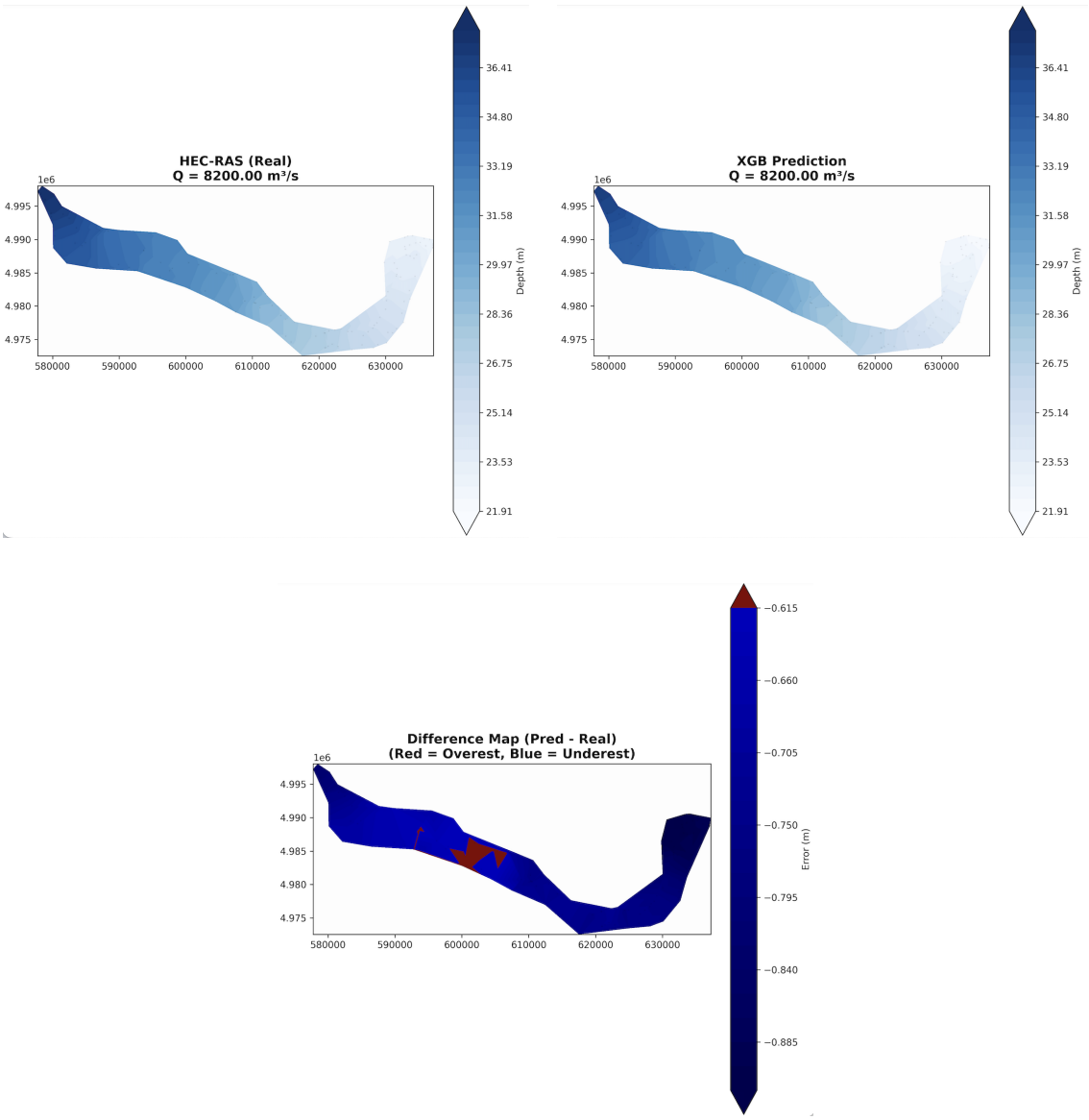


Figure 34: Spatial Error Analysis: XGBoost Difference Map for Q = 8,200 m^3/s (Testing Set). The blue gradient indicates a systematic underestimation of water depth

This sharp discontinuity, where "big error zones" border "low error zones", that are visible not only in this specific case but in most of the contour error maps created, is attributable to two factors:

- Topographic Gradients: The largest errors frequently align with the river banks or steep embankment features. In these zones of high geometric gradient, a minor deviation in the predicted water surface elevation translates into a significant discrepancy in depth, creating "hotspots" of error along the channel margins.
- Algorithmic Partitioning: The "patchy" texture of the difference map visually confirms the recursive partitioning mechanism of the XGBoost algorithm. Unlike the smooth, continuous surface of a hydraulic fluid, the decision tree model divides the river geometry into discrete hyper-rectangles. Consequently, the error often jumps abruptly across the boundaries of these spatial partitions, failing to capture the smooth hydraulic gradient of the free surface (Hastie et al., 2009).

Moving to the extrapolation set ($Q = 14,800$ and $Q = 15,000\ m^3/s$), the spatial maps provide visual confirmation of the "saturation bias" identified in the scatter plots. Saturation bias refers to the structural limitation where a decision tree model cannot predict a value higher than the maximum target observed in the training data (H_{train_max}). Consequently, when the physical water level rises beyond this historical maximum during an extreme flood, the model's prediction remains "saturated" or capped at the lower training value

In the difference maps for Q=14,820 (Figure 35) and Q=15,000 (Figure 36), this phenomenon is visually dominant. The error field is covered in deep blue areas, indicating a pervasive negative bias (underestimation).

Critically, the maximum error occurs on the valley floor path of the channel (deepest point). This is physically consistent with the saturation hypothesis: since the flow depth increases most dramatically during a flood, the gap between the rising real water level and the "capped" model prediction value is largest in the main channel, resulting in the deep blue artifacts observed.

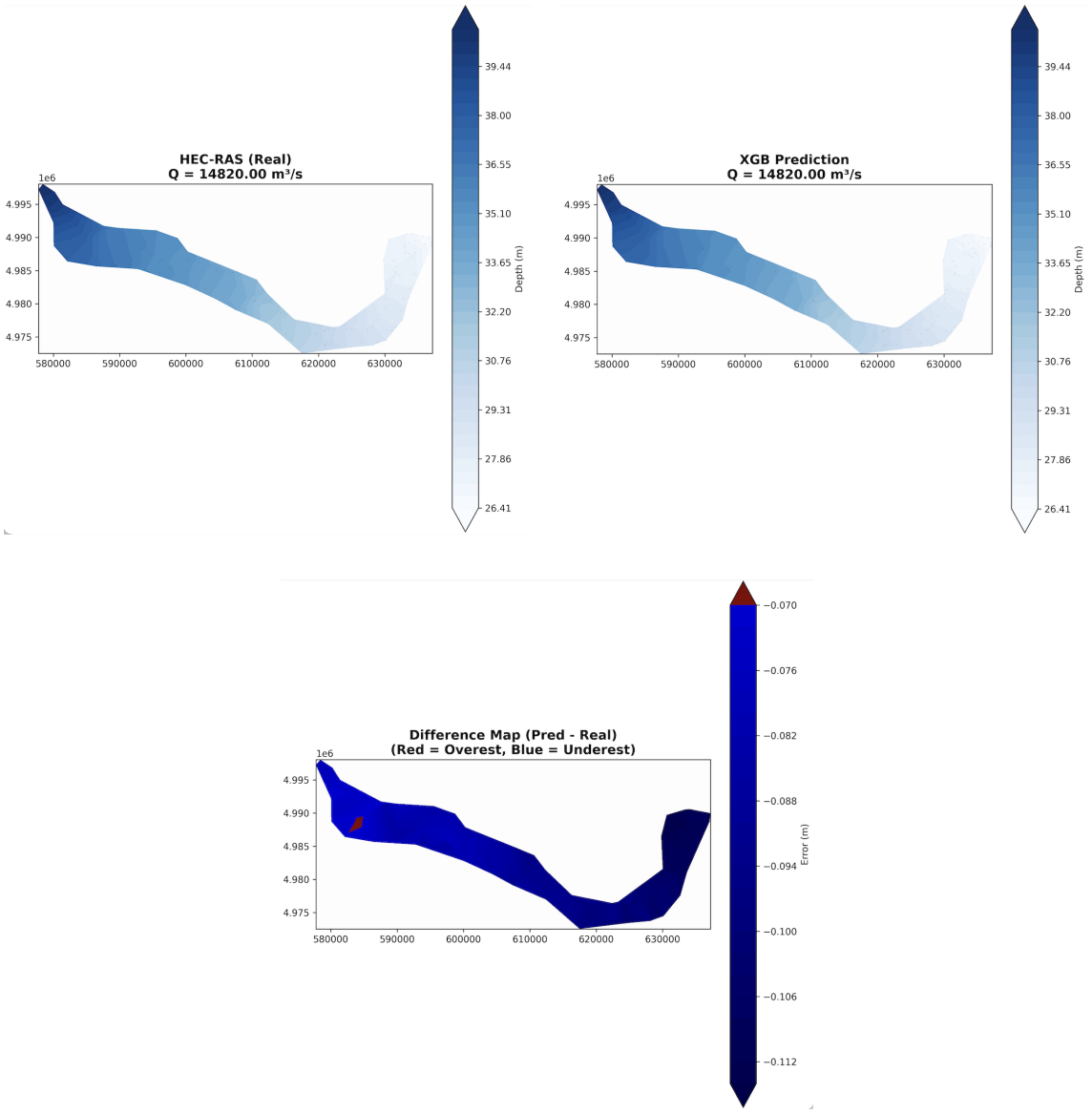


Figure 35: XGBoost Difference Map (Q = 14,820 m^3/s): Persistent underestimation across the main channel

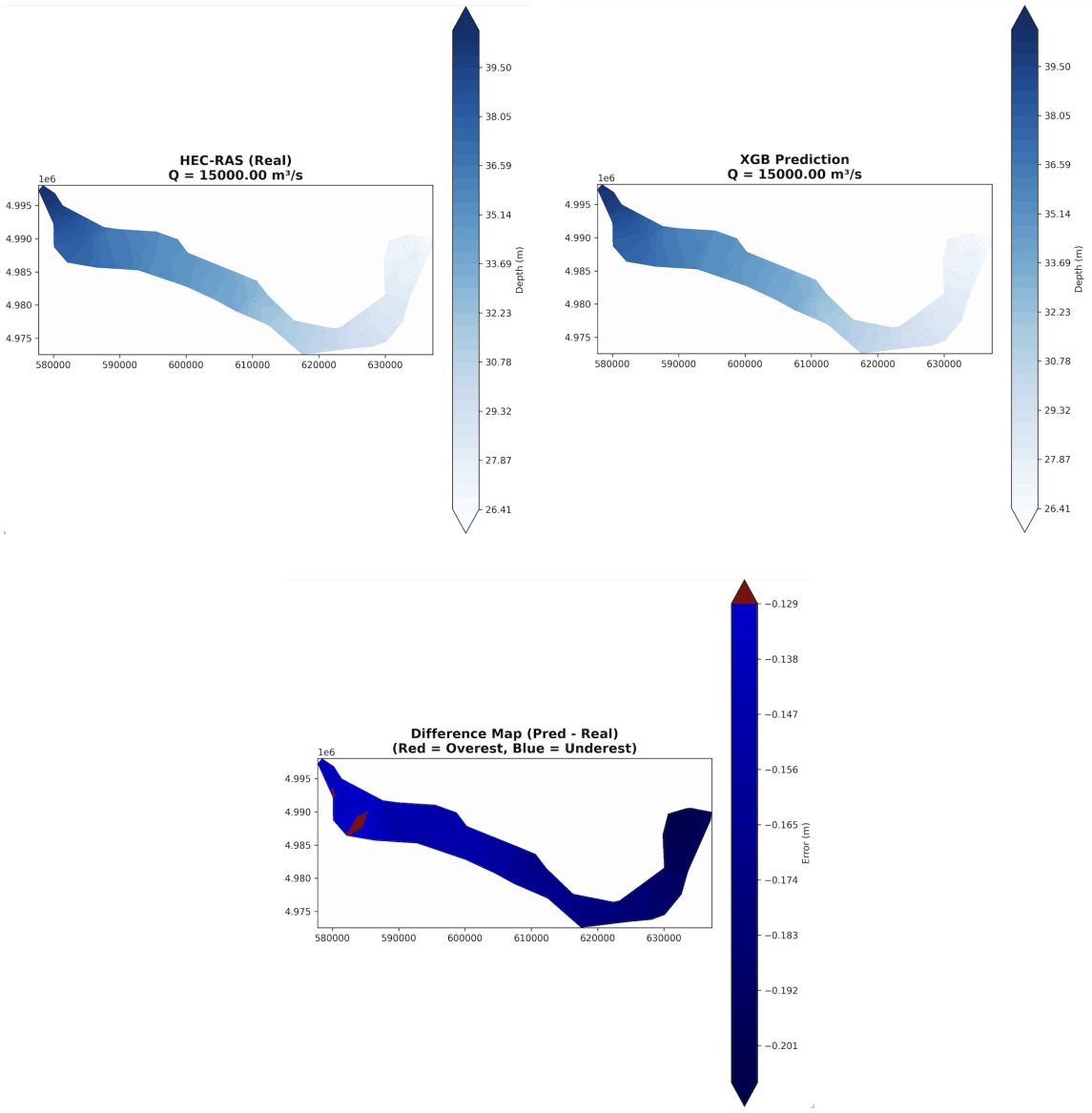


Figure 36: XGBoost Difference Map (Q = 15,000 m^3/s): The "saturation" effect of the decision trees is still most pronounced.

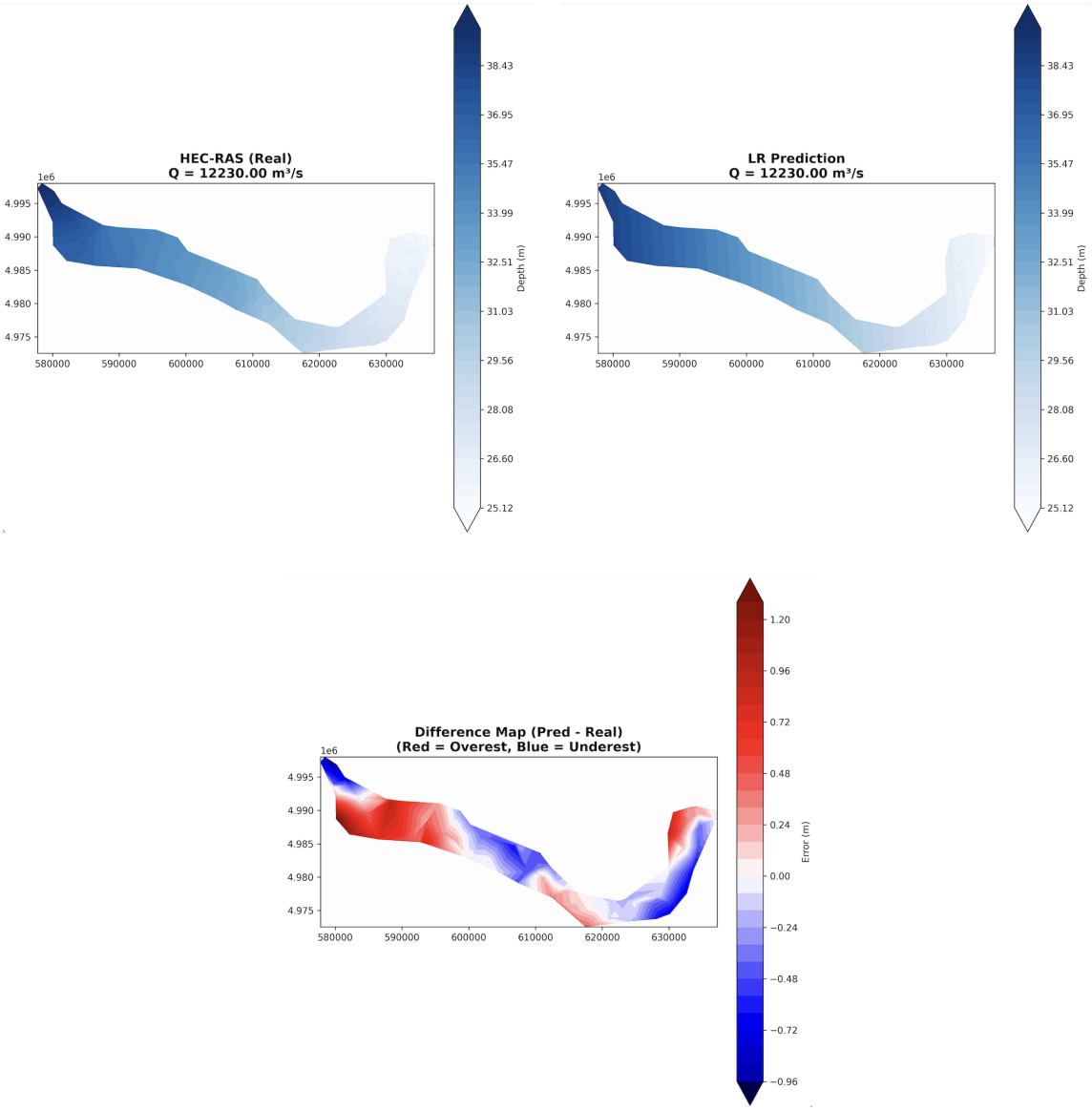


Figure 37: MLR Difference Map (Q = 12,230 m^3/s): Broad, smooth error zones caused by the model's structural rigidity

Finally, comparing the spatial outputs of the different architectures tested in this study:

- The Multiple Linear Regression produced the most spatially incoherent maps (e.g., Figure 37) considered the "worst" spatial maps. This is due to its planar assumption that forced a rigid surface over a complex, curving topography, resulting in massive, spatially correlated error zones that ignored local hydraulic features entirely.
- As observed in Figures 34 and in Figure 36, the tree-based models produce highly detailed but "noisy" maps. While they successfully capture the general shape of the inundation (Low Bias in interpolation), their output is marred by the artificial "staircase" artifacts of the decision trees, making them less physically realistic for hydraulic mapping.
- In terms of physical consistency, the Gaussian Process Regression (GPR) yields the superior error maps (e.g., Figure 38). Due to the smoothness properties of the Matérn kernel (Rasmussen and Williams, 2006), the GPR error fields do not exhibit the sharp, artificial discontinuities seen in the XGBoost results. Instead, the error transitions smoothly across the floodplain, mirroring the continuous nature of water. Furthermore, during extrapolation (Figure 39, while GPR errors do increase, they remain spatially homogeneous (centered around zero) rather than exhibiting the distinct "blue channel" systematic bias seen in the tree-based models.

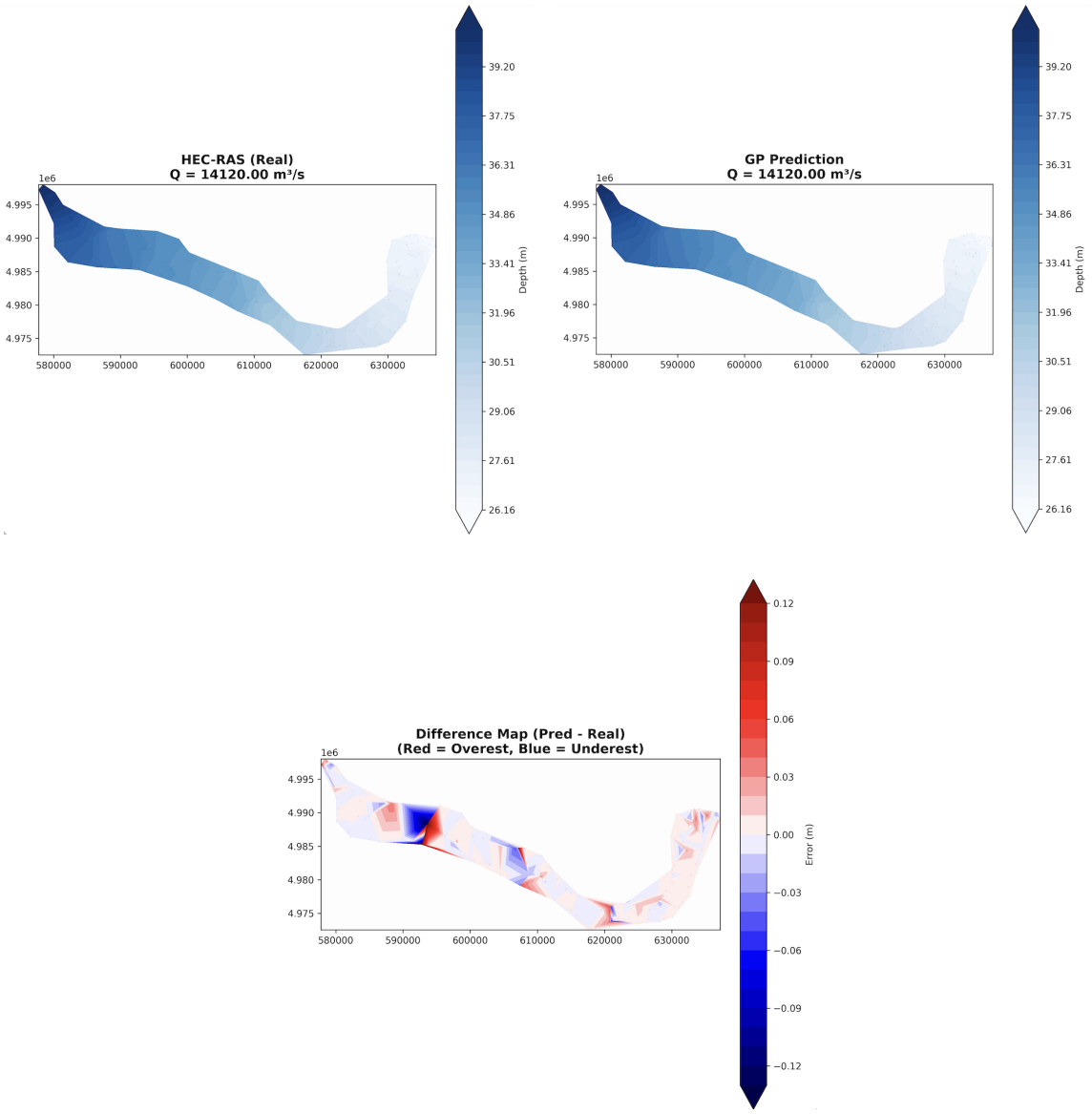


Figure 38: GPR Difference Map (Q = 14,120 m^3/s)

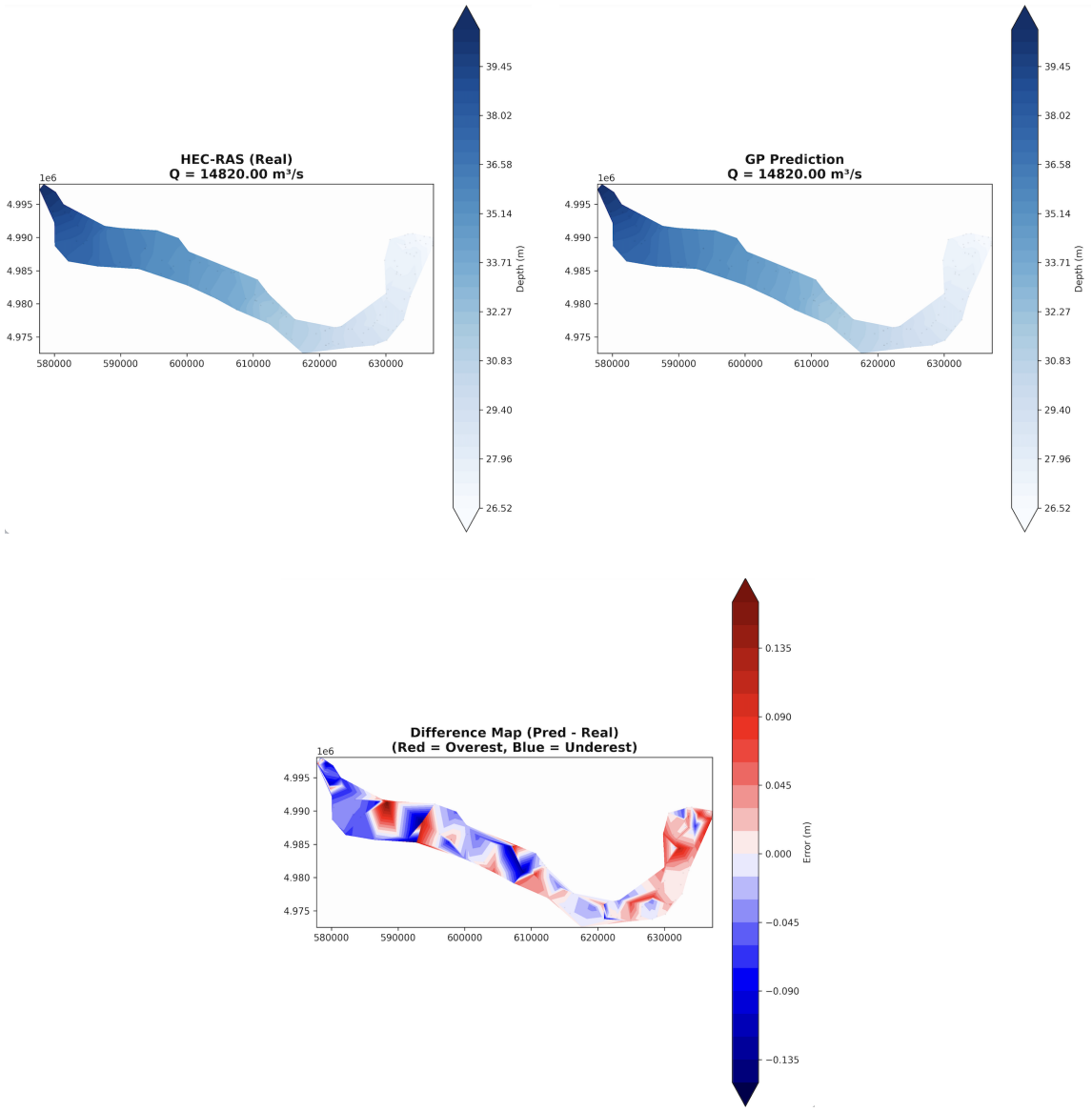


Figure 39: GPR Difference Map (Q = 14,820 m^3/s)

This spatial evidence reinforces the final selection of GPR as the optimal model: it is the only architecture that produces a flood map that "looks" and behaves like a fluid: continuous, smooth, and robust to geometric variations.

4.4 Operational Impact Visualization: Inundation Mapping

The final phase of the results analysis shifts the focus from statistical accuracy to operational applicability. While the spatial difference maps (discussed in Section 4.3) quantify mathematical errors, they do not immediately convey the physical consequences of a flood event. To bridge this gap, Inundation Maps were generated following the methodology described in Section 2.4, translating the predicted water surface elevations into tangible observable risk scenarios for urban and agricultural sectors.

The analysis begins with a macroscopic perspective of the entire modeled reach. Figure 40 illustrates the inundation extent for the total studied area (approx. 90 km of the Po River) for a flood event with a discharge equal to $Q = 4970\ m^3/s$. This comprehensive visualization confirms that the machine learning pipeline successfully reconstructs the primary hydraulic corridor, correctly confining the flow within the major embankment system even at high discharges.

As it can be seen in the Figure 40 and Figure 41, the spatial distribution of the flooded area is not uniform along the reach. As observed in the two global maps, the inundation extent widens significantly in the central and downstream sections. This phenomenon is hydraulically consistent with the river's morphological characteristics. In these sections, the channel exhibits high sinuosity (meandering). The presence of these meanders increases the flow path length, thereby reducing the effective longitudinal slope and increasing hydraulic resistance (Chow, 1959). This inevitably reduces flow velocity. In accordance with the principle of continuity ($Q = v \cdot A$), this deceleration forces an increase in the wetted cross-sectional area to maintain discharge (Sturm, 2001). This hydraulic compensation results in higher water surface elevations and a wider expansion of flood waters onto the adjacent floodplain.

Figure 40: Global Inundation Map: Modeled Po River reach from the HEC-RAS $Q = 8200\ m^3/s$ flood event simulation results, overlaid on satellite imagery.

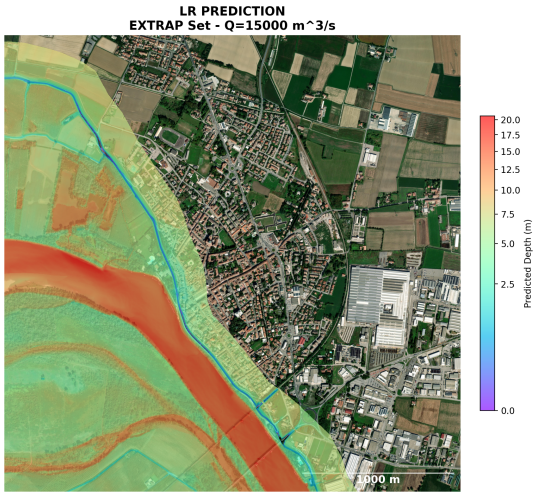
Figure 41: Global Inundation Map: Modeled Po River reach from the HEC-RAS $Q = 8200$ m^3/s flood event simulation results, overlaid on satellite imagery.

However, while these global maps are essential for validating general flow consistency, their scale is too broad to identify localized vulnerabilities, such as specific breached levees or threatened infrastructure. To provide a concrete assessment of risk, the study narrowed its focus to the specific area of Casalmaggiore. This selection was driven by a dual necessity:

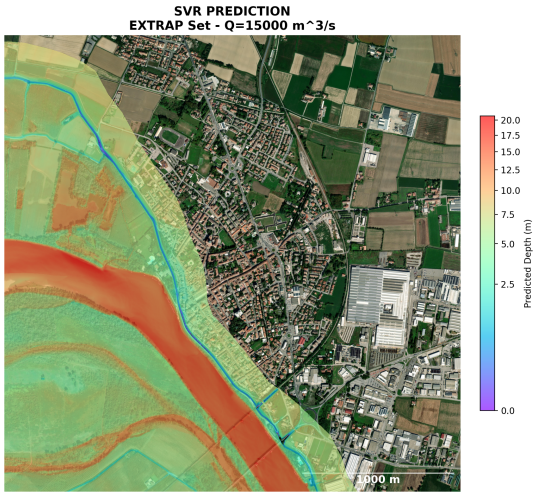
- Computational Efficiency: Generating high-resolution ($1m \times 1m$) georeferenced raster maps for the entire 90 km reach for every simulation (20+ discharge scenarios across 5 models) imposes a prohibitive computational burden. By defining a targeted Region of Interest (ROI), the processing time was optimized, allowing for rapid generation of detailed maps.

- Risk Identification: Casalmaggiore represents a critical interface between the hydraulic domain and human activity. It contains a complex mix of residential zones, industrial facilities, and agricultural fields. Focusing on this area allows for a granular inspection of the "wet/dry" boundaries, making it possible to explicitly verify whether specific assets, such as the town center or key agricultural plots, would be affected by the predicted flood levels.

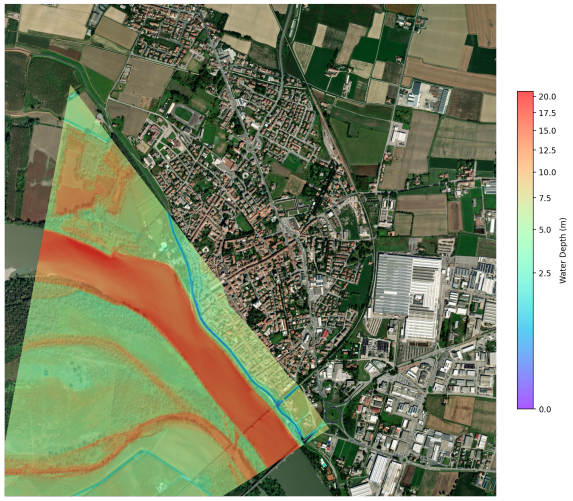
The resulting inundation maps were generated for all five machine learning models (LR, SVR, RF, XGB, GPR) across both the Testing and Extrapolation regimes. Specifically, Figure 43 presents a comparative example of the operational output produced by the Linear Regression (LR) (Fig. 42a) and Support Vector Regression (SVR) (Fig. 42b) models during an extreme extrapolation event ($Q = 15,000\, m^3/s$). These are also compared to the inundation map produced in Python based on the data extracted from the HEC-RAS simulation run with the Upstream Boundary equal to $Q = 15,000\, m^3/s$.



(a) LR Inundation Map ($Q = 15{,}000\, m^3/s$)



(b) SVR Inundation Map ($Q = 15{,}000\, m^3/s$)

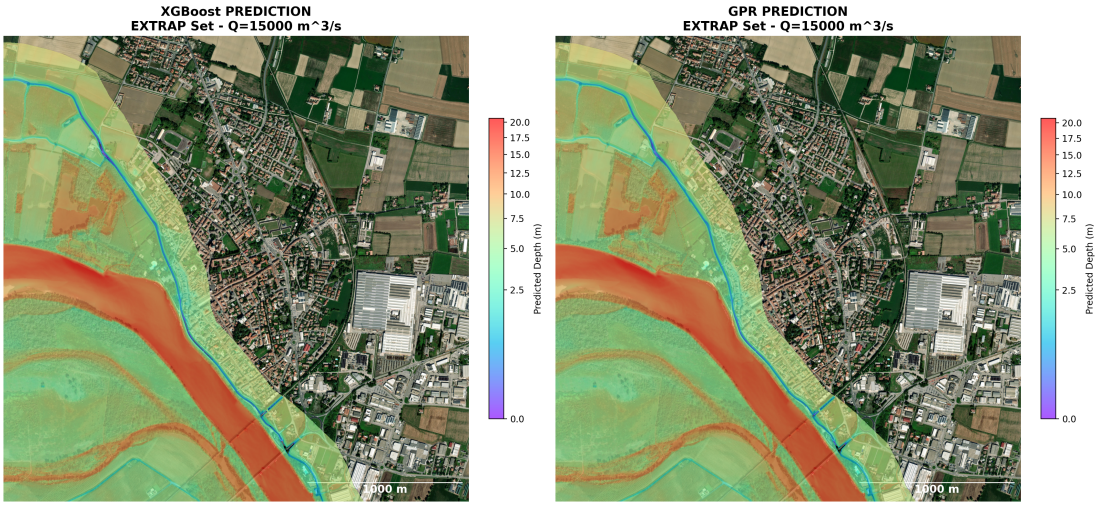


(c) HEC-RAS Simulation Inundation Map ($Q = 15{,}000\, m^3/s$)

Figure 42: Localized Inundation Maps (Casalmaggiore Focus Area): These maps translate the raw hydraulic depth predictions into operational risk scenarios. The visualization, comparing the physically-based HEC-RAS reference map with the Machine Learning predictions, reveals distinct morphological differences.

In these visualizations, the abstract error metrics discussed previously translate into physical realities. For instance, despite the statistical differences, both maps clearly identify the main channel (red zones, depths > 10 m) and the active floodplain (green/yellow zones). However, a critical visual comparison between the HEC-RAS reference map and the ML-generated predictions reveals notable morphological differences. The inundation extent plotted directly from the HEC-RAS data exhibits a more discrete, granular surface with sharper boundaries, strictly adhering to the local topographic constraints of the hydraulic mesh. In contrast, the Machine Learning models tend to generate a smoother, more continuous water surface. While this mathematical smoothing effectively captures the global flood dynamics, it can result in a slightly broader spatial coverage along the floodplain margins, as the algorithms interpolate across localized dry patches or minor embankments that the physical model explicitly resolves. Nevertheless, this confirms that the developed models are not just mathematical exercises but viable tools. By producing these maps, the study demonstrates that it is possible to generate detailed, georeferenced hazard scenarios in near real-time, providing civil protection authorities with the immediate intelligence needed to assess whether a specific town or agricultural district is at risk.

Further analysis of the Casalmaggiore maps reveals critical operational distinctions between the architectures. For insight The XGBoost map (Figure 43a)depicts a more contained flood extent compared to the GPR (Figure 43b) for a $Q = 15000\, m^3/s$ flood event. This is the direct operational consequence of the "saturation bias": because the tree model underestimates the peak channel stage, it calculates less lateral spillover. Consequently, for example, a civil protection plan based solely on the XGBoost map would unintentionally underestimate the necessary evacuation zone.



(a) XGBoost Inundation Map ($Q = 15,000\,m^3/s$) (b) GPR Inundation Map ($Q = 15,000\,m^3/s$)

Figure 43: Localized Inundation Maps (Casalmaggiore Focus Area): Differences between XGBoost Inundation Map and GPR Inundation map for a $Q = 15,000\,m^3/s$ flood event

Ultimately, the distinct "saturation bias" exhibited by the tree-based models creates a systematic blind spot for decision-makers, effectively masking the true potential of catastrophic flooding in extrapolated scenarios and leading to unintentional underestimates of necessary evacuation zones. Conversely, the Gaussian Process Regression, while sacrificing some local precision, maintains the physical continuity of the flood wave and provides a necessary safety margin through its uncertainty estimates. This fundamental trade-off between the ability to fit historical data and the capacity to robustly predict unprecedented events constitutes the pivotal finding of this study. These conflicting performance characteristics set the stage for the Discussion, where the operational feasibility of deploying these architectures within real-time Early Warning Systems is critically evaluated.

5 Discussion

This research compared five machine learning architectures and revealed a clear compromise between achieving high prediction accuracy within the range of conditions represented in the training data and maintaining reliable performance when models are applied to previously unseen operational scenarios.

While tree-based ensembles, specifically Random Forest and XGBoost, showed superior accuracy when interpolating within the training domain, their structural limitation in extrapolating beyond observed values produced a critical "saturation bias" during extreme flood events. In fact, despite the higher computational cost and slightly lower interpolation accuracy, the Gaussian Process Regression (GPR) model emerged as the preferred architecture for operational Early Warning Systems (EWS) in this context. This preference is mainly related to its ability to preserve physical plausibility, avoiding artificial upper limits in predicted water levels while simultaneously providing a rigorous quantification of predictive uncertainty (σ), which is essential for risk-informed decision-making. The capacity of GPR to generate smooth and continuous flood surfaces that resemble fluid dynamic behaviour contrasts with the discontinuous and locally fragmented patterns produced by the recursive partitioning structure of tree-based models (Rasmussen and Williams, 2006).

5.1 Model Efficacy and Selection

The primary objective of this research was to determine how effectively Machine Learning algorithms can predict Water Surface Elevations (WSE) to generate flood inundation maps, and to utilize comparative performance metrics to justify the adoption of a specific model for real-time forecasting. The findings of this study directly answer these inquiries, demonstrating that ML surrogates can indeed replicate high-fidelity hydrodynamic models with remarkable speed, though their operational safety depends entirely on the chosen architectural framework.

The systematic comparison of the five models provided a clear resolution to the first research objective. The statistical evaluation, utilizing RMSE, MAE, R^2, and MBE, whose values are reported in Table 5, revealed a fundamental dichotomy between interpolation precision and extrapolation robustness. As hypothesized, the Multiple Linear Regression (MLR) baseline proved insufficient due to its rigid parametric structure, which underfitted the non-linear hydraulic transitions. In contrast, the tree-based ensembles, particularly XGBoost and Random Forest, excelled in the interpolation regime, achieving near-perfect fidelity ($R^2 > 0.99$) and outperforming other architectures in pure numerical precision within the training bounds. However, when tasked with predicting extreme, unobserved flood events (the extrapolation set), the spatial and statistical results revealed a critical structural vulnerability in these algorithms: a "saturation ceiling" that led to dangerous, systematic underestimations of the water depth.

These performance results directly inform the final component of the research question: justifying the adoption of a specific model. The comparative metrics establish Gaussian Process Regression (GPR) as the most physically consistent choice for disaster risk reduction. While it incurs a higher training cost, GPR overcame the limitations of both the linear and tree-based models by maintaining physical consistency and continuous trends even in uncharted hydraulic territory (e.g., the 500-year return period). Furthermore, the spatial comparative analysis demonstrated that GPR maps behave most like a continuous fluid, avoiding the artificial, patchy fragmentation of decision trees. Most importantly, GPR provides a native quantification of uncertainty (σ), fulfilling the critical need for a probabilistic approach to real-time flood mapping. Therefore, the empirical evidence of this study confidently validates the GPR architecture as the optimal surrogate for operational flood forecasting.

5.2 The "mechanic" nature of physical modeling

Looking past purely predictive accuracy, the operational viability of these architectures depends strongly on their computational efficiency during real-world crisis scenarios. In this study, the Machine Learning workflow showed a clear advantage in the speed of post-processing and inundation mapping, which proved unexpectedly demanding when performed using traditional approaches. Although HEC-RAS offers a reliable hydraulic modelling engine, the sequence of running simulations, exporting outputs, and superimposing them onto satellite imagery is often mechanical and time-intensive, creating a significant bottleneck during emergency response operations.

Nevertheless, this speed advantage must be interpreted considering the specific constraints of the experimental framework. The hydraulic benchmarks adopted in this research were generated using Steady Flow simulations, which are computationally less demanding than the fully dynamic Unsteady Flow simulations required for real-time flood wave routing. For this reason, the difference in execution time observed in this study is likely conservative. In a fully operational environment involving dynamic routing, the time reduction achieved through ML surrogates would probably be substantially greater (Qureshi et al., 2025).

The largest efficiency improvement was observed during the generation of localized inundation maps (e.g., Casalmaggiore). As shown in Figure 44, the ML algorithms were able to produce high-resolution, georeferenced flood inundation maps almost instantaneously. In contrast, attempts to reproduce the same mapping workflow using standard GIS software (QGIS) revealed significant computational constraints

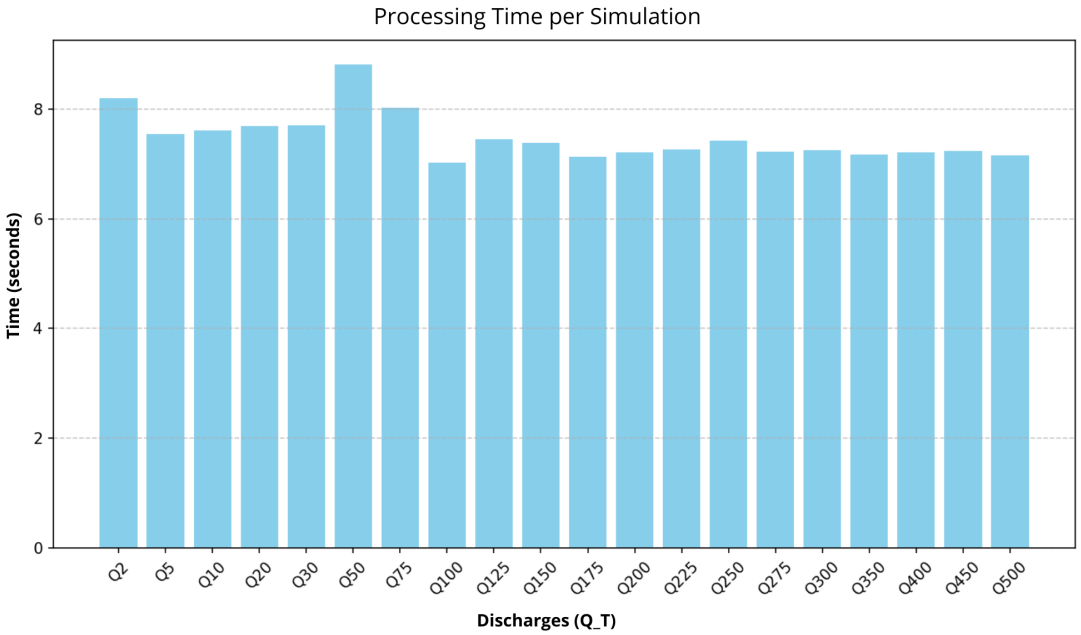


Figure 44: Computational Time Histogram for the inundation map of the 1500m x 15000m Area of Casalmaggiore: 20 Inundation Maps were computed to show the flood wave for each of the 20 simulation, one for each of the 20 discharges values considered in this study, produced in HEC-RAS

Fast interpolation tools in QGIS generated topologically inaccurate, 'broad' flood zones. On the other hand, higher-accuracy methods required over 10 minutes per scenario. In some cases, this even led to complete system failure due to memory overload. To better contextualize these performance results, it should be noted that all computations were performed on a standard workstation, specifically a 2020 MacBook Air M1 (8GB Unified Memory, 256GB SSD). This result demonstrates that ML surrogates can provide high-quality inundation mapping without relying on dedicated high-performance computing infrastructure. Such accessibility indicates that these tools could potentially be deployed on standard laptops within field command centres, reducing dependence on heavy server-side processing.

5.3 Operational Integration

The practical value of these models ultimately lies in their capacity to support Early Warning Systems (EWS) during time-critical emergency situations. As highlighted by the IPCC, an effective warning system is not solely defined by hydraulic precision, but also by its ability to provide clear and actionable information during the critical "golden hour" of emergency response. Within this framework, the transition from the deterministic, single-value predictions typical of tree-based models to the probabilistic formulation offered by Gaussian Process Regression (GPR) represents a substantial improvement in operational safety.

In practical applications, traditional EWS often operate as threshold-based systems, activating alarms only when predicted water levels exceed predefined limits. This binary "safe or unsafe" logic can be problematic, as it provides no indication of the confidence associated with the prediction. The GPR framework addresses this limitation by supplying a spatial "uncertainty map" (σ) alongside flood extent predictions. This additional information allows civil protection agencies to adopt risk-based decision strategies. Instead of relying on a single deterministic estimate, emergency planners can evaluate potential worst-case conditions, for example by considering the upper bound of the confidence interval, and plan evacuations according to the probability of flooding. This approach helps bridge a critical limitation identified by Teng et al. (2017), effectively transforming the modelling framework from a simple alarm system into a decision-support tool capable of visualizing spatial risk distribution.

The high computational efficiency of ML models also enables the implementation of ensemble-based forecasting strategies. Traditional hydraulic simulations such as HEC-RAS are often too computationally intensive to be executed repeatedly during rapidly evolving flood events, typically limiting operators to a small number of forecast scenarios. In contrast, ML surrogates capable of generating predictions within milliseconds could allow the rapid evaluation of numerous alternative scenarios, including variations in rainfall timing or soil saturation conditions. As discussed by Wang et al. (2025), the capability to explore multiple potential flood evolutions is essential for managing the increasing complexity and uncertainty associated with contemporary flood hazards. In this context, the ML framework proposed in this study does not simply replicate traditional hydraulic modelling, but enhances it by providing the speed and flexibility required to support timely community protection under increasingly variable climatic conditions.

5.4 Experimental Constraints and Limitations

However, the transition from this steady-state experimental framework to a fully operational and dynamic EWS introduces several challenges that must be carefully considered. A major limitation of this study is its reliance on steady-flow simulations, in which water surface elevation is directly linked to a constant peak discharge (Q). Real flood events are inherently dynamic and transient. Water levels are influenced not only by instantaneous discharge but also by the overall hydrograph shape and by hysteresis effects, where the rising and falling limbs of the flood wave produce different stages for identical discharge values. As highlighted by Wang et al. (2025), operational flood forecasting systems typically require the integration of multiple flood drivers together with temporal characteristics to provide reliable predictions.

The steady-state assumption also affects the parametrization of hydraulic roughness. The current modelling framework adopts a fixed Manning's n coefficient derived from a single calibration event. In reality, hydraulic resistance within the Po River floodplains shows strong seasonal variability. Dense maize or poplar cultivation during summer significantly increases surface fric-

tion compared to winter conditions, when floodplains are often characterized by exposed bare soil. Training the ML surrogates on a single roughness configuration makes them insensitive to these seasonal changes. An operational EWS would ideally include additional descriptors, such as vegetation density or a seasonality indicator, to represent the resulting variability in flood wave celerity and stage throughout the agricultural cycle. Although the GPR model performed best in this simplified setup, applying it to a fully dynamic, real-time environment requires a substantially expanded training dataset including time-dependent hydrographs and variable Manning's n values. Such an extension would inevitably increase feature space complexity and computational demand.

A fundamental limitation of this study, and of surrogate modeling in general, is the cascading nature of uncertainty. Because the machine learning models were trained on synthetic outputs generated by the HEC-RAS hydraulic model rather than observed historical flood extents, the ML architectures effectively operate as a "surrogate of a surrogate". Consequently, any epistemic or uncertainty embedded within the physical modeling phase is inherently transferred into the ML predictions. To rigorously interpret the predictive confidence of the surrogate models, particularly the uncertainty metric (σ) generated by the GPR, it is necessary to explicitly classify the sources of uncertainty into a three-tiered hierarchy:

- Input Data Uncertainty: This first foundational tier encompasses errors in the physical data used to build the baseline model. Specifically, this includes geometric and topographic inaccuracies within the LiDAR-derived Digital Terrain Model (DTM) and bathymetric surveys. If the underlying 2-meter resolution topography fails to capture micro-topographic features relevant to flood routing, this baseline error propagates through all subsequent modeling phases.

- Hydraulic Implementation Uncertainty: This second tier represents the epistemic and aleatoric errors inherent in the physical HEC-RAS model itself. Epistemically, the 1D model formulation relies on simplified momentum balances and approximations of energy loss that may not fully capture complex 2D or 3D hydrodynamic behaviors. Clearly, the uncertainties of the parameters remain high; while the Manning's roughness coefficient (n) was rigorously calibrated against the high-water marks of the past, the hydraulic resistance naturally exhibits spatial variability and fluctuates temporally due to the changing flow regimes and seasonal landscape/vegetation variations.

- Surrogate ML Deviation: The final tier of uncertainty is the specific mathematical deviation between the machine learning prediction and the HEC-RAS model output. It is critical to note that the variance (σ) provided by the GPR model specifically quantifies this third level of uncertainty. It measures the algorithm's confidence in emulating the HEC-RAS benchmark, but it does not account for the foundational geometric (first tier) or hydraulic (second tier) uncertainties inherent in the real-world flood event.

Therefore, while the high accuracy metrics obtained (e.g., $R^2 \approx 0.99$) validate the surrogate's capability to instantly replicate the numerical behavior of HEC-RAS, an operational Early Warning System (EWS) must eventually account for all three tiers to provide a complete quantification of physical flood risk.

The computational burden associated with GPR also introduces relevant scalability limitations. While prediction time after training remains negligible and satisfies the rapid-response requirements of EWS applications, training complexity increases cubically ($O(N^3)$) with dataset size. To maintain computational feasibility, the training dataset in this study was intentionally limited to 15 representative hydraulic simulations. A fully operational forecasting system would likely require hundreds or even thousands of synthetic scenarios in order to capture a broader range of flood hydrographs and potential breach conditions (Rasmussen and Williams, 2006). This training demand contrasts with the computational efficiency of XGBoost, which has proven particularly effective in managing large tabular datasets with lower computational overhead (Ni et al., 2020). Future developments may therefore benefit from exploring sparse Gaussian Process formulations or hybrid modelling approaches that combine the computational efficiency of gradient boosting with the probabilistic calibration typical of Bayesian methods.

The relatively small size of the training dataset ($N = 15$ simulations) also introduces a potential bias in the comparative model evaluation. Gaussian Processes are well known to perform effectively under small-data conditions because their Bayesian formulation imposes smoothness constraints through prior assumptions (Rasmussen and Williams, 2006). Conversely, Deep Learning architectures and ensemble tree methods such as XGBoost generally require large training datasets to adequately resolve feature relationships and reduce risks of overfitting or saturation near data boundaries (Teng et al., 2017). For this reason, the superior extrapolation performance of GPR observed in this study may partly reflect the limited dataset size. With a substantially larger training corpus, the saturation behaviour observed in XGBoost could potentially be reduced simply by providing training samples closer to extreme hydraulic conditions, which might alter the performance ranking obtained in this analysis.

When viewing these findings within the broader field of hydro-informatics, it is essential to contextualize these findings within the limitations of the current experimental design. While this thesis successfully demonstrated the feasibility of ML surrogates, the input variable space was intentionally simplified. For example, recent studies such as Wang et al. (2025) have developed emulators that incorporate multiple concurrent flood drivers, including tributary inflows, tidal interactions, and variations in hydrograph shape. In contrast, the present study adopted a simplified steady-state framework mainly driven by upstream discharge (Q) and spatial coordinates (X, Y). Although this reduction in input variables helped isolate the mechanical behaviour of the algorithms, it restricts the model's ability to represent the complex multivariate interactions that typically characterize compound flooding events in real-world conditions. As noted

by Wang et al. (2025), expanding the feature space to include these "multiple flood drivers" can quickly increase the complexity of interpolation-based approaches, indicating that future developments of this framework should integrate additional hydrological variables to achieve a level of fidelity comparable to comprehensive flood library methodologies.

These results are also strongly conditioned by the specific morphological characteristics of the selected case study. The machine learning models were trained and validated exclusively on the middle reach of the Po River (Cremona-Borgoforte), a river section characterized by a low water-surface gradient of 0.17% and by a transitional morphology between braided and sinuous channel patterns. The performance hierarchy observed in this work, where GPR shows improved extrapolation capability compared to tree-based models, may not be directly transferable to different river typologies. For instance, steep mountain catchments or highly urbanized and channelized rivers typically exhibit flashier and less diffusive flow regimes. Under these conditions, the "smoothness" assumption imposed by the Matérn kernel adopted in the GPR could lead to excessive smoothing of relevant hydraulic discontinuities, such as abrupt hydraulic jumps or rapid stage variations. In a similar perspective, Qureshi et al. (2025) highlights the importance of integrating AI-based models with morphodynamic assessments in order to account for evolving flow regimes and riverbed adjustments. The present study assumes a static bed configuration and therefore neglects sediment transport processes, which may significantly modify the stage–discharge relationship over time. For this reason, the identification of GPR as the most suitable modelling framework should be considered valid primarily for large, low-gradient alluvial rivers morphologically comparable to the Po River, while its applicability to substantially different geomorphological settings still requires further validation.

A key limitation of this study is related to the nature of the training dataset. The machine learning models were trained not on observed historical flood extents, but on synthetic outputs generated by the HEC-RAS hydraulic model. As a result, the ML architectures effectively operate as a 'surrogate of a surrogate.' Any epistemic uncertainty embedded in the hydraulic simulations, including uncertainties in the calibration of Manning's roughness coefficients (n) or inaccuracies in LiDAR-derived topography, is inherently transferred into the ML predictions. For this reason, the high accuracy metrics obtained (e.g., $R^2 \approx 0.99$) primarily reflect the capability of the models to reproduce the numerical behaviour of HEC-RAS rather than their capacity to represent the physical dynamics of real flood events. A more rigorous validation framework would require direct comparison of ML predictions with satellite-observed flood extents derived from historical events that were not included in the HEC-RAS calibration process.

6 Conclusions

Modern flood management still faces a major operational limitation: physically based hydraulic models require substantial computational time, while emergency response demands predictions that are available almost immediately. By developing and rigorously testing five Machine Learning surrogates against a high-fidelity HEC-RAS model of the Po River, this study results indicate that data-driven surrogate models can substantially reduce computational time while preserving an acceptable level of hydraulic accuracy.

The comparison highlighted an important trade-off between models that excel in statistical interpolation and those that better preserve physically consistent behavior. The tree-based ensembles, particularly XGBoost and Random Forest, proved to be exceptional interpolators. Within the bounds of the training data, they reconstructed flood depths with near-perfect fidelity ($R^2 > 0.99$), outperforming other architectures in pure numerical precision. However, an important limitation emerged. The analysis highlighted a structural limitation of these algorithms: their structural inability to extrapolate. When faced with extreme discharges that had never happened before, precisely the scenarios where an Early Warning System is most needed, these models hit a "saturation ceiling", effectively capping the water levels and potentially underestimating catastrophic , making these model extremely dangerous for EWS during record-breaking floods. In contrast, Gaussian Process Regression (GPR) provided the most physically consistent behavior among the tested models for disaster risk reduction. Although it incurs a higher training cost and offers slightly lower precision in low-flow regimes, it was preferred because it mimics the physical behavior of fluid dynamics, maintaining continuous and plausible trends even in uncharted territory. More importantly, the GPR provides a native quantification of uncertainty (σ). This permits flood maps to be interpreted probabilistically rather than deterministically, allowing decision-makers to visualize not just where the water might go, but how confident the model is about that prediction.

The operational contribution of this research extends beyond mere computational speed; it offers a concrete framework for upgrading Early Warning Systems (EWS) from deterministic monitoring to probabilistic forecasting. Currently, civil protection protocols often rely on a single hydraulic simulation triggered by a confirmed upstream discharge (Teng et al. (2017) and Wang et al. (2025)). The workflow that it is proposed in this thesis allows for a fundamental restructuring of this process through Real-Time Ensemble Mapping.

In a practical operational setting, this surrogate model could be directly coupled with upstream hydrological forecasts (e.g., rainfall-runoff models) that typically provide a range of potential discharge outcomes rather than a single value. Because the GPR surrogate can generate a complete inundation map in milliseconds, it becomes feasible to process the entire spread of a weather forecast ensemble (Wang et al., 2025). For example, if a meteorological model predicts a peak discharge at Cremona between $4,000 m^3/s$ and $6,000 m^3/s$ with varying proba-

bilities, the surrogate can instantly generate hundreds of inundation scenarios corresponding to this distribution. Instead of waiting for a single HEC-RAS simulation, operators can dynamically produce Probabilistic Inundation Maps. These maps would visualize not just the extent of the water, but the probability of inundation for every pixel in the domain, making possible to visualize exactly which buildings or areas under risk. Moreover, the lightweight nature of this architecture allows for flexible deployment. While traditional hydraulic modelling requires workstation-grade hardware often locked away in central offices, the GPR model trained in this thesis requires negligible RAM. It can run on a standard field laptop. This means that local emergency commanders could assess "what-if" scenarios directly on-site using standard laptops or tablets, without requiring continuous connectivity to a central server. This capability effectively bridges the gap between high-level hydraulic science and on-the-ground emergency response, providing actionable, risk-informed intelligence when it is most critical.

While this work shows promising results, the path toward a fully operational "Digital Twin" of the Po River, or at least the section studied, requires addressing the limitations identified in the experimental design, discussed in the previous paragraph. Future research should prioritize three key areas:

- To capture the true transient nature of floods, the training data must move beyond constant discharge (Q) inputs. Future iterations must incorporate time-varying hydrographs to account for volume, duration, and hysteresis effects, aligning with the "multiple-driver" approach advocated by Wang et al. (2025).

- Since the river is not a static concrete channel, it will be essential to integrate seasonal variables (e.g., vegetation density index) and morphodynamic updates (e.g., bed elevation changes due to previuos flood events) to prevent the model from becoming obsolete as the riverbed evolves, a necessity highlighted by Qureshi et al. (2025).

- The most intriguing next step is to test these algorithms on a completely different river typology. This study focused on a large, low-gradient alluvial river; however, it remains to be seen if the "smoothness" assumption of the GPR holds in steeper, flashier mountain catchments or highly urbanized channels. Replicating this methodology on a hydraulically distinct river would verify if the GPR is a universal solution or if different morphologies require different ML architectures.

- To overcome the cubic computational complexity ($O(N^3)$) of the GPR while retaining its probabilistic benefits, future work should explore Sparse Gaussian Processes or Deep Probabilistic Learning. These methods could allow the model to ingest thousands of training simulations, necessary for a robust operational system, without the prohibitive training times observed in this study (Rasmussen and Williams, 2006).

Overall, the results indicate that Machine Learning cannot fully replace physically based hydrodynamic modelling, but it can significantly enhance its operational applicability. By carefully selecting architectures that respect the physics of the problem, prioritizing robustness over raw accuracy, it is possible to build tools that not only predict the future of our rivers but help secure the safety of the communities living alongside them.

Declaration of Generative AI Use

During the preparation of this thesis, generative artificial intelligence (AI) tools were used to support the research process. Specifically, AI was employed to assist in the preliminary screening of scientific documents to assess their relevance to the research topic, to help reformulate and improve the clarity of written text, and to support code development by identifying and correcting programming errors.

The AI tools were not used to generate original scientific results, conduct independent analyses, or formulate research conclusions. All methodological choices, data processing, modeling decisions, interpretations, and final results presented in this thesis are the sole responsibility of the author.

References

Robert J Abrahart and Linda See. Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments. *Hydrological Processes*, 14(11-12):2157–2172, 2000.

Agenzia Interregionale per il fiume PO (AIPO). Fiume Po da Cremona al Mare. Idrovia Padana, 2021. Dati relativi a fondali e navigabilità del tratto.

M. Amadio. *Flood Risk Assessment in the Po river basin under a Climate Change scenario*. Master's thesis, Università Ca' Foscari Venezia, 2013.

Autorità di Bacino del Fiume Po. Catasto arginature maestre del fiume po: Da foce tanaro all'incile del po di goro. Atlante cartografico generale, Autorità di Bacino del Fiume Po, Parma, novembre 2004.

Autorità di Bacino Distrettuale del Fiume Po. *2. IL FIUME PO. Ambiente Naturale*. 2009. URL `https://www.adbpo.it/download/CartaItticaPo2009/pdf/FiumePo_AmbienteNaturale.pdf`. Riferimento al cambio di regime morfologico e pendenza (progressione 463 km).

Autorità di Bacino Distrettuale del Fiume Po. *Piano di Gestione del Rischio Alluvioni (PGRA) 2021-2027 – Distretto Idrografico del Fiume Po*. Autorità di Bacino Distrettuale del Fiume Po, 2021a.

Autorità di Bacino Distrettuale del Fiume Po. Problematiche e criticità - Capitolo 4. Piano Stralcio per l'Assetto Idrogeologico (PAI), 2021b. Documentazione Tecnica.

C. Bradford Barber, David P. Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22(4):469–483, 1996.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. ISBN 978-0387310732.

L. Bradimarte and G. Di Baldassarre. Uncertainty in design flood profiles derived by hydraulic modelling. *Hydrology Research - An International Journal*, pages 753–761, 2012.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Gary W Brunner. *HEC-RAS River Analysis System: Hydraulic Reference Manual, Version 5.0*. US Army Corps of Engineers, Hydrologic Engineering Center, Davis, CA, 2016.

Gary W. Brunner. *HEC-RAS, River Analysis System User's Manual Version 6.0*. U.S. Army Corps of Engineers, Hydrologic Engineering Center, Davis, CA, 2021.

A. et al Catsellarin. Optimal cross-section spacing in preissmann scheme 1d hydrodynamic models. *Journal of Hydraulic Engineering*, (135):95–105, 2009.

A. et al Catsellarin. Floodplain management strategies for ood attenuation in the river po. *River Research and Application*, (27):1037–1047, 2011.

T. Chai and R.R. Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7(3): 1247–1250, 2014.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Ven Te Chow. *Open-channel hydraulics*. McGraw-Hill, New York, 1959.

E. Coppola and F. Giorgi. Climate change projections for the Po River basin. *Natural Hazards and Earth System Sciences*, 10(12):2603–2615, 2010. doi: 10.5194/nhess-10-2603-2010.

Core Writing Team, H. Lee, and J. Romero, editors. *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC, Geneva, Switzerland, 2023. doi: 10.59327/IPCC/AR6-9789291691647.

M. Cozzi and F. Giani. Climate change scenarios and hydrological simulation for the Po River Basin. *Advances in Science and Research*, 5(1):103–108, 2011. doi: 10.5194/asr-5-103-2011.

Herbert Edelsbrunner, David Kirkpatrick, and Raimund Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4):551–559, 1983.

Environmental Systems Research Institute (ESRI). *ArcGIS Pro Documentation: Fundamentals of TIN Surfaces*. Esri, Redlands, CA, 2023. Available at: `https://pro.arcgis.com/en/pro-app/latest/help/data/tin/fundamentals-of-tin-surfaces.htm`.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

David C Garen. Improved techniques in regression-based streamflow volume forecasting. *Journal of Water Resources Planning and Management*, 118(6):654–670, 1992.

Sean Gillies et al. *Rasterio: geospatial raster I/O for Python programmers*. Mapbox, 2013. URL `https://github.com/rasterio/rasterio`.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

IBM Cloud Education. What is boosting?, 2020a. URL https://www.ibm.com/think/topics/boosting. Accessed: 2025-01-29.

IBM Cloud Education. What is random forest?, 2020b. URL https://www.ibm.com/think/topics/random-forest. Accessed: 2025-01-29.

Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA). Consumo di suolo, dinamiche territoriali e servizi ecosistemici. Edizione 2018, 2018. URL https://www.isprambiente.gov.it/files2018/pubblicazioni/rapporti/Rapporto{_}territorio{_}web.pdf.

Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1143. Montreal, Canada, 1995.

B. Lastoria and et al. Rapporto sulle condizioni di pericolosità da alluvione in italia e indicatori di rischio associati. *Rapporti ISPRA*, 2021.

Zaw Latt and Hartmut Wittenberg. Improving flood forecasting in a developing country: A comparative study of stepwise multiple linear regression and artificial neural network. *Water Resources Management*, 28(8):2109–2128, 2014.

Der-Tsai Lee and Bruce J Schachter. Two algorithms for constructing a delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3):219–242, 1980.

David R Legates and Gregory J McCabe Jr. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water resources research*, 35(1):233–241, 1999.

Ministero dell'Ambiente e della Sicurezza Energetica. Visualizzatore Mappe di Pericolosità e Rischio di Alluvioni, 2025. URL https://shorturl.at/oEijq. Consultato in data 5 Novembre 2025.

Alberto Montanari. Hydrology of the Po River: looking for changing patterns in river discharge. *Hydrology and Earth System Sciences*, 16(10):3739–3747, 2012.

Amir Mosavi, Pinar Ozturk, and Kwok-wing Chau. Flood prediction using machine learning models: Literature review. *Water*, 10(11):1536, 2018.

NHI. Development and validation of an xgboost-based model [insert full title from pdf]. *Clinical and Translational Science*, 18(1):e70172, 2024. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11895769/.

L. Ni, D. Wang, J. Wu, Y. Wang, Y. Tao, and J. Zhang. Streamflow forecasting using extreme gradient boosting. *Journal of Hydrology*, 586:124901, 2020.

NVIDIA Corporation. Glossary: Xgboost, 2024. URL https://www.nvidia.com/en-us/glossary/xgboost/. Accessed: 2025-01-29.

Thomas K Peucker, Robert J Fowler, James J Little, and David M Mark. The triangulated irregular network. In *Proceedings of the Digital Terrain Models (DTM) Symposium*, pages 516–532. American Society of Photogrammetry, 1978.

M. Pregnolato, P. F. Giordano, D. Panici, L. J. Prendergast, and M. P. Limongelli. A comparison of the UK and Italian national risk-based guidelines for assessing hydraulic actions on bridges. *Structure and Infrastructure Engineering*, 18(9):1297–1317, 2022. doi: 10.1080/15732479.2021.1923053.

Provincia di Cremona. Piano di Emergenza Fiume Po, 2012. Settore Ambiente e Territorio.

Provincia di Mantova. Programma Provinciale di Previsione e Prevenzione Protezione Civile (Rischio Idraulico), 2020. Settore Ambiente e Territorio/Protezione Civile.

Mohammad Uzair Anwar Qureshi, Afshin Amiri, Isa Ebtehaj, Silvio José Guimere, Juraj Cunderlik, and Hossein Bonakdari. Coupling HEC-RAS and AI for River Morphodynamics Assessment Under Changing Flow Regimes: Enhancing Disaster Preparedness for the Ottawa River. *Hydrology*, 12(2):25, 2025. doi: 10.3390/hydrology12020025.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14:199–222, 2004.

R. Stedinger and V. W. Griffis. Flood frequency analysis in the united states: Time to update. *Journal of Hydrologic Engineering*, pages 199–204, 2008.

Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, New York, 1999.

Terry W Sturm. *Open Channel Hydraulics*. McGraw-Hill, 2001.

P. Tarolli and et al. Why the 2022 Po River drought is the worst in the past two centuries. *PNAS Nexus*, 2(8):pgad251, 2023. doi: 10.1093/pnasnexus/pgad251.

J. Teng, A.J. Jakeman, J. Vaze, B.F.W. Croke, D. Dutta, and S. Kim. Flood inundation modelling: A review of methods, recent advances and uncertainty analysis. *Environmental Modelling & Software*, 90:201–216, 2017. doi: 10.1016/j.envsoft.2017.01.006.

J. Teng, J. Vaze, S. Kim, D. Dutta, A. Jakeman, and B. Croke. Enhancing the capability of a simple, computationally efficient, conceptual flood inundation model in hydrologically complex terrain. *Water Resources Management*, 33(2):831–845, 2019. doi: 10.1007/s11269-018-2146-7.

U.S. Army Corps of Engineers. *HEC-RAS Mapper User's Manual Version 6.4.1*. Hydrologic Engineering Center, Davis, CA, 2023.

C Vezzani. Drought management and early warning in the Po River Basin. Technical report, CAZALAC / Po River Basin Authority, 2014.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

W. Wang, Q. Wang, R. Nathan, and C. Velasco-Forero. Rapid prediction of flood inundation by interpolation between flood library maps for real-time applications. *Journal of Hydrology*, 609:127735, 2022. doi: 10.1016/j.jhydrol.2022.127735.

Wen Wang, Q.J. Wang, and Rory Nathan. Gaussian process regression on multiple drivers and attributes for rapid prediction of maximum flood inundation extent and depth. *Journal of Hydrology*, 649:132476, 2025. doi: 10.1016/j.jhydrol.2025.132476.

Cort J Willmott. On the validation of models. *Physical Geography*, 2(2):184–194, 1981.

Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30 (1):79–82, 2005.

XGBoost Developers. *Introduction to Boosted Trees*. DMLC, 2024. URL `https://xgboost.readthedocs.io/en/stable/tutorials/model.html`. Accessed: 2025-01-29.

Davide Zanchettin et al. The Po River Basin: Managing a complex system. *Regional Hydrological Impacts of Climatic Change*, 2008.