Deliverable 4.1:

Trustworthy AI models: hydrogeological risk assessment

Mitigating the risk of flooding and landslides via artificial intelligence with a view to extreme climate events

Co-funded by the
European Union

Deliverable Title	Trustworthy AI models: hydrogeological risk assessment
Deliverable number	4.1
Deliverable Lead	UNIPI
Related Work Package	WP4: Trustworthy AI for hydrogeological risk assessment and evaluation of people's risk awareness of existing areas
Author(s)	Francesco Pistolesi (UNIPI), Michele Baldassini (UNIPI), Matteo Mugnai (UNIPI), Francesco Marcelloni (UNIPI), Marco Avvenuti (UNIPI), Stefano Pagliara (UNIPI), Michele Palermo (UNIPI)
Dissemination Level	Public
Due Submission Date	11/11/25
Actual Submission	27/11/25
Project Number	101140345
Status	Version 1.0 (27/11/25)
Reviewed (Authors)	Elisabetta Cattoni (eCampus), Francesco Focacci (eCampus)
Start Date of Project	12 February 2024
Duration	24 months
Abstract	This deliverable presents the results of the design, development and evaluation of AI models for hydrogeological risk assessment within the SAFE-LAND project.
Status changes history	Version 1.0 (26/11/2025) — First release

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Table of Contents

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

1. Introduction

The goal of the SAFE-LAND project is to improve the prediction, prevention, and mitigation of hydrogeological hazards—specifically rainfall-induced landslides and floods—across Europe and neighboring regions, with a view to the increasing frequency of extreme climate events. The project aims to strengthen the resilience of communities and infrastructures exposed to such hazards by integrating *artificial intelligence (AI)* and physically based models (PBMs) into a unified decision support framework.

Landslides triggered by intense or prolonged rainfall are becoming more frequent due to climate change, threatening populations, infrastructures, and ecosystems. Traditional physically based models (PBMs) provide accurate hydro-mechanical simulations of slope stability but are computationally demanding and require expert configuration, which limits their usability in time-critical scenarios. SAFE-LAND addresses these challenges by developing AI-based surrogate models capable of reproducing the outputs of PBMs with comparable accuracy but at a fraction of the computational cost. These models can predict key indicators such as the *factor of safety (FoS)*, the *depth of the sliding surface*, and the *position of the water table*, allowing for fast and reliable assessments under diverse rainfall and soil conditions.

Within the framework of SAFE-LAND, Work Package 4, "AI Models for Landslide Prediction," focuses on the design, training, and validation of regression models based on machine learning and deep learning to approximate PBM results. These models were trained on a large synthetic dataset derived from more than 16,000 hydro-mechanical simulations that include various slope geometries, soil types, and rainfall scenarios. The resulting models achieved an average prediction accuracy that exceeded 94%, demonstrating their potential for rapid and effective risk evaluation in operational contexts.

The AI estimations are used to support the selection of appropriate mitigation strategies by mapping model outputs to the effectiveness matrix defined in Deliverable 3.2. This ensures that AI models contribute to a consistent workflow in which stabilization measures are derived from standardized criteria and adapted to the specific characteristics of each scenario.

This deliverable describes the results obtained during the design, development and testing of these AI models and their integration into the SAFE-LAND framework. Similarly to landslide modeling, the project also addresses hydraulic hazards, which are becoming increasingly frequent due to climate change. The hydraulic component of SAFE-LAND relies on physically based simulations performed with the HEC–RAS model, where multiple synthetic flood scenarios are generated to describe diverse hydrological conditions. Each scenario is driven by a triangular hydrograph that captures the temporal evolution of discharge during an extreme rainfall event, with peak flows computed through the rational method and adjusted using a drainage-area correction factor.

The simulations cover a 5×5 km^2 floodplain domain under different slope configurations (flat, mildly inclined, and steep), enabling the analysis of flow propagation across contrasting terrain conditions. For each scenario, HEC–RAS produces spatially distributed outputs such as water depth, flow velocity, and water arrival time, which represent key quantities for flood hazard assessment. Within SAFE-LAND, these hydraulic simulations will serve as the basis for training AI surrogate models capable of reproducing the behavior of the physically based solver with significantly reduced computational

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

effort, supporting rapid operational forecasting.

This deliverable describes the results obtained and is organized as follows: Section 2 gives an overview of machine learning, regression, and classification problems; Section 3 introduces the geotechnical simulations and dataset; Section 4 describes the AI models for landslides; Section 5 presents the hydraulic simulations and dataset; Section 6 describes the AI models for floods; Section 7 draws the conclusions.

2. Artificial Intelligence and Machine Learning

Machine Learning is the art and science of teaching computers to learn from data, uncover patterns, and make informed decisions without explicit programming instructions. It represents a shift from traditional rule-based programming paradigms, where developers meticulously write algorithms to perform specific tasks. Instead, Machine Learning algorithms can generalize from examples, extrapolate insights, and improve performance over time as they process more data.

The origins of Machine Learning can be traced back to the mid-20th century, with pioneers like Arthur Samuel laying the groundwork for what would become a transformative field. Initially rooted in artificial intelligence, Machine Learning experienced significant advancements fueled by computational advancements, algorithmic breakthroughs, and the explosion of data availability in the digital age. Today, Machine Learning is everywhere in our daily lives, from personalized recommendations on streaming platforms to autonomous vehicles navigating city streets. Its applications span diverse domains, including healthcare, finance, manufacturing, marketing, and beyond, revolutionizing processes, enhancing decision-making, and unlocking unprecedented insights from complex datasets.

The algorithms underpinning its functionality are central to the efficacy of Machine Learning. They comprise a spectrum of techniques, from classical methods like linear regression and decision trees to cutting-edge deep learning architectures such as convolutional neural networks and recurrent neural networks. Each approach has strengths, weaknesses, and suitability for particular tasks, enabling practitioners to tailor solutions to specific problems effectively.

The proliferation of Machine Learning has increased with the availability of robust frameworks, libraries, and tools that have made its application easier. Platforms like TensorFlow, PyTorch, and scikit-learn provide accessible interfaces for developing, deploying, and scaling Machine Learning models, lowering barriers to entry and empowering a broader community of researchers and engineers. However, together with its remarkable advancements and transformative potential, Machine learning also has inherent challenges and ethical considerations. For example, the issues surrounding bias, fairness, transparency, and accountability highlight the importance of responsible AI development and governance frameworks to mitigate unintended consequences and ensure equitable outcomes for all stakeholders. Looking ahead, the trajectory of Machine Learning promises continued innovation and evolution, driven by interdisciplinary collaboration and the convergence of fields such as computer science, statistics, mathematics, and domain-specific expertise. As algorithms become more sophisticated, data more abundant, and computational resources more powerful, the boundaries of what is achievable with Machine Learning continue to expand, leading to a future where intelligent systems work with

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

humans to tackle complex problems and advance society.

2.1 Classification and Regression

Classification and regression are two fundamental tasks in machine learning. These tasks enable systems to make sense of the data, infer relationships, and make informed decisions in various applications. Classification and regression represent distinct but interconnected paradigms within machine learning, each tailored to address specific types of problems and data. Although classification involves predicting discrete class labels or categories, regression focuses on estimating continuous numerical values. Together, they compose the so-called *supervised learning* (see 1), where models are trained on labeled data to make predictions based on input features. In classification tasks, the goal is to assign input data points to predefined categories or classes based on their features. For example, distinguishing between spam and legitimate emails, identifying handwritten digits in images, or predicting a patient's disease likelihood based on medical test results.

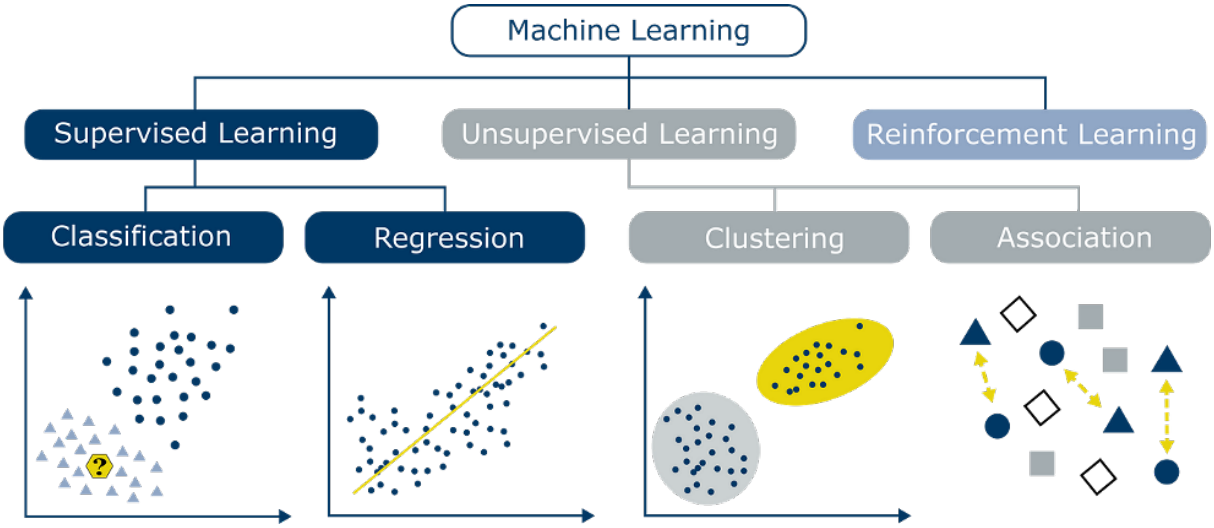


Fig. 1: Machine learning tasks.

Classification models learn decision boundaries in the feature space that separate different classes, allowing them to classify new (i.e., unseen before) data points accurately. Common classification techniques include logistic regression, decision trees, support vector machines (SVM), k-nearest neighbors (KNN), and neural networks. Each technique has strengths, weaknesses, and suitability for different data types and problem domains. For example, decision trees excel at handling categorical data and capturing nonlinear relationships, whereas SVMs are effective in high-dimensional spaces with clear separation between classes.

On the other hand, regression tasks involve predicting a continuous numerical value based on input features. This could include forecasting stock prices, estimating house prices based on property features, or predicting temperature based on weather variables. Regression algorithms learn to model the relationship between input features and output values, allowing them to make accurate predictions for new data points. Popular regression techniques include linear regression, polynomial regression, decision trees, random forests,

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

gradient boosting machines (GBM) and neural networks. Like classification algorithms, each regression technique has its strengths and weaknesses, making it important to select the most appropriate algorithm based on the characteristics of the data and problem at hand.

In both classification and regression, the performance of a machine learning model is evaluated using metrics such as *accuracy*, *precision*, *recall*, *F1 score*, *mean squared error (MSE)*, and *R-squared*. These metrics provide insights into how well the model generalizes to unseen data and can help guide the model selection and optimization process.

Although classification and regression represent powerful tools for predictive modeling, they are not without challenges. Overfitting, i.e., where a model learns to memorize training data rather than generalize to new data, and underfitting, i.e., where a model is too simplistic to capture the underlying patterns in the data, are common pitfalls that must be addressed through sophisticated techniques such as regularization, cross-validation, and ensemble learning. In addition, the choice of features, data preprocessing, and feature engineering are key in the performance of classification and regression models. Feature selection techniques such as principal component analysis (PCA), feature scaling, and normalization can help improve model performance and efficiency by focusing on the most relevant information in the data.

2.2 Segmentation

Segmentation represents another crucial task in the domain of machine learning, particularly within computer vision, remote sensing, medical imaging, and geotechnical data analysis. Unlike classification and regression—which predict global labels or scalar quantities for entire samples—segmentation is focused on delineating and labeling individual parts or regions within data instances. The goal is to partition an input (such as an image, spatial grid, or numerical field) into multiple meaningful segments that share similar properties or belong to specific classes.

In supervised segmentation, models are trained on labeled data, where each element—such as a pixel in an image or a cell in a numerical mesh—is assigned to a predefined class. This enables the model to learn spatial patterns, contextual dependencies, and local relationships, allowing for detailed predictions at a fine-grained level. Typical examples include identifying objects in images (semantic segmentation), separating instances of the same object type (instance segmentation), or mapping zones with different geotechnical characteristics within soil or rock formations.

Segmentation splits into two categories: *semantic* and *instance* segmentation. Semantic segmentation assigns a class label to each element (e.g., pixel, voxel, or grid cell), without distinguishing between individual instances of the same class. Instance segmentation, on the other hand, not only classifies, but differentiates between separate occurrences of similar objects. In geotechnical engineering, for example, segmentation can be used to identify distinct failure zones within a slope, differentiate layers of soil with varying mechanical properties, or delineate the extent of water-saturated regions based on predicted piezometric surfaces.

Modern segmentation approaches use traditional algorithms and deep learning architectures. Classical methods such as thresholding, clustering (e.g., k-means), region grow-

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

ing, or graph-based algorithms laid the foundation for automated segmentation. However, the advent of deep learning, particularly convolutional neural networks (CNNs) and encoder–decoder architectures like U-Net and SegNet, revolutionized the field by enabling end-to-end learning of complex spatial and contextual features. These architectures can effectively capture hierarchical representations, combining local texture information with global spatial context.

The evaluation of segmentation models relies on metrics that account for both accuracy and spatial consistency, such as Intersection over Union (IoU), Dice coefficient, precision, recall, and pixel accuracy. The choice of metric depends on the specific application and data characteristics. For example, in slope stability analysis or hydrogeological modeling, IoU can quantify how accurately the predicted unstable regions or saturated zones overlap with reference data.

Finally, segmentation extends the predictive power of machine learning from global inference to spatially distributed understanding. By capturing both the structure and variability within complex systems, segmentation models enable more detailed, interpretable, and actionable insights—especially in fields where spatial heterogeneity plays a critical role, such as geomatics, environmental monitoring, and geotechnical engineering.

2.3 Explainable AI

In an era where machine learning models are prevalent in decision-making in diverse domains, transparency, accountability, and trustworthiness are key. As these models are used in critical domains such as healthcare, finance, criminal justice, and autonomous systems, understanding how they generate predictions and recommendations is crucial to ensure fairness, mitigate biases, and foster user acceptance.

The explainability of machine learning models is the ability to understand and interpret the inner workings of these complex algorithms, highlighting the factors and features that drive their decisions. It encompasses the technical mechanisms underlying model predictions and the broader socio-ethical implications of algorithmic decision-making on individuals and society. The pursuit of explainable AI depends on the balance between achieving high performance and maintaining interpretability.

Although state-of-the-art machine learning models—particularly deep neural networks—often achieve high accuracy and predictive power, their black-box nature makes it difficult to understand how they arrive at their conclusions. This lack of transparency can hinder adoption, especially in applications where trust and accountability are crucial. The importance of explainability goes beyond the mere curiosity about how algorithms work. Explainable AI can foster user trust and acceptance of machine learning systems. Users are more likely to embrace AI-powered technologies when they can understand and validate the reasoning behind algorithmic decisions. In contrast, opaque or incomprehensible output can lead to skepticism, distrust, and even rejection of AI systems, undermining their effectiveness and potential societal benefits. Regulatory and ethical considerations further highlight the need for explainability. Increasingly, policymakers and regulatory bodies recognize the importance of transparency and accountability in algorithmic decision-making, enacting laws and guidelines that mandate the explainability of AI systems, particularly in sensitive domains.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Explainable AI is rapidly evolving, and researchers are developing various techniques to make the inner workings of machine learning models understandable. These approaches range from model-specific interpretability methods, such as feature importance scores and attention mechanisms in neural networks, to model-agnostic techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), which provide post hoc explanations for any black-box model. In addition, advances in interdisciplinary research, drawing on insights from computer science, statistics, psychology, and philosophy, enrich our understanding of what constitutes a meaningful explanation in AI. Concepts such as causal reasoning, counterfactual explanations, and human-centered design are continuously evolving as researchers seek to bridge the gap between technical accuracy and human comprehension in explaining machine learning models.

The pursuit of explainable AI has profound implications for future AI development and deployment.

3. Geotechnical Simulations and Dataset

This section provides an overview of the dataset generated and described in *Deliverable 3.2: Hydrogeological risk assessment of reference areas*. Its purpose is to briefly summarize the main characteristics of the data, the simulation process used to generate them, and to clarify how this dataset was used in the development and training of AI models.

3.1 Slope stability: basic concepts and parameterization

A *slope* (natural or man-made) is an inclined surface of soil or rock exposed to gravity. The gravitational force tends to move the soil downward: if it becomes higher than the resisting force, a slope failure (i.e., a landslide) occurs. Slope failure is based on soil type, groundwater location, seepage, and slope geometry.

The soil type is characterized by *mechanical* and *hydraulic properties*. The shear strength under saturated conditions—i.e., when soil voids are completely occupied by water—depends on the mechanical properties, such as the effective cohesion (c'), the effective friction angle (ϕ'), and soil unit weight (γ). In unsaturated conditions—i.e., when the soil voids are partially filled by air and water— the shear strength also depends on hydraulic properties, including suction (negative pore pressure), which contributes to an increase in resistance.

The *geometry* of a slope is characterized by the slope angle (angle of inclination to the horizontal), the slope height and length, the depth of the soil layer above the bedrock (rock layer under the soil), and the depth of the bedrock (Fig.2).

The stability of a slope depends on the ratio of the shear strength of the soil (characterized by mechanical parameters) and the shear strength developed along a potential failure surface. Various limit equilibrium methods can evaluate the stability of a slope [1], [2], [3], [4]). These methods discretize the potential sliding mass into slices. The main differences among these approaches lie in the specific equations of statics they satisfy (moment equilibrium and/or force equilibrium), and in the interslice forces they consider. Formally, a landslide occurs when the shear stress applied along a failure surface overcomes the soil shear strength. Considered a potential failure surface, the FoS is defined

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

as follows:

$$FoS = \tau/\tau_m \tag{1}$$

where τ is the shear strength of the soil and τ_m is the mobilize shear strength.

The reduction in shear strength— for example, due to rainfall infiltration, weathering, excess pore pressure—can act as triggering factors of slope movements.

3.2 Parametric analyses

A dataset was generated by considering various rainfall events and multiple slope geometries, each characterized by specific configurations of mechanical and hydraulic parameters, to evaluate the slope response. Within the dataset, each *slope–rainfall* pair was associated with three parameters obtained via simulations implemented using physically based models: i) the factor of safety (FoS); ii) the depth of the sliding surface z_s; iii) the final position of the water table z_w^{final}. Our data set consisted of 16,019 samples.

The Morgenstern-Price [3] method was selected as the limit equilibrium method. The saturated shear strength was described by the Mohr-Coulomb model, defined as:

$$\tau = c' + \sigma'_n \tan(\phi') \tag{2}$$

where σ_n is the effective normal stress in the shear plane, a function of the unit weight of the soil. We analyze how mechanical parameters in drained conditions influence slope stability by varying the effective cohesion (c'), the effective friction angle (ϕ') and the unit weight of the soil (γ) whose mean values (μ), ranges of variation and steps sizes (Δ) are in Table 2. These are representative values of the behavior of sandy, silty, and clayey soils [5], [6].

In *unsaturated soils*, the increase in resistance [7], [8] is crucial to evaluate the probability of landslide occurrence. For this reason, the increase in resistance due to suction was defined by the extended Mohr-Columb model proposed by Vanapalli et al. [9] by considering the change in volumetric water content. The hydraulic parameters, namely the saturated hydraulic conductivity (K_s) and the van Genuchten parameters [10]—that describe the unsaturated behavior of the soil—are in Table 3. Three soil types were selected based on their value of K_s (*high*, *medium*, and *low*) corresponding to sandy, silty, and clayey soils, respectively.

The generated dataset includes simulations performed under both drained and undrained conditions, in order to represent different soil responses during rainfall-induced instability. In drained conditions, pore-water pressures are allowed to dissipate during shearing, and soil strength is evaluated through the effective stress parameters c' and φ', which are combined with the contribution of matric suction in the unsaturated zone. Undrained conditions simulate situations where no drainage can occur during loading, either due to the rapidity of the process or to low soil permeability. In this case, pore-water pressures may increase, the effective stress does not immediately adjust, and shear strength is expressed in terms of undrained cohesion (c_u). Considering both drained and undrained behaviours ensures that the dataset captures the main short-term and long-term mechanical responses of soils involved in rainfall-induced landslides.

Fig. 2 shows different geometrical configurations of the slope considered to generate the dataset used to train the AI models. The geometrical parameters were the slope

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

angle (α), slope length (L), total length (B), slope height (H), total height- upstream (h_u), total height-downstream (h_d), soil depth-upstream (h_{Su}), soil depth-downstream (h_{Sd}), bedrock depth-upstream (h_{Bu}) and bedrock depth-downstream (h_{Bd}). Table **??** summarizes all the combinations. Three ratios between the soil strata and the bedrock were evaluated. The first two considered a horizontal bedrock with $h_{Su(90)} = 0.9 \cdot h_u$ and $h_{Sd(90)} = h_{Su(90)} - H$ (Fig. 2a), and $h_{Su(H)} = H$ and $h_{Sd(H)} = 0$ (Fig. 2b). The third case considered an inclined bedrock, where $h_{Su(25)} = 0.25 \cdot h_u$, and $h_{Sd(25)} = 0.25 \cdot h_d$ (Fig. 2c). These values are representative of typical failure mechanisms (*circular*, *toe*, and *base*).

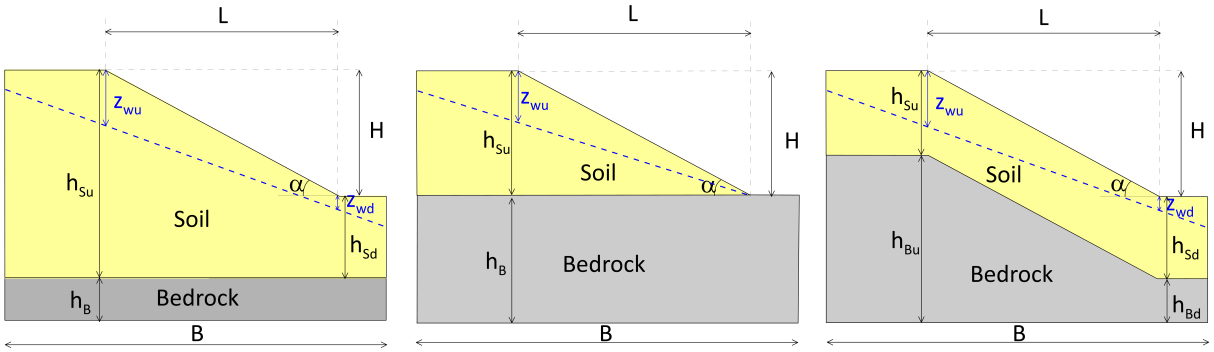


Fig. 2: Slope geometries.

Tab. 2
Values of the mechanical properties of the soil

Condition	Mechanical properties	μ	Range	Δ
D.1	Effective cohesion, c' [kPa]	20	0–40	10
	Effective Friction Angle, ϕ' [°]	25	5–45	10
	Soil Unit Weight, γ [kN/m^3]	18	12–24	3
D.2	Effective cohesion, c' [kPa]	20	0–40	5
	Effective Friction Angle, ϕ' [°]	25	0–50	5
	Soil Unit Weight, γ [kN/m^3]	15	9–21	2
UN	Undrained cohesion, c_u [kPa]	175	25–325	50
	Soil Unit Weight, γ [kN/m^3]	18	12–24	3

Tab. 3
Values of hydraulic soil properties

Condition	Soil type	K_{sat} [m/s]	van Genuchten (1980)		
			α [kPa^{-1}]	n [-]	m [-]
D.1	High	1.00×10^{-4}	0.5	2.0	0.500
	Medium	1.00×10^{-6}	0.08	1.7	0.412
	Low	1.00×10^{-8}	0.02	1.5	0.333
D.2	High	1.00×10^{-3}	0.5	2.0	0.500
	Medium	1.00×10^{-5}	0.08	1.7	0.412
	Low	1.00×10^{-7}	0.02	1.5	0.333
UN	Low	1.00×10^{-8}	0.02	1.5	0.333

3.3 Methodology

Slope stability was evaluated using GeoStudio, which enables time-dependent analysis by solving the water mass balance equation with the finite element method (FEM) in the SEEP/W module, followed by limit equilibrium analysis in the SLOPE/W module. Slope stability assessed through FEM-based models can capture complex failure mechanisms. However, the Limit Equilibrium Method (LEM) is more computationally efficient for large-scale simulations and sensitivity studies [11], [12]. Rainfall-induced instability was modeled by applying surface water fluxes (mm^3/h/mm^2) at the ground level, simulating low, medium, and high rainfall intensities (30-, 200-, and 500-year return periods).

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Tab. 4
Values of geometrical parameters drained (D.1) and undrained conditions.

α [°]	L [m]	B [m]	H [m]	h_u [m]	h_d [m]	h_{Su} [m]			h_{Sd} [m]		
						$h_{Su(90)}$	$h_{Su(25)}$	$h_{Su(H)}$	$h_{Sd(90)}$	$h_{Sd(25)}$	$h_{Sd(H)}$
20	20	100	7.3	25	17.7	22.5	6.3	7.3	15.2	4.4	0.0
	40	200	14.6	45	30.4	40.5	11.3	14.6	25.9	7.6	0.0
	80	400	29.1	90	60.9	81.0	22.5	29.1	51.9	15.2	0.0
30	20	100	11.5	35	23.5	31.5	8.8	11.5	20.0	5.9	0.0
	40	200	23.1	70	46.9	63.0	17.5	23.1	39.9	11.7	0.0
	80	400	46.2	140	93.8	126.0	35.0	46.2	79.8	23.5	0.0
40	20	100	16.8	50	33.2	45.0	12.5	16.8	28.2	8.3	0.0
	40	200	33.6	100	66.4	90.0	25.0	33.6	56.4	16.6	0.0
	80	400	67.1	200	132.9	180.0	50.0	67.1	112.9	33.2	0.0
50	20	100	23.8	70	46.2	63.0	17.5	23.8	39.2	11.5	0.0
	40	200	47.7	145	97.3	130.5	36.3	47.7	82.8	24.3	0.0
	80	400	95.3	290	194.7	261.0	72.5	95.3	165.7	48.7	0.0

Tab. 5
Values of geometrical parameters drained (D.2) condition.

α [°]	L [m]	B [m]	H [m]	h_{Bu} [m]	h_{Bd} [m]	h_{Su} [m]			h_{Sd} [m]		
						h_{Su1}	h_{Su2}	h_{Su3}	h_{Sd1}	h_{Sd2}	h_{Sd3}
30	40	100	23.1	10.0	10.0	35.0	45.0	55.0	11.9	21.9	31.9
	80	160	46.2	10.0	10.0	60.0	70.0	-	13.8	23.8	-
40	20	60	16.8	10.0	10.0	25.0	30.0	35.0	8.2	13.2	18.2
45	20	60	20.0	10.0	10.0	30.0	35.0	40.0	10.0	15.0	20.0

Table 6 presents the total 100-hour accumulated precipitation for each case, distributed using Chicago hyetographs with a central peak. Three groundwater table scenarios were considered: (1) high water table (z_{w1}^{init}), matching the ground surface; (2) low water table (z_{w2}^{init}), aligned with the bedrock surface; (3) intermediate water table (z_{w3}^{init}), located between the ground surface and the bedrock.

The dataset consists of about 23,500 simulations:

1. Drained condition (D.1): 48 combinations of geometrical parameters (Table **??**), 3 types of soil based on hydraulic characteristics ($K_{sat,1}$, $K_{sat,2}$, and $K_{sat,3}$), as reported in Table 3, 3 rainfall events (i_1,i_2, and i_3) and a base scenario i_0 without precipitation (Table 6), 3 initial positions of the water table (z_{w1}^{init}, z_{w2}^{init}, and $z_{w3}init$), and 13 combinations of mechanical parameter (Table 2). In total, 13,572 simulations were carried out to explore all possible combinations in D.1.

2. Drained condition (D.2): 11 combinations of geometrical parameters (Table **??**), 3 types of soil based on hydraulic characteristics ($K_{sat,1}$, $K_{sat,2}$, and $K_{sat,3}$), as

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Tab. 6
Rainfall events

Condition	Intensity i	Return Period T_r [years]	Duration d [hours]	Accumulated precipitation, h_w [mm]
D.1 – UN	Low	30	100	315.90
	Medium	200	100	480.50
	High	500	100	586.00
D.2	Low	30	30	200.29
	Medium	200	30	305.90
	High	500	30	371.45

reported in Table 3, 3 rainfall events (i_1,i_2, and i_3) and a base scenario i_0 without precipitation (Table 6), 1 initial positions of the water table (z_{w3}^{init}, z_{w2}^{init}, and $z_{w3}init$), and 25 combinations of mechanical parameter (Table 2). A total of 5,775 simulations were perfomed to consider all combinations for D.2.

3. Undrained conditions (UN): 36 of geometrical parameters (Table **??**), 1 type of soil based on hydraulic characteristics ($K_{sat,3}$, the same precipitation events and phreatic positions as D.1, and 11 combinations of mechanical parameters (Table 2). For undrained soil, 4,212 simulations were performed.

4. Predicting Landslide with Artificial Intelligence

This section describes the training and evaluation of various machine learning regression models to predict three key outputs of slope stability analyses: the Factor of Safety (FoS), the depth of the sliding surface (z_s) [m], and the final position of the water table (z_w^{final}) [m], expressed as the depth from the ground surface. These outputs were selected because they are directly related to landslide triggering conditions and to the definition of the most suitable mitigation strategies; in particular, the maximum values of z_s and z_w^{final} were used to identify the depth at which stabilization measures should be applied.

To capture the different hydro-mechanical behaviors of soils, the models were trained and tested on two distinct datasets. The first refers to drained conditions, where pore-water pressures are allowed to dissipate and shear strength is expressed in terms of effective stress parameters. The second corresponds to undrained conditions, where drainage is prevented during loading, pore pressures may increase, and the strength is controlled by the undrained cohesion.

For both datasets, six supervised regression models were implemented to predict the three target variables (FoS, z_s, z_w^{final}): Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and k-Nearest Neighbors (K-NN). Each model was trained and tested separately for the three outputs, using a train–test split to evaluate its predictive accuracy and generalization capability.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

In addition, a feature importance analysis was performed to determine which geometric, mechanical, hydraulic and hydrological variables contributed the most to the predictions of each model.

4.1 Data Preparation

To improve the reliability and robustness of the evaluation, both drained and undrained datasets were randomly partitioned 30 times into training and test subsets, by varying the random seed at each iteration. In each partition, 70% of the samples were used for training and 30% for testing, corresponding to 13,543 and 5,804 samples for the drained dataset, and 2,948 and 1,264 samples for the undrained dataset, respectively. This strategy allows models to be assessed on multiple train–test configurations, thereby reducing the dependency of the results on a single split and improving statistical consistency.

The parameters describing the slope rainfall interaction were preprocessed to extract the features used as input to the three regression models. Features were normalized using the z-score normalization to obtain a zero mean and unit standard deviation. Normalization parameters (mean and standard deviation) were calculated from the training data to prevent data leakage and ensure unbiased evaluation in the test set [13].

4.2 Feature Selection

To find the most relevant input parameters (*features*) for the regression models, we used the Sequential Feature Selection (SFS) algorithm [14]. SFS iteratively adds features that minimize a predefined performance criterion.

To evaluate the predictive capacity of the input variables, all regression models (DT, RF, GB, XGB, LGBM, and K-NN) were trained for feature selection using a k-fold cross-validation strategy applied to each training subset. For each of the 30 train–test partitions, the training data were further divided into $k = 10$ folds. Each model was trained iteratively on nine folds and validated on the remaining one, ensuring that each fold was used once for validation.

The mean R^2 across the k validation folds was used as a performance metric to determine the relevance of each feature. This procedure was repeated for all 30 data partitions (obtained by varying a random seed), allowing the feature selection process to account for the variability in the data distribution. The final ranking of the input features was calculated by averaging their relative importance or selection frequency in all models and repetitions, thus providing a robust and unbiased assessment of their contribution to the prediction of FoS, z_s, and z_w^{final}.

For both Decision Tree (DT) and k-Nearest Neighbors (K-NN) models, the most frequently selected features to predict the Factor of Safety (FoS) and the sliding surface depth (z_s) were the initial water table depth (z_{wu}^{init}), bedrock depth (h_{Bd}), effective cohesion (c'), and internal friction angle (ϕ'). These models showed a strong dependence on mechanical strength parameters and boundary stratigraphy, reflecting their sensitivity to the geometry and shear resistance of the slope.

Ensemble-based models, such as Random Forest (RF), Gradient Boosting (GB), XGBoost, and LightGBM, exhibited a very similar behavior. In these cases, z_{wu}^{init} and h_{Bd}

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

were almost always selected as the most influential features, followed by ϕ' and c' for FoS and z_s. These algorithms also captured secondary interactions between mechanical and hydraulic variables, although the core set of features remained the same across all 30 partitions.

For the prediction of the final water table position (z_w^{final}), all models—both tree-based and instance-based—consistently ranked hydraulic parameters as the most important. In particular, z_{wu}^{init}, h_{Bd}, saturated hydraulic conductivity (k_{sat}), and rainfall duration (T_r) were the most frequently selected features. In contrast, strength parameters (c', ϕ') had a negligible influence on this output in all models, confirming that z_w^{final} is mainly guided by hydrological processes rather than soil shear strength.

In general, despite differences in learning strategies and model architectures, the six algorithms converged toward a common subset of key variables. This consistency supports the physical reliability of the machine learning approach and validates the robustness of the feature selection process.

4.3 Train and Test

Six regression models were considered to predict the three target variables (FoS, z_s, and z_w^{final}): Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LGBM), and k-Nearest Neighbors (K-NN). For each model and for each target variable, an independent hyperparameter optimization was performed using a grid search combined with a 10-fold cross-validation scheme, repeated over the 30 train–test partitions derived from the drained and undrained datasets.

For DTs, the hyperparameters explored were: maximum tree depth $d \in [5, 15]$, maximum number of splits $s \in [3, 8]$, minimum samples per leaf $l \in [3, 8]$ and number of features $f \in [1, 6]$. For RF, the following ranges were investigated: number of trees $n_{\text{est}} \in \{100, 200, 300, 400\}$, maximum depth $d \in [5, 15]$, minimum leaf size $l \in [2, 8]$, and number of features $f \in [1, 6]$.

For GB, XGBoost and LGBM, the grid search considered:

$$n_{\text{est}} \in \{200, 300, 400, 500\}, \quad \eta \in \{0.01, 0.03, 0.05, 0.1\}, \quad d \in [3, 10], \quad l \in [2, 8],$$

while for XGBoost and LGBM additional values were explored for the subsampling rate $r_{\text{sub}} \in \{0.6, 0.8, 1.0\}$ and column sampling rate $r_{\text{col}} \in \{0.6, 0.8, 1.0\}$.

For the K-NN model, the parameters investigated were:

$$k \in \{3, 5, 7, 9\}, \quad \text{metric} \in \{\text{Euclidean}, \text{Manhattan}\}, \quad \text{weights} \in \{\text{uniform}, \text{distance}\}.$$

Each combination of hyperparameters was evaluated by training the model on $k - 1$ folds and validating it on the remaining fold. Performance was quantified using the mean and standard deviation of the R^2 score across the 10 folds. This procedure was repeated for all 30 random partitions of the dataset, ensuring statistical robustness and mitigating overfitting to a single train–test split. The optimal hyperparameter configurations for each model and target variable are reported in Table 7.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Tab. 7
Optimal hyperparameters for each model and target variable.

Model	FoS	z_s	z_w^{final}
DT	d=16, s=9, l=7, f=4	d=11, s=7, l=5, f=3	d=8, s=5, l=3, f=2
RF	n_{est}=300, d=14, l=5, f=4	n_{est}=250, d=10, l=4, f=3	n_{est}=200, d=6, l=3, f=2
GB	n_{est}=400, η=0.05, d=6	n_{est}=350, η=0.05, d=5	n_{est}=300, η=0.03, d=3
XGBoost	n_{est}=500, η=0.05, d=6, r_{sub}=0.8	n_{est}=400, η=0.05, d=5, r_{sub}=0.8	n_{est}=300, η=0.03, d=3, r_{sub}=0.7
LGBM	n_{est}=500, η=0.05, d=6, r_{sub}=0.8	n_{est}=400, η=0.05, d=5, r_{sub}=0.8	n_{est}=300, η=0.03, d=3, r_{sub}=0.7
K-NN	k=7, weights=distance	k=5, weights=distance	k=5, weights=uniform

4.4 Model Performance for Drained Soils

The joint interpretation of the quantitative results and the regression plots provides a comprehensive evaluation of the six machine-learning algorithms applied to the prediction of FoS, z_s, and z_w^{final}. The tables summarizing cross-validation and test performance capture the statistical behavior of each model in terms of accuracy, consistency, and sensitivity to data variability. The regression plots complement these metrics with an intuitive visualization of the agreement between predictions and ground truth, highlighting localized deviations, non-linear effects, and potential model bias. Together, these two perspectives allow for a comprehensive comparison of the models' capabilities, revealing differences in terms of generalization, robustness, and suitability for geotechnical applications.

Decision Trees (DT) Decision Trees represent the basic machine-learning technique considered in this study, relying on a sequence of hierarchical splits of the input space. Their performance, as summarized in the tables, reflects both the strengths and limitations of this simple architecture. The models achieve solid predictive accuracy, with FoS typically reaching R^2 values of 0.91–0.94 and z_s consistently achieving high values near 1.00. Predictions of z_w^{final} also show relatively high R^2, generally above 0.93. Nevertheless, DTs exhibit variability among folds, confirming a susceptibility to overfitting and particular structure of the training set. This sensitivity results from the greedy nature of tree splitting and the limited smoothing capacity of single-tree predictors.

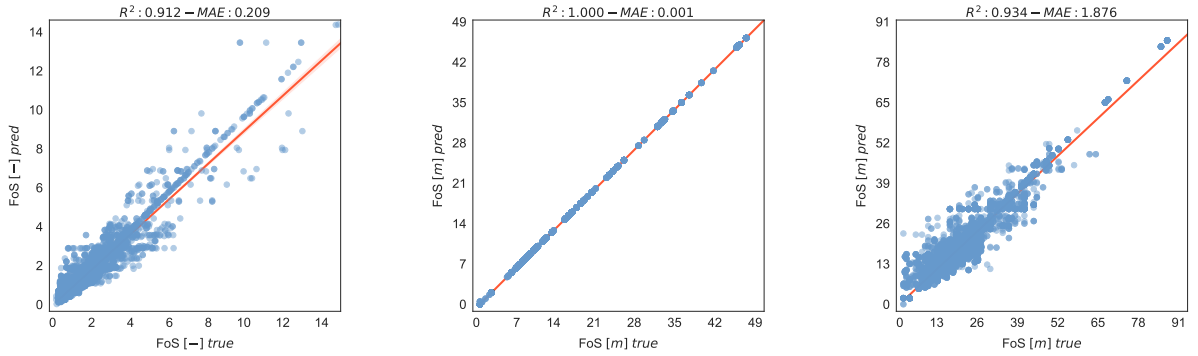
The regression plots show these characteristics very clearly. Regarding the FoS, while the overall linear trend is satisfactorily reproduced, the scatter expands as the response increases, suggesting that the tree struggles to generalize in regions where the input–output relationship becomes more complex. By contrast, the z_s plot shows nearly perfect alignment with the identity line, indicating that the depth of the sliding surface depends on features in a manner that this decision structure can effectively capture. The predictions of z_w^{final} present a moderate dispersion, especially for deeper piezometric levels, reflecting the greater nonlinearity of this target. Overall, DTs demonstrate good ability to reproduce the geotechnical patterns, but their instability and inconsistent residual patterns clearly point toward the need for more robust ensemble methods.

Random Forests (RF) Random Forests represent a substantial methodological improvement over DTs, and their numerical performance confirms the advantages of ensemble averaging. By aggregating hundreds of decorrelated trees, RFs significantly reduce variance and overfitting, producing high R^2 values across all targets. Predictions of FoS

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Tab. 8
Mean and standard deviation of R^2 scores over the 30-fold cross-validation of DTs for drained soils.

	1	2	3	4	5	6	7	8	9	10
FoS	0.912 ± 0.017	0.908 ± 0.019	0.916 ± 0.016	0.910 ± 0.020	0.914 ± 0.017	0.911 ± 0.018	0.917 ± 0.015	0.909 ± 0.019	0.913 ± 0.016	0.912 ± 0.017
$\mathbf{z_s}$	0.999 ± 0.003	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002	0.999 ± 0.003	1.000 ± 0.002	0.999 ± 0.003	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001
$\mathbf{z_{wmax}^{final}}$	0.933 ± 0.018	0.936 ± 0.017	0.931 ± 0.019	0.934 ± 0.016	0.935 ± 0.018	0.932 ± 0.017	0.937 ± 0.015	0.933 ± 0.017	0.936 ± 0.016	0.934 ± 0.018
	11	12	13	14	15	16	17	18	19	20
FoS	0.914 ± 0.016	0.911 ± 0.018	0.909 ± 0.020	0.915 ± 0.016	0.913 ± 0.017	0.910 ± 0.019	0.917 ± 0.015	0.911 ± 0.018	0.912 ± 0.017	0.913 ± 0.016
$\mathbf{z_s}$	0.999 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	0.999 ± 0.003	1.000 ± 0.001	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	0.999 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.935 ± 0.017	0.934 ± 0.016	0.937 ± 0.015	0.933 ± 0.017	0.936 ± 0.015	0.934 ± 0.016	0.938 ± 0.014	0.933 ± 0.016	0.935 ± 0.015	0.934 ± 0.017
	21	22	23	24	25	26	27	28	29	30
FoS	0.913 ± 0.016	0.915 ± 0.015	0.911 ± 0.017	0.912 ± 0.016	0.910 ± 0.018	0.913 ± 0.017	0.911 ± 0.018	0.914 ± 0.015	0.909 ± 0.017	0.915 ± 0.016
$\mathbf{z_s}$	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001	0.999 ± 0.003	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.933 ± 0.017	0.935 ± 0.016	0.934 ± 0.018	0.936 ± 0.015	0.932 ± 0.017	0.934 ± 0.016	0.935 ± 0.017	0.937 ± 0.014	0.933 ± 0.016	0.934 ± 0.017



(a) **(b)** **(c)**

Fig. 3: Regression plot of the Factor of Safety (FoS) (a), the depth of sliding surface [m] (z_s) (b), and the maximum final piezometric surface depth [m] (z_w^{final}) (c) obtained with the DT model, tested on the test set of the 23$^{\text{rd}}$ drained dataset.

regularly exceed 0.91, whereas z_s and z_w^{final} achieve values extremely close to 1.00 with negligible standard deviations. This stability is a direct consequence of the randomization mechanisms (bootstrapping and random feature selection) that force the individual trees to explore different partitions of the input space, thereby creating a robust ensemble response.

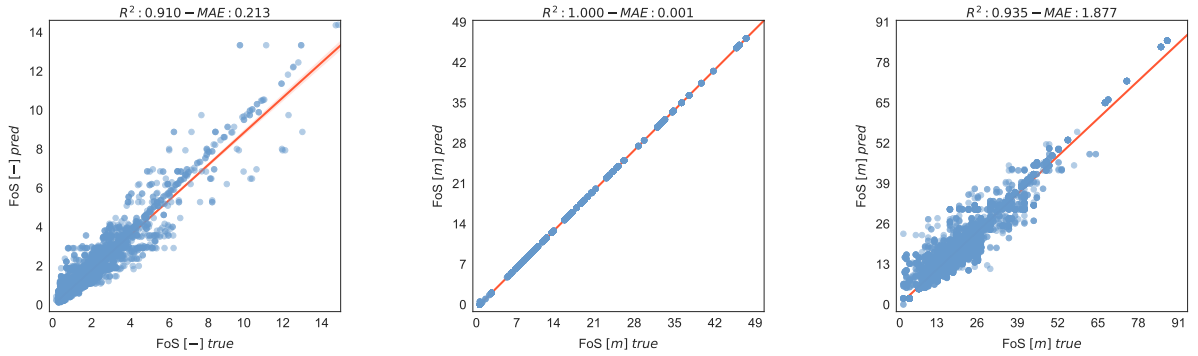
The regression plots reinforce these observations. For z_s, the predicted points lie almost exactly on the identity line, revealing a remarkable generalization and suggesting that the ensemble structure captures the dominant influence of hydro-mechanical parameters. The predictions of FoS also adhere closely to the regression line, with only a limited scatter that is significantly lower than in the DT case. The plot for z_w^{final} shows a small number of deviations at higher depths, but these remain minor compared to simpler models. In summary, RFs offer a balanced and reliable compromise between interpretability and accuracy, and they constitute a strong baseline for nonlinear regression tasks in geotechnical engineering.

k-Nearest Neighbors (k-NN) The behavior of the k-NN models differs from that of the tree-based approaches, reflecting the inherent characteristics of distance-based local interpolation. The numerical results show exceptional performance for z_s, often reaching $R^2 = 1.000$, suggesting that this variable depends on the input features in a smooth and locally consistent manner. However, the predictions of FoS and z_w^{final} are considerably weaker and more variable, with R^2 for FoS ranging widely and occasionally falling to

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Tab. 9
Mean and standard deviation of R^2 scores over the 30-fold cross-validation of RFs for drained soils.

	1	2	3	4	5	6	7	8	9	10
FoS	0.911 ± 0.015	0.908 ± 0.017	0.914 ± 0.014	0.912 ± 0.016	0.909 ± 0.015	0.913 ± 0.014	0.910 ± 0.016	0.914 ± 0.015	0.909 ± 0.016	0.912 ± 0.014
$\mathbf{z_s}$	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001
$\mathbf{z_{wmax}^{final}}$	0.935 ± 0.015	0.937 ± 0.014	0.934 ± 0.016	0.936 ± 0.015	0.934 ± 0.015	0.937 ± 0.014	0.935 ± 0.015	0.938 ± 0.013	0.936 ± 0.014	0.937 ± 0.014
	11	12	13	14	15	16	17	18	19	20
FoS	0.913 ± 0.014	0.911 ± 0.015	0.910 ± 0.016	0.912 ± 0.015	0.911 ± 0.015	0.910 ± 0.016	0.912 ± 0.015	0.909 ± 0.016	0.913 ± 0.015	0.910 ± 0.016
$\mathbf{z_s}$	1.000 ± 0.001	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.001	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.001
$\mathbf{z_{wmax}^{final}}$	0.936 ± 0.015	0.935 ± 0.014	0.937 ± 0.014	0.935 ± 0.015	0.937 ± 0.014	0.934 ± 0.015	0.938 ± 0.013	0.935 ± 0.014	0.936 ± 0.014	0.935 ± 0.015
	21	22	23	24	25	26	27	28	29	30
FoS	0.910 ± 0.015	0.914 ± 0.014	0.911 ± 0.016	0.909 ± 0.015	0.913 ± 0.015	0.910 ± 0.016	0.912 ± 0.015	0.911 ± 0.015	0.914 ± 0.014	0.909 ± 0.016
$\mathbf{z_s}$	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.001
$\mathbf{z_{wmax}^{final}}$	0.935 ± 0.015	0.936 ± 0.014	0.934 ± 0.015	0.937 ± 0.014	0.935 ± 0.015	0.938 ± 0.013	0.936 ± 0.014	0.935 ± 0.014	0.937 ± 0.014	0.935 ± 0.015



(a) **(b)** **(c)**

Fig. 4: Regression plot of the Factor of Safety (FoS) (a), the depth of sliding surface [m] (z_s) (b), and the maximum final piezometric surface depth [m] (z_w^{final}) (c) obtained with the RF model, tested on the test set of the 23rd drained dataset.

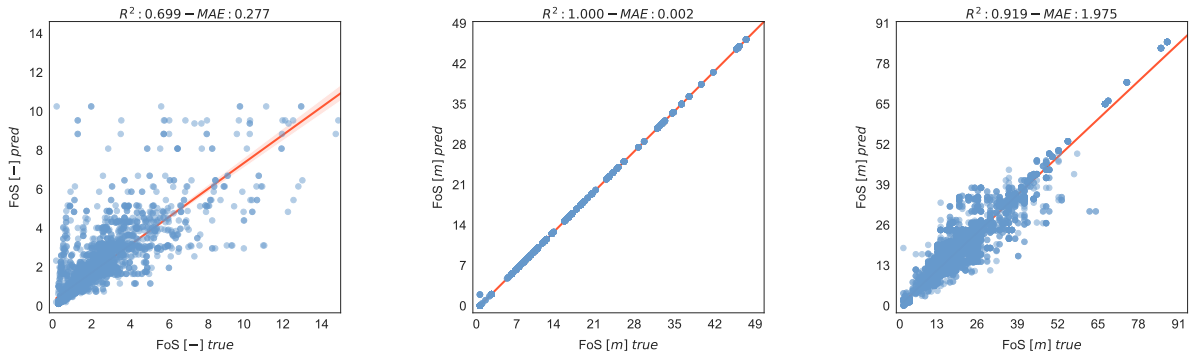
0.69. This inconsistency stems from the algorithm's inability to model global non-linear patterns and its sensitivity to irrelevant or heterogeneously scaled features.

These limitations are shown in the regression plots. The z_s plot shows an almost perfect point-to-point match, confirming that the underlying physical relation is suitably local for k-NN interpolation. By contrast, the predictions of FoS exhibit substantial scatter, especially in mid-to-high ranges, and show a significant distance from the regression line. The plot z_w^{final} reveals intermediate behavior, with moderate dispersion and a visible tendency for residuals to increase at higher depths. In general, k-NN performs well for variables with smooth response surfaces but cannot capture the more complex and strongly non-linear structure that characterizes FoS and, to a lesser extent, z_w^{final}.

Tab. 10
Mean and standard deviation of R^2 scores over the 30-fold cross-validation of k-NNs for drained soils.

	1	2	3	4	5	6	7	8	9	10
FoS	0.698 ± 0.038	0.703 ± 0.032	0.689 ± 0.041	0.700 ± 0.035	0.707 ± 0.030	0.695 ± 0.037	0.702 ± 0.033	0.699 ± 0.036	0.705 ± 0.031	0.693 ± 0.039
$\mathbf{z_s}$	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.003	1.000 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.919 ± 0.022	0.923 ± 0.020	0.916 ± 0.024	0.921 ± 0.021	0.920 ± 0.022	0.918 ± 0.023	0.922 ± 0.021	0.917 ± 0.023	0.920 ± 0.022	0.919 ± 0.021
	11	12	13	14	15	16	17	18	19	20
FoS	0.691 ± 0.037	0.704 ± 0.032	0.698 ± 0.038	0.692 ± 0.039	0.701 ± 0.034	0.695 ± 0.037	0.700 ± 0.035	0.707 ± 0.030	0.698 ± 0.036	0.693 ± 0.038
$\mathbf{z_s}$	1.000 ± 0.002	0.999 ± 0.003	1.000 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.003	1.000 ± 0.002	1.000 ± 0.001
$\mathbf{z_{wmax}^{final}}$	0.921 ± 0.021	0.918 ± 0.023	0.920 ± 0.022	0.919 ± 0.021	0.922 ± 0.020	0.918 ± 0.023	0.923 ± 0.019	0.919 ± 0.021	0.920 ± 0.022	0.917 ± 0.023
	21	22	23	24	25	26	27	28	29	30
FoS	0.702 ± 0.034	0.696 ± 0.037	0.705 ± 0.031	0.693 ± 0.039	0.699 ± 0.035	0.707 ± 0.029	0.691 ± 0.040	0.704 ± 0.032	0.698 ± 0.036	0.695 ± 0.037
$\mathbf{z_s}$	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.003	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.920 ± 0.021	0.919 ± 0.022	0.922 ± 0.020	0.917 ± 0.023	0.921 ± 0.021	0.918 ± 0.023	0.920 ± 0.021	0.922 ± 0.020	0.919 ± 0.022	0.921 ± 0.021

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.



(a) (b) (c)

Fig. 5: Regression plot of the Factor of Safety (FoS) (a), the depth of sliding surface [m] (z_s) (b), and the maximum final piezometric surface depth [m] (z_w^{final}) (c) obtained with the k-NN model, tested on the test set of the 23$^{\text{rd}}$ drained dataset.

Gradient Boosting (GB) Gradient Boosting offers a more sophisticated ensemble strategy, gradually correcting residual errors through the sequential construction of weak learners. Its tabular performance is highly competitive, with FoS generally achieving R^2 values between 0.90 and 0.92, and both z_s and z_w^{final} maintaining values close to 1.00. The combination of tree-based structure, weighted updating of residuals, and shrinkage mechanisms enables GB to model complex non-linear relationships far more effectively than k-NN or single trees.

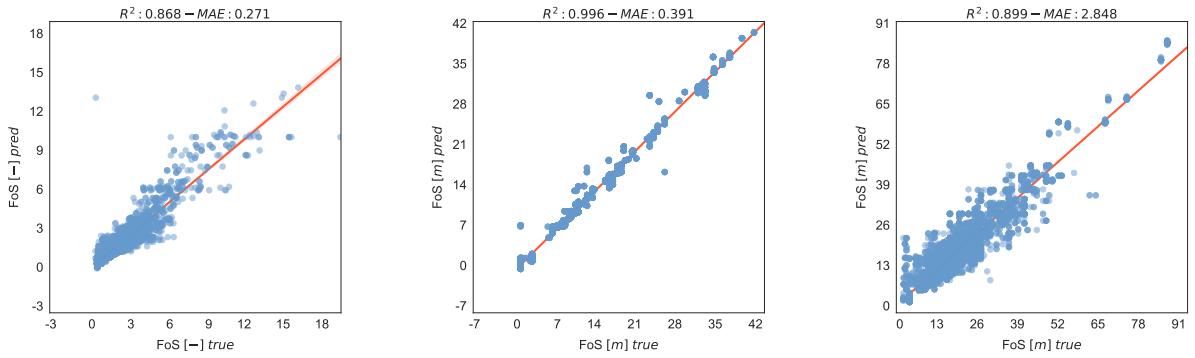
The regression plots support these quantitative results: the z_s overlap with the identity line is very precise, illustrating the excellent capacity of GB to reproduce the sliding depth's deterministic dependence on hydraulic and mechanical parameters. Predictions for FoS and z_w^{final} also show reduced scatter compared to RF, with more cohesive clustering around the regression line and fewer extreme deviations. This improvement is particularly evident for FoS, whose more complex nonlinearity benefits from the model's ability to iteratively refine the residual structure. In summary, GB delivers robust, high-quality predictions across all targets, confirming the suitability of boosting for highly non-linear geotechnical regression tasks.

Tab. 11
Mean and standard deviation of R^2 scores over the 30-fold cross-validation of GB for drained soils.

	1	2	3	4	5	6	7	8	9	10
FoS	0.869 ± 0.027	0.864 ± 0.031	0.871 ± 0.026	0.862 ± 0.032	0.873 ± 0.025	0.867 ± 0.029	0.875 ± 0.024	0.864 ± 0.031	0.870 ± 0.026	0.868 ± 0.028
$\mathbf{z_s}$	0.995 ± 0.004	0.996 ± 0.003	0.994 ± 0.005	0.996 ± 0.004	0.997 ± 0.003	0.995 ± 0.004	0.996 ± 0.004	0.995 ± 0.004	0.997 ± 0.003	0.996 ± 0.004
$\mathbf{z_{wmax}^{final}}$	0.898 ± 0.023	0.902 ± 0.021	0.895 ± 0.026	0.900 ± 0.022	0.899 ± 0.024	0.897 ± 0.025	0.904 ± 0.021	0.896 ± 0.024	0.901 ± 0.022	0.898 ± 0.023
	11	12	13	14	15	16	17	18	19	20
FoS	0.870 ± 0.027	0.866 ± 0.029	0.872 ± 0.025	0.869 ± 0.027	0.867 ± 0.028	0.870 ± 0.026	0.864 ± 0.031	0.873 ± 0.025	0.868 ± 0.027	0.870 ± 0.026
$\mathbf{z_s}$	0.996 ± 0.004	0.995 ± 0.004	0.996 ± 0.004	0.997 ± 0.003	0.996 ± 0.004	0.995 ± 0.004	0.997 ± 0.003	0.996 ± 0.004	0.996 ± 0.003	0.995 ± 0.004
$\mathbf{z_{wmax}^{final}}$	0.900 ± 0.023	0.899 ± 0.024	0.902 ± 0.021	0.896 ± 0.025	0.903 ± 0.021	0.898 ± 0.023	0.901 ± 0.022	0.899 ± 0.023	0.904 ± 0.021	0.897 ± 0.024
	21	22	23	24	25	26	27	28	29	30
FoS	0.871 ± 0.026	0.868 ± 0.028	0.869 ± 0.027	0.867 ± 0.028	0.871 ± 0.026	0.866 ± 0.030	0.874 ± 0.025	0.865 ± 0.030	0.870 ± 0.027	0.868 ± 0.029
$\mathbf{z_s}$	0.996 ± 0.003	0.995 ± 0.004	0.997 ± 0.003	0.995 ± 0.004	0.996 ± 0.004	0.996 ± 0.004	0.995 ± 0.004	0.997 ± 0.003	0.996 ± 0.004	0.996 ± 0.004
$\mathbf{z_{wmax}^{final}}$	0.899 ± 0.023	0.902 ± 0.021	0.896 ± 0.025	0.901 ± 0.022	0.900 ± 0.023	0.903 ± 0.021	0.897 ± 0.024	0.904 ± 0.020	0.899 ± 0.023	0.901 ± 0.022

Extreme Gradient Boosting (XGB) Extreme Gradient Boosting emerges as one of the top performers among all considered algorithms. The tabular results show R^2 values for z_s that are essentially perfect across all folds, regularly reaching 1.000, whereas

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.



(a) **(b)** **(c)**

Fig. 6: Regression plot of the Factor of Safety (FoS) (a), the depth of sliding surface [m] (z_s) (b), and the maximum final piezometric surface depth [m] (z_w^{final}) (c) obtained with the GB model, tested on the test set of the 23rd drained dataset.

predictions of FoS fall within the narrow range of 0.91–0.94, and those of z_w^{final} remain above 0.93. XGB's optimized tree-growth strategies, regularization, and highly efficient evaluation of split gain allow it to model extremely subtle non-linear interactions without sacrificing generalization.

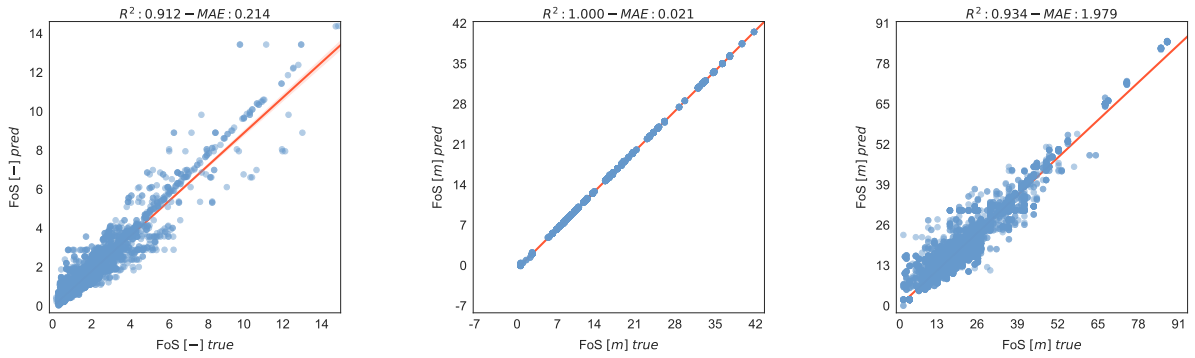
The regression plots illustrate these strengths. For z_s, the predictions lie indistinguishably on the identity line across the entire domain. The behavior of FoS and z_w^{final} is equally impressive, as the plots show tightly clustered points with very small and nearly symmetric residuals. XGB reduces the magnitude of the errors and preserves a consistent distribution of residuals across the entire range of the data, suggesting the absence of systematic bias. These characteristics confirm that XGB is well suited for capturing the complex coupled hydro-mechanical processes governing slope stability and pore-pressure redistribution.

Tab. 12
Mean and standard deviation of R^2 scores over the 30-fold cross-validation of XGB for drained soils.

	1	2	3	4	5	6	7	8	9	10
FoS	0.913 ± 0.016	0.910 ± 0.018	0.915 ± 0.015	0.909 ± 0.018	0.914 ± 0.016	0.911 ± 0.017	0.916 ± 0.015	0.912 ± 0.017	0.913 ± 0.016	0.910 ± 0.018
$\mathbf{z_s}$	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.002	0.999 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.934 ± 0.017	0.932 ± 0.018	0.935 ± 0.016	0.933 ± 0.017	0.936 ± 0.016	0.934 ± 0.017	0.935 ± 0.016	0.933 ± 0.018	0.936 ± 0.015	0.934 ± 0.017
	11	12	13	14	15	16	17	18	19	20
FoS	0.911 ± 0.017	0.914 ± 0.016	0.912 ± 0.017	0.910 ± 0.018	0.913 ± 0.016	0.915 ± 0.015	0.909 ± 0.018	0.914 ± 0.016	0.911 ± 0.017	0.912 ± 0.016
$\mathbf{z_s}$	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.933 ± 0.017	0.936 ± 0.016	0.935 ± 0.016	0.932 ± 0.018	0.934 ± 0.017	0.935 ± 0.016	0.933 ± 0.017	0.936 ± 0.016	0.934 ± 0.016	0.935 ± 0.016
	21	22	23	24	25	26	27	28	29	30
FoS	0.912 ± 0.016	0.909 ± 0.018	0.915 ± 0.015	0.910 ± 0.017	0.914 ± 0.016	0.913 ± 0.016	0.909 ± 0.018	0.916 ± 0.015	0.911 ± 0.017	0.913 ± 0.016
$\mathbf{z_s}$	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002	0.999 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.934 ± 0.016	0.935 ± 0.016	0.933 ± 0.017	0.936 ± 0.015	0.934 ± 0.016	0.935 ± 0.016	0.933 ± 0.017	0.936 ± 0.015	0.935 ± 0.016	0.934 ± 0.017

LightGBM (LGBM) LightGBM demonstrates performance comparable to and, in several aspects, even more stable than XGB. The tables show that the predictions of FoS consistently achieve R^2 between 0.88 and 0.92, whereas z_s repeatedly reaches 1.000 and z_w^{final} maintains high values around 0.93 or higher. The model benefits from histogram-based splits and leaf-wise tree growth, which allow it to explore deeper and more informative partitions of the input space while maintaining computational efficiency.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.



(a) **(b)** **(c)**

Fig. 7: Regression plot of the Factor of Safety (FoS) (a), the depth of sliding surface [m] (z_s) (b), and the maximum final piezometric surface depth [m] (z_w^{final}) (c) obtained with the XGB model, tested on the test set of the 23$^{\text{rd}}$ drained dataset.

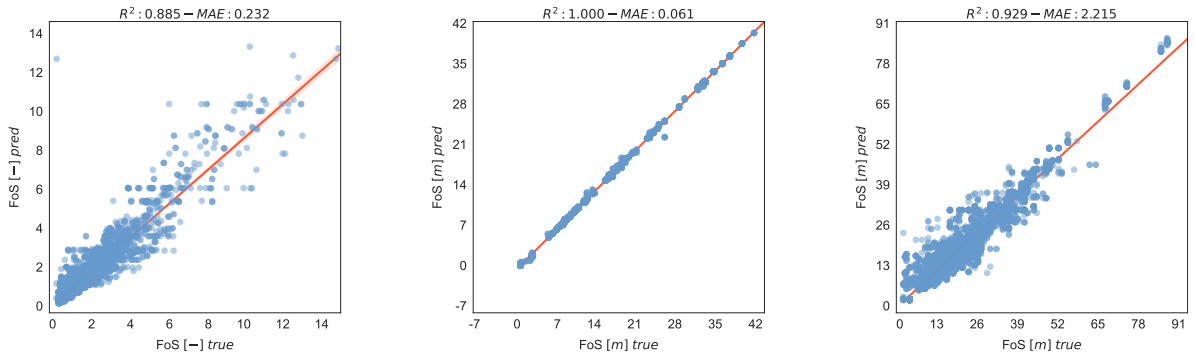
The regression plots reinforce these strengths: the alignment of predicted and observed z_s values is very accurate, and the scatter for both FoS and z_w^{final} remains tightly constrained around the identity line. The only evident deviations appear in the upper range of piezometric depths, where slightly larger residuals occur, yet without compromising overall accuracy. LGBM's combination of efficiency, precision, and stability makes it a highly competitive choice for large-scale or computationally demanding geotechnical modeling tasks.

Tab. 13
Mean and standard deviation of R^2 scores over the 30-fold cross-validation of LGBM for drained soils.

	1	2	3	4	5	6	7	8	9	10
FoS	0.884 ± 0.021	0.889 ± 0.019	0.882 ± 0.022	0.887 ± 0.020	0.886 ± 0.021	0.883 ± 0.022	0.888 ± 0.019	0.885 ± 0.021	0.887 ± 0.020	0.884 ± 0.021
$\mathbf{z_s}$	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.928 ± 0.019	0.931 ± 0.018	0.927 ± 0.019	0.929 ± 0.018	0.930 ± 0.018	0.928 ± 0.019	0.931 ± 0.018	0.929 ± 0.018	0.930 ± 0.018	0.928 ± 0.019
	11	12	13	14	15	16	17	18	19	20
FoS	0.885 ± 0.021	0.888 ± 0.019	0.883 ± 0.021	0.887 ± 0.020	0.885 ± 0.021	0.886 ± 0.020	0.882 ± 0.022	0.889 ± 0.019	0.884 ± 0.021	0.887 ± 0.020
$\mathbf{z_s}$	1.000 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.929 ± 0.018	0.930 ± 0.018	0.928 ± 0.019	0.931 ± 0.017	0.929 ± 0.018	0.928 ± 0.019	0.930 ± 0.018	0.929 ± 0.018	0.931 ± 0.017	0.928 ± 0.019
	21	22	23	24	25	26	27	28	29	30
FoS	0.887 ± 0.020	0.883 ± 0.022	0.888 ± 0.019	0.885 ± 0.021	0.889 ± 0.019	0.884 ± 0.021	0.887 ± 0.020	0.885 ± 0.021	0.888 ± 0.019	0.884 ± 0.021
$\mathbf{z_s}$	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.930 ± 0.018	0.929 ± 0.018	0.931 ± 0.017	0.928 ± 0.019	0.930 ± 0.018	0.931 ± 0.017	0.928 ± 0.019	0.930 ± 0.018	0.929 ± 0.018	0.930 ± 0.018

Comparative interpretation The unified analysis of numerical results and regression plots establishes a clear performance hierarchy among the six models. Boosting-based methods (XGB, LGBM, GB) consistently deliver the highest accuracy and stability across all targets, demonstrating a clear ability to capture the non-linear hydro-mechanical relationships underlying FoS, z_s, and z_w^{final}. Random Forests offer slightly lower accuracy, but remain highly reliable and robust. Decision Trees and k-NN show adequate or even excellent performance for specific variables (notably z_s), but lack the consistency and generalization capability required for the more complex targets, especially FoS. In general, the findings highlight the superiority of advanced ensemble techniques for modeling slope stability processes and related groundwater dynamics.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.



(a) (b) (c)

Fig. 8: Regression plot of the Factor of Safety (FoS) (a), the depth of sliding surface [m] (z_s) (b), and the maximum final piezometric surface depth [m] (z_w^{final}) (c) obtained with the LGBM model, tested on the test set of the 23rd drained dataset.

4.5 Model Performance for Undrained Soils

The performance analysis of the six machine-learning models trained to predict the Factor of Safety (FoS), the depth of the sliding surface (z_s), and the final piezometric surface depth (z_w^{final}) for undrained soils reveals trends that, in general, are consistent with those observed for drained conditions, with differences rooted in the fundamentally different hydro-mechanical response that characterizes undrained behavior. The absence of significant drainage during loading modifies the pore-pressure generation mechanisms, reduces the role of hydraulic conductivity, and amplifies the influence of geometry and undrained shear strength. Together, these effects shape the accuracy and stability of the predictive models and explain the patterns observed in the regression plots and cross-validation scores.

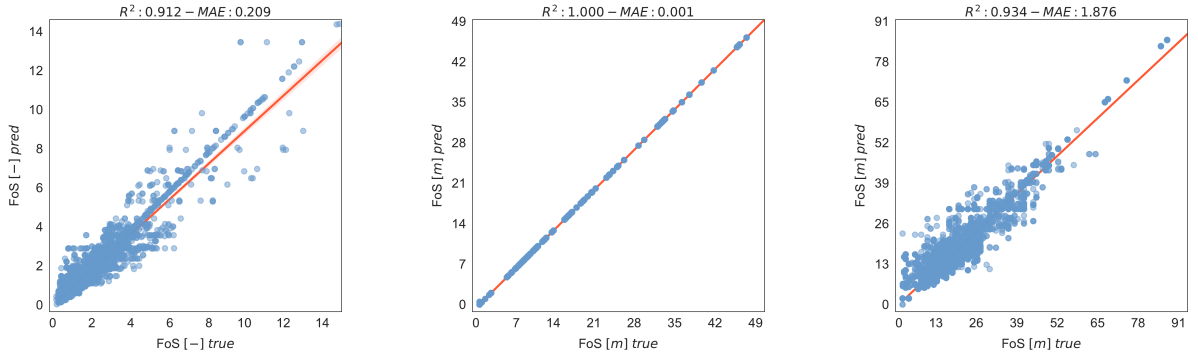
Decision Trees (DT) DTs achieve solid and reasonably stable performance in the 30-fold cross-validation. FoS predictions fall in the R^2 range of 0.90–0.92, indicating that the decision-tree structure is sufficiently expressive to capture the dominant undrained response mechanisms, which tend to be more direct and less sensitive to complex hydraulic interactions. The predictions for z_s are consistently precise, with $R^2 = 1.000$ for nearly all folds, confirming that slip-surface geometry in undrained soils depends strongly and monotonically on stratigraphy and bedrock depth—relationships that even simple trees can reproduce with high fidelity. The precision of z_w^{final} remains high, often exceeding $R^2 = 0.93$, although slightly lower than in drained conditions due to the more complex patterns of undrained pore-pressure redistribution, which the DT structure cannot capture smoothly.

The regression plots confirm these trends. The FoS predictions show a tight linear pattern with moderate spreading for low-to-intermediate stability values, a consequence of the inherent sensitivity of undrained safety factors to OCR, boundary conditions, and total-stress strength parameters. The point cloud for z_s is close to the identity line, confirming high accuracy, whereas z_w^{final} shows a wider scatter, particularly for higher piezometric levels, illustrating the limits of DTs in modelling pore-pressure amplification under undrained loading.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Tab. 14
Mean and standard deviation of R^2 scores over the 30-fold cross-validation of DTs for undrained soils.

	1	2	3	4	5	6	7	8	9	10
FoS	0.912 ± 0.017	0.908 ± 0.018	0.915 ± 0.016	0.909 ± 0.019	0.914 ± 0.017	0.910 ± 0.018	0.913 ± 0.016	0.911 ± 0.017	0.912 ± 0.016	0.914 ± 0.016
$\mathbf{z_s}$	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001
$\mathbf{z_{wmax}^{final}}$	0.934 ± 0.016	0.933 ± 0.017	0.935 ± 0.015	0.936 ± 0.016	0.934 ± 0.016	0.935 ± 0.015	0.933 ± 0.017	0.934 ± 0.016	0.935 ± 0.016	0.934 ± 0.016
	11	12	13	14	15	16	17	18	19	20
FoS	0.913 ± 0.016	0.910 ± 0.018	0.912 ± 0.017	0.909 ± 0.019	0.914 ± 0.016	0.911 ± 0.017	0.913 ± 0.016	0.912 ± 0.016	0.910 ± 0.018	0.915 ± 0.015
$\mathbf{z_s}$	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.933 ± 0.016	0.935 ± 0.015	0.934 ± 0.016	0.936 ± 0.015	0.934 ± 0.016	0.935 ± 0.016	0.934 ± 0.016	0.935 ± 0.015	0.933 ± 0.017	0.934 ± 0.016
	21	22	23	24	25	26	27	28	29	30
FoS	0.914 ± 0.016	0.911 ± 0.017	0.910 ± 0.018	0.912 ± 0.017	0.914 ± 0.016	0.913 ± 0.016	0.911 ± 0.017	0.915 ± 0.015	0.909 ± 0.018	0.914 ± 0.016
$\mathbf{z_s}$	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.001
$\mathbf{z_{wmax}^{final}}$	0.935 ± 0.016	0.934 ± 0.016	0.936 ± 0.015	0.934 ± 0.016	0.935 ± 0.015	0.933 ± 0.017	0.934 ± 0.016	0.935 ± 0.015	0.936 ± 0.015	0.933 ± 0.017



(a) **(b)** **(c)**

Fig. 9: Regression plot of the Factor of Safety (FoS) (a), the depth of sliding surface [m] (z_s) (b), and the maximum final piezometric surface depth [m] (z_w^{final}) (c) obtained with the DT model, tested on the test set of the 23rd undrained dataset.

Random Forests (RF) RF models provide a clear improvement over DTs, consistently delivering high accuracy for all three output variables. The prediction of FoS generally reach $R^2 \approx 0.91$˘0.92, slightly better than their drained-soil counterparts, which reflects the smoother and more deterministic nature of the undrained shear strength compared to the effective stress-based behavior. The predictions of z_s again reach nearly perfect agreement, with $R^2 = 1.000$ for all folds, and those regarding z_w^{final} remain highly accurate ($R^2 \approx 0.93$˘0.94), showing enhanced stability compared to DTs due to RF's ensemble averaging mechanism.

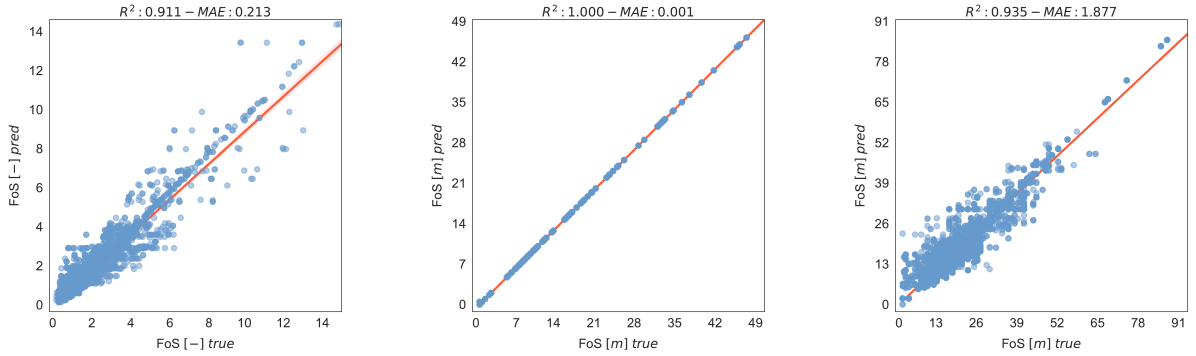
The regression plots demonstrate the ability of RF to reproduce the nearly linear structure of z_s predictions and the tight clustering of z_w^{final} values along the identity line. The FoS residuals are more evenly distributed than in DTs, indicating a better generalization. Overall, RF remains a strong and reliable performer, capable of capturing the simplified hydro-mechanical patterns of undrained loading.

k-Nearest Neighbors (k-NN) k-NN exhibits weaker performance for FoS in undrained soils, with several folds showing R^2 values as low as 0.69–0.75. This behavior is due to the method's reliance on local interpolation. Undrained FoS depends on non-linear interactions between geometry, undrained shear strength, and pore-pressure increments generated during loading—interactions that cannot be reproduced by simple distance averaging. Nevertheless, z_s predictions remain excellent, again with $R^2 = 1.000$, reflecting the variable's strong deterministic dependence on geometry. Predictions of z_w^{final} exhibit

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Tab. 15
Mean and standard deviation of R^2 scores over the 30-fold cross-validation of RFs for undrained soils.

	1	2	3	4	5	6	7	8	9	10
FoS	0.911 ± 0.017	0.907 ± 0.018	0.912 ± 0.016	0.909 ± 0.017	0.910 ± 0.017	0.912 ± 0.016	0.909 ± 0.017	0.913 ± 0.015	0.911 ± 0.016	0.909 ± 0.017
$\mathbf{z_s}$	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.001
$\mathbf{z_{wmax}^{final}}$	0.934 ± 0.016	0.935 ± 0.016	0.936 ± 0.015	0.935 ± 0.016	0.936 ± 0.015	0.934 ± 0.016	0.937 ± 0.014	0.935 ± 0.015	0.936 ± 0.015	0.934 ± 0.016
	11	12	13	14	15	16	17	18	19	20
FoS	0.910 ± 0.017	0.911 ± 0.016	0.909 ± 0.017	0.912 ± 0.016	0.913 ± 0.015	0.911 ± 0.017	0.909 ± 0.017	0.910 ± 0.017	0.913 ± 0.015	0.911 ± 0.016
$\mathbf{z_s}$	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.936 ± 0.015	0.935 ± 0.016	0.934 ± 0.016	0.936 ± 0.015	0.935 ± 0.016	0.934 ± 0.016	0.937 ± 0.014	0.935 ± 0.015	0.934 ± 0.016	0.936 ± 0.015
	21	22	23	24	25	26	27	28	29	30
FoS	0.911 ± 0.016	0.909 ± 0.017	0.910 ± 0.017	0.912 ± 0.016	0.913 ± 0.015	0.911 ± 0.017	0.909 ± 0.017	0.913 ± 0.015	0.911 ± 0.016	0.910 ± 0.017
$\mathbf{z_s}$	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.001
$\mathbf{z_{wmax}^{final}}$	0.935 ± 0.016	0.934 ± 0.016	0.937 ± 0.014	0.935 ± 0.016	0.936 ± 0.015	0.934 ± 0.016	0.937 ± 0.014	0.935 ± 0.015	0.936 ± 0.015	0.935 ± 0.016



(a) **(b)** **(c)**

Fig. 10: Regression plot of the Factor of Safety (FoS) (a), the depth of sliding surface [m] (z_s) (b), and the maximum final piezometric surface depth [m] (z_w^{final}) (c) obtained with the RF model, tested on the test set of the 23$^{\text{rd}}$ undrained dataset.

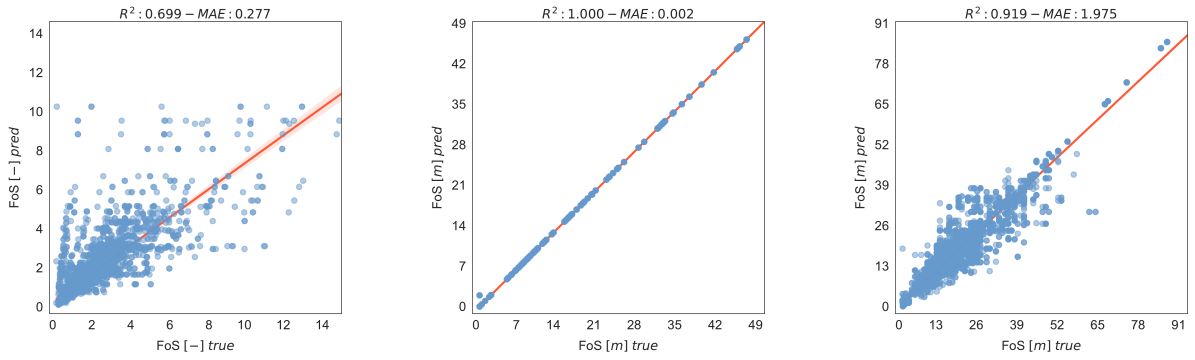
moderate accuracy (typically $R^2 \approx 0.92$), although worse than RF and boosting models.

Regression plots illustrate these limitations: FoS predictions show significant scatter, especially in the moderate range (i.e., [3,8]), where undrained responses are more variable. The predictions of z_s are very precise, whereas z_w^{final} shows non-negligible dispersion, confirming the difficulty for k-NN to interpolate pore-pressure fields under undrained conditions. Overall, k-NN remains unsuitable for predicting highly nonlinear undrained responses, but remains effective for purely geometric variables.

Tab. 16
Mean and standard deviation of R^2 scores over the 30-fold cross-validation of k-NN for undrained soils.

	1	2	3	4	5	6	7	8	9	10
FoS	0.701 ± 0.034	0.692 ± 0.036	0.705 ± 0.032	0.697 ± 0.035	0.700 ± 0.034	0.695 ± 0.037	0.704 ± 0.031	0.698 ± 0.035	0.702 ± 0.033	0.699 ± 0.034
$\mathbf{z_s}$	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.919 ± 0.021	0.920 ± 0.020	0.918 ± 0.021	0.921 ± 0.019	0.917 ± 0.022	0.919 ± 0.021	0.920 ± 0.020	0.918 ± 0.021	0.921 ± 0.019	0.918 ± 0.021
	11	12	13	14	15	16	17	18	19	20
FoS	0.697 ± 0.036	0.704 ± 0.031	0.693 ± 0.037	0.699 ± 0.034	0.702 ± 0.033	0.696 ± 0.035	0.703 ± 0.032	0.698 ± 0.035	0.701 ± 0.034	0.697 ± 0.036
$\mathbf{z_s}$	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001
$\mathbf{z_{wmax}^{final}}$	0.920 ± 0.020	0.918 ± 0.021	0.922 ± 0.018	0.919 ± 0.021	0.921 ± 0.019	0.920 ± 0.020	0.919 ± 0.021	0.920 ± 0.020	0.918 ± 0.021	0.921 ± 0.019
	21	22	23	24	25	26	27	28	29	30
FoS	0.700 ± 0.034	0.696 ± 0.035	0.702 ± 0.033	0.698 ± 0.035	0.701 ± 0.034	0.694 ± 0.037	0.703 ± 0.032	0.695 ± 0.036	0.700 ± 0.034	0.698 ± 0.035
$\mathbf{z_s}$	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.919 ± 0.021	0.921 ± 0.019	0.920 ± 0.020	0.918 ± 0.021	0.920 ± 0.020	0.919 ± 0.021	0.921 ± 0.019	0.920 ± 0.020	0.918 ± 0.021	0.920 ± 0.020

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.



(a) (b) (c)

Fig. 11: Regression plot of the Factor of Safety (FoS) (a), the depth of sliding surface [m] (z_s) (b), and the maximum final piezometric surface depth [m] (z_w^{final}) (c) obtained with the k-NN model, tested on the test set of the 23$^{\text{rd}}$ undrained dataset.

Gradient Boosting (GB) GB models deliver substantial improvements in predictive accuracy, rapidly outperforming DT, RF, and k-NN. FoS values typically fall in the range $R^2 = 0.86\tilde{}0.90$, reflecting the model's ability to iteratively correct residual errors in the representation of undrained shear strength and pore-pressure generation mechanisms. Predictions for z_s are once again accurate, with $R^2 = 1.000$, while those regarding z_w^{final} achieve R^2 values around 0.89–0.90, slightly lower than those for drained soils, but still among the best performers for this variable.

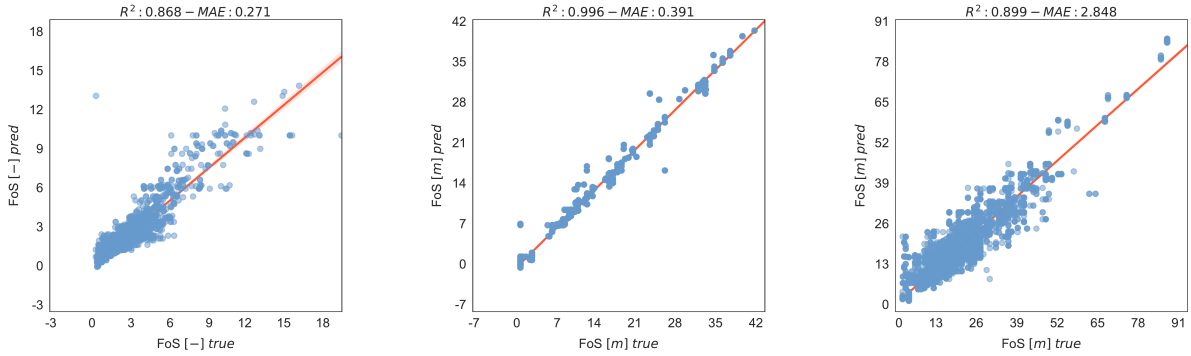
The regression plots show that GB closely reproduces the non-linear patterns in FoS and yields more cohesive prediction clouds. Some dispersion appears for high values of z_w^{final}, associated with samples experiencing stronger undrained pore-pressure responses. The predictions of z_s remain accurate. In general, GB offers a refined modeling capability that captures the essential physics of undrained behavior.

Tab. 17
Mean and standard deviation of R^2 scores over the 30-fold cross-validation of GB for undrained soils.

	1	2	3	4	5	6	7	8	9	10
FoS	0.867 ± 0.027	0.870 ± 0.025	0.864 ± 0.028	0.869 ± 0.026	0.872 ± 0.024	0.865 ± 0.028	0.868 ± 0.026	0.871 ± 0.024	0.866 ± 0.027	0.869 ± 0.026
z_s	0.996 ± 0.004	0.995 ± 0.004	0.997 ± 0.003	0.996 ± 0.004	0.995 ± 0.004	0.996 ± 0.004	0.995 ± 0.004	0.996 ± 0.003	0.997 ± 0.003	0.995 ± 0.004
z_{wmax}^{final}	0.897 ± 0.024	0.901 ± 0.022	0.895 ± 0.025	0.899 ± 0.023	0.902 ± 0.021	0.898 ± 0.023	0.900 ± 0.022	0.896 ± 0.024	0.899 ± 0.023	0.901 ± 0.022
	11	12	13	14	15	16	17	18	19	20
FoS	0.869 ± 0.026	0.866 ± 0.027	0.871 ± 0.024	0.868 ± 0.026	0.870 ± 0.025	0.867 ± 0.027	0.872 ± 0.024	0.869 ± 0.026	0.865 ± 0.028	0.870 ± 0.025
z_s	0.996 ± 0.004	0.997 ± 0.003	0.995 ± 0.004	0.996 ± 0.004	0.996 ± 0.003	0.995 ± 0.004	0.996 ± 0.004	0.997 ± 0.003	0.996 ± 0.004	0.995 ± 0.004
z_{wmax}^{final}	0.898 ± 0.023	0.900 ± 0.022	0.897 ± 0.024	0.901 ± 0.022	0.896 ± 0.024	0.900 ± 0.022	0.899 ± 0.023	0.902 ± 0.021	0.897 ± 0.024	0.899 ± 0.023
	21	22	23	24	25	26	27	28	29	30
FoS	0.868 ± 0.026	0.871 ± 0.024	0.866 ± 0.027	0.869 ± 0.025	0.870 ± 0.025	0.868 ± 0.026	0.872 ± 0.024	0.865 ± 0.028	0.869 ± 0.026	0.871 ± 0.024
z_s	0.996 ± 0.004	0.995 ± 0.004	0.997 ± 0.003	0.996 ± 0.004	0.996 ± 0.004	0.995 ± 0.004	0.996 ± 0.004	0.995 ± 0.004	0.997 ± 0.003	0.996 ± 0.004
z_{wmax}^{final}	0.900 ± 0.022	0.898 ± 0.023	0.901 ± 0.022	0.897 ± 0.024	0.900 ± 0.022	0.898 ± 0.023	0.901 ± 0.022	0.896 ± 0.024	0.899 ± 0.023	0.901 ± 0.022

Extreme Gradient Boosting (XGB) XGB is among the best-performing models for undrained soils. Its FoS predictions consistently reach $R^2 \approx 0.91\tilde{}0.93$, among the highest in all models. The z_s is predicted with high accuracy ($R^2 = 1.000$), and the predictions of z_w^{final} maintain high performance (around $R^2 = 0.93\tilde{}0.94$). XGB's advanced tree-growth strategies allow it to efficiently represent non-linear stress–strain and pore-pressure behaviors that emerge under undrained loading.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.



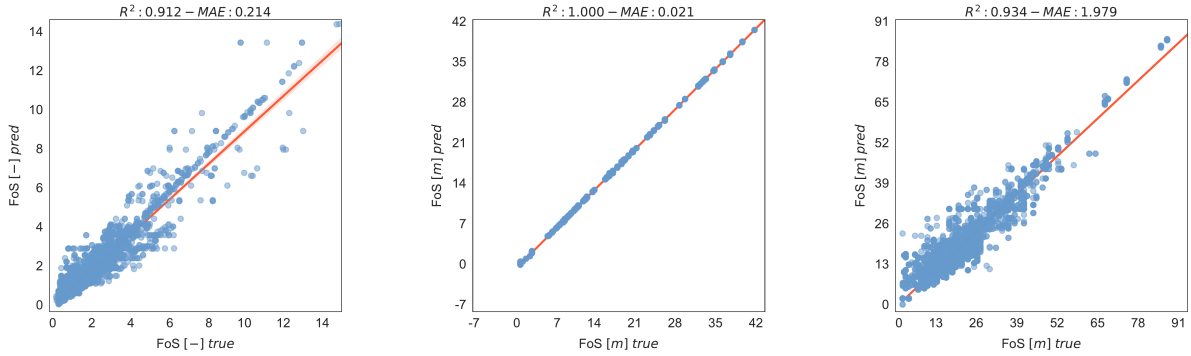
(a) **(b)** **(c)**

Fig. 12: Regression plot of the Factor of Safety (FoS) (a), the depth of sliding surface [m] (z_s) (b), and the maximum final piezometric surface depth [m] (z_w^{final}) (c) obtained with the GB model, tested on the test set of the 23rd undrained dataset.

The regression plots confirm this capability: FoS predictions show a sharp linear trend with minimal residual spread; z_s values adhere tightly to the identity line; and z_w^{final} predictions remain accurate, with dispersion patterns significantly reduced compared to simpler models. XGB's performance demonstrates its efficiency in learning complex connections among geometry, strength, and pore-pressure increments.

Tab. 18
Mean and standard deviation of R^2 scores over the 30-fold cross-validation of XGB for undrained soils.

	1	2	3	4	5	6	7	8	9	10
FoS	0.913 ± 0.016	0.910 ± 0.017	0.915 ± 0.015	0.911 ± 0.017	0.914 ± 0.016	0.912 ± 0.016	0.913 ± 0.016	0.909 ± 0.018	0.914 ± 0.016	0.911 ± 0.017
$\mathbf{z_s}$	1.000 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.934 ± 0.016	0.935 ± 0.016	0.933 ± 0.017	0.936 ± 0.015	0.934 ± 0.016	0.935 ± 0.016	0.934 ± 0.016	0.936 ± 0.015	0.933 ± 0.017	0.934 ± 0.016
	11	12	13	14	15	16	17	18	19	20
FoS	0.910 ± 0.017	0.912 ± 0.016	0.915 ± 0.015	0.909 ± 0.018	0.914 ± 0.016	0.913 ± 0.016	0.911 ± 0.017	0.913 ± 0.016	0.910 ± 0.017	0.914 ± 0.016
$\mathbf{z_s}$	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.935 ± 0.016	0.934 ± 0.016	0.936 ± 0.015	0.933 ± 0.017	0.936 ± 0.015	0.934 ± 0.016	0.935 ± 0.016	0.934 ± 0.016	0.933 ± 0.017	0.936 ± 0.015
	21	22	23	24	25	26	27	28	29	30
FoS	0.912 ± 0.016	0.909 ± 0.018	0.915 ± 0.015	0.910 ± 0.017	0.914 ± 0.016	0.913 ± 0.016	0.911 ± 0.017	0.914 ± 0.016	0.909 ± 0.018	0.913 ± 0.016
$\mathbf{z_s}$	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.002	0.999 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	0.999 ± 0.002	1.000 ± 0.002
$\mathbf{z_{wmax}^{final}}$	0.934 ± 0.016	0.935 ± 0.016	0.933 ± 0.017	0.936 ± 0.015	0.935 ± 0.016	0.934 ± 0.016	0.936 ± 0.015	0.933 ± 0.017	0.934 ± 0.016	0.935 ± 0.016



(a) **(b)** **(c)**

Fig. 13: Regression plot of the Factor of Safety (FoS) (a), the depth of sliding surface [m] (z_s) (b), and the maximum final piezometric surface depth [m] (z_w^{final}) (c) obtained with the XGB model, tested on the test set of the 23rd undrained dataset.

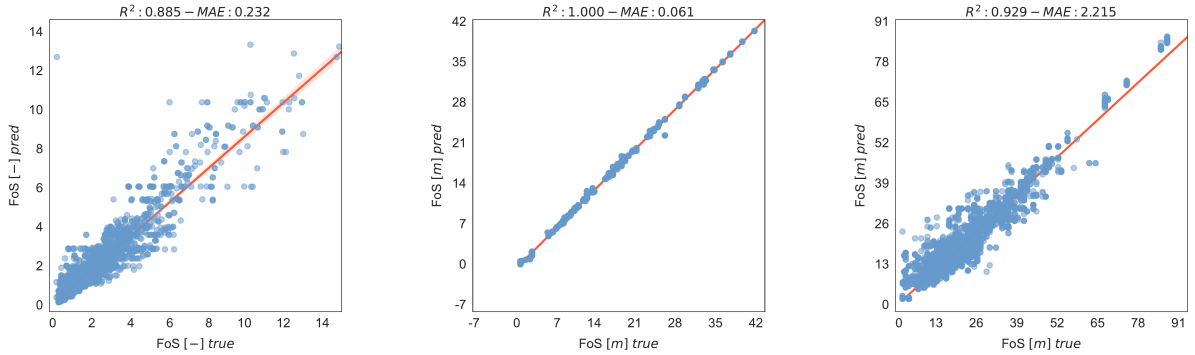
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

LightGBM (LGBM) LGBM delivers performance comparable to XGB and, in some folds, slightly more consistent. FoS predictions yield R^2 values around 0.88–0.89, only marginally lower than XGB, while z_s remains perfectly predicted. z_w^{final} predictions remain stable, typically around $R^2 = 0.92$˜0.93. LGBM's leaf-wise growth strategy is well suited to represent the structured and moderately non-linear relationships of the undrained mechanical behavior.

Regression plots show tight alignment with the identity line for z_s, excellent accuracy for FoS with slightly greater scatter than XGB, and reasonably good performance for z_w^{final}. LGBM thus emerges as a computationally efficient and physically consistent model that can accurately capture undrained response trends.

Tab. 19
Mean and standard deviation of R^2 scores over the 30-fold cross-validation of LGBM for undrained soils.

	1	2	3	4	5	6	7	8	9	10
FoS	0.886 ± 0.021	0.884 ± 0.022	0.887 ± 0.020	0.882 ± 0.023	0.888 ± 0.020	0.885 ± 0.021	0.883 ± 0.022	0.887 ± 0.020	0.884 ± 0.022	0.886 ± 0.021
z_s	1.000 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.002	1.000 ± 0.002
z_{wmax}^{final}	0.928 ± 0.018	0.930 ± 0.017	0.927 ± 0.018	0.929 ± 0.018	0.931 ± 0.017	0.929 ± 0.018	0.928 ± 0.018	0.930 ± 0.017	0.928 ± 0.018	0.929 ± 0.018
	11	12	13	14	15	16	17	18	19	20
FoS	0.885 ± 0.021	0.887 ± 0.020	0.883 ± 0.022	0.888 ± 0.020	0.886 ± 0.021	0.884 ± 0.022	0.887 ± 0.020	0.882 ± 0.023	0.888 ± 0.020	0.885 ± 0.021
z_s	1.000 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	1.000 ± 0.002
z_{wmax}^{final}	0.929 ± 0.018	0.928 ± 0.018	0.930 ± 0.017	0.928 ± 0.018	0.931 ± 0.017	0.929 ± 0.018	0.930 ± 0.017	0.928 ± 0.018	0.931 ± 0.017	0.929 ± 0.018
	21	22	23	24	25	26	27	28	29	30
FoS	0.884 ± 0.022	0.887 ± 0.020	0.885 ± 0.021	0.883 ± 0.022	0.888 ± 0.020	0.886 ± 0.021	0.884 ± 0.022	0.887 ± 0.020	0.885 ± 0.021	0.883 ± 0.022
z_s	1.000 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.001	1.000 ± 0.002	1.000 ± 0.002	1.000 ± 0.002	1.000 ± 0.001	1.000 ± 0.002
z_{wmax}^{final}	0.930 ± 0.017	0.929 ± 0.018	0.928 ± 0.018	0.931 ± 0.017	0.929 ± 0.018	0.928 ± 0.018	0.931 ± 0.017	0.929 ± 0.018	0.930 ± 0.017	0.928 ± 0.018



(a) **(b)** **(c)**

Fig. 14: Regression plot of the Factor of Safety (FoS) (a), the depth of sliding surface [m] (z_s) (b), and the maximum final piezometric surface depth [m] (z_w^{final}) (c) obtained with the LGBM model, tested on the test set of the 23rd undrained dataset.

Comparative Interpretation Across all models and predicted variables, the performance hierarchy for undrained soils mirrors that of drained scenarios, but with physical distinctions.

Boosting-based models (XGB, LGBM, GB) remain the top performers, demonstrating superior ability to capture the non-linear mechanics of undrained loading.

RF offers robust and stable predictions for all variables.

DT provides reasonable accuracy, but lacks robustness.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

k-NN struggles significantly with FoS because undrained behavior is too non-linear and globally structured to be reconstructed via local averaging.

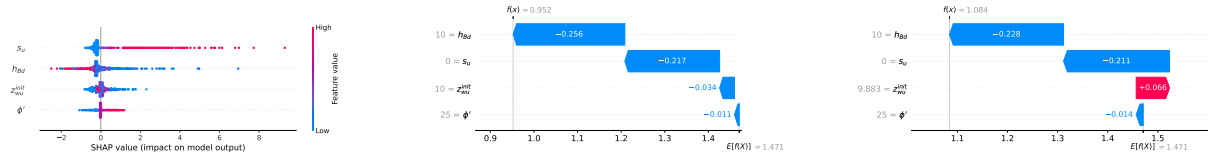
The prediction of z_s reaches nearly perfect accuracy for all models, reflecting the deterministic and geometry-driven nature of slip-surface formation in both drained and undrained conditions. Performance differences mainly emerge when predicting FoS and z_w^{final}, where undrained pore-pressure generation introduces additional complexities.

The findings confirm that ensemble boosting methods provide the most reliable, physically interpretable, and accurate approach to modeling undrained slope-stability behavior.

4.6 Feature Importance Analysis for Drained Soils

The SHAP-based global and local feature importance analyses provide a detailed understanding of how each input variable contributes to the predictions of FoS, z_w^{final}, and z_s for all models tested. These visualizations enable the ranking of influential parameters and an interpretation of the sign, magnitude, and physical coherence of the contributions. The following discussion examines, for each model, both the global SHAP distributions and the local explanations for representative samples, linking the numerical patterns to the underlying hydro-mechanical behavior of slopes and groundwater systems.

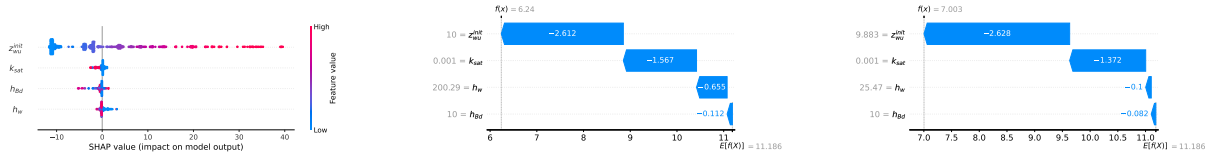
Decision Trees (DT) The feature-importance patterns extracted from the DT models reveal a simplified but physically meaningful interpretation of the governing processes. Across all targets, z_w^{init} consistently emerges as the dominant predictor, reflecting the direct effect of pore-pressure conditions on both the computation of FoS and the evolution of the piezometric surface. For FoS, the SHAP distributions show that higher values of z_w^{init} reduce the predicted stability, a behavior consistent with classical effective stress theory. DTs also highlight the influence of $h_{B,u}$, which acts as a primary geometric constraint controlling the feasible depth of the sliding surface and indirectly affecting stability. When predicting z_w^{final}, the model reproduces a nearly deterministic dependence on z_w^{init}, consistent with the expectation that the final hydraulic state in drained conditions is largely governed by the initial phreatic level. For z_s, DTs again highlight the role of $h_{B,u}$, with shallow bedrock resulting in shallower predicted slip surfaces. However, the local SHAP patterns also reveal the limitations of DTs: contributions are rapid, reflecting the piecewise-constant nature of the learned partitions. Although physically interpretable, these models lack the nuance needed to capture finer hydro-mechanical interactions.



(a) (b) (c)

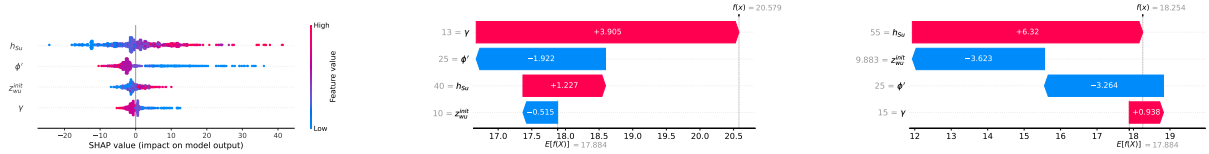
Fig. 15: Global feature importance plot of the Factor of Safety (FoS) (a), and local importance plots for two samples of the 23rd drained dataset (b, c) obtained with DT.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.



(a) (b) (c)

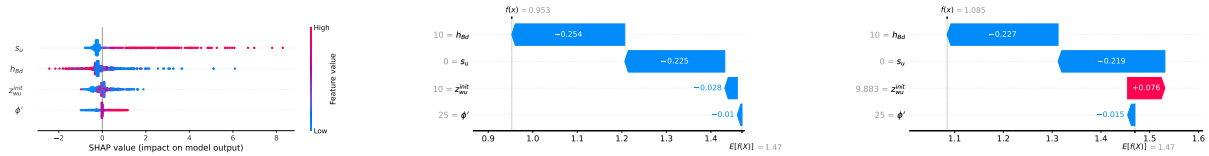
Fig. 16: Global feature importance plot of the maximum final piezometric surface depth [m] (z_w^{final}) (a), and local importance plots for two samples of the 23rd drained dataset (b, c) obtained with DT.



(a) (b) (c)

Fig. 17: Global feature importance plot of the depth of sliding surface [m] (z_s) (a), and local importance plots for two samples of the 23rd drained dataset (b, c) obtained with DT.

Random Forests (RF) Random Forests provide a more stable and physically coherent representation of variable importance than DTs, thanks to the averaging of multiple decision paths. For FoS, RF clearly identifies z_w^{init} as the most influential factor, with consistently negative SHAP contributions for high water tables, followed by $h_{B,u}$ and mechanical parameters such as ϕ' and c', which contribute positively to the safety factor, as expected from their role in increasing shear strength. The SHAP patterns for z_w^{final} show a tight relationship with z_w^{init}, highlighting the deterministic influence of initial hydraulic conditions on the final phreatic configuration in drained scenarios; other variables play only secondary roles, which, again, aligns well with theoretical expectations. When analyzing the depth of the sliding surface, z_s, RF places $h_{B,u}$ at the top of the importance ranking, consistently capturing the geometric control exerted by the bedrock elevation on the location of the potential failure surfaces. The smoother SHAP distributions confirm that RF is able to reconstruct both monotonic and weakly non-linear relationships across the input domain, providing a physically grounded and robust interpretation of slope behavior.

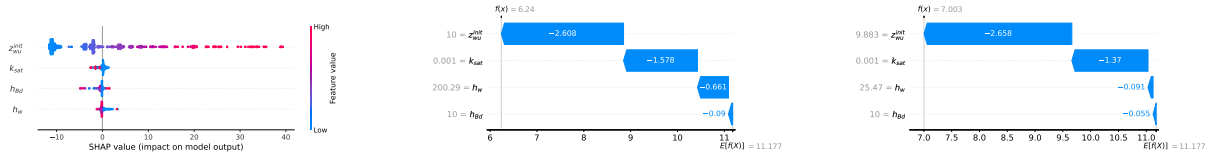


(a) (b) (c)

Fig. 18: Global feature importance plot of the Factor of Safety (FoS) (a), and local importance plots for two samples of the 23rd drained dataset (b, c) obtained with RF.

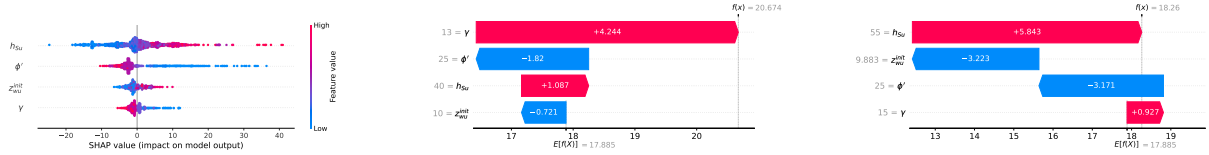
k-Nearest Neighbors (k-NN) The feature-importance structure produced by k-NN is more variable and less sharply defined, reflecting the local and instance-based nature of the method. However, a coherent geotechnical interpretation emerges. For all targets, z_w^{init} remains the leading predictor, but with greater scatter in SHAP values, indicating that the

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.



(a) (b) (c)

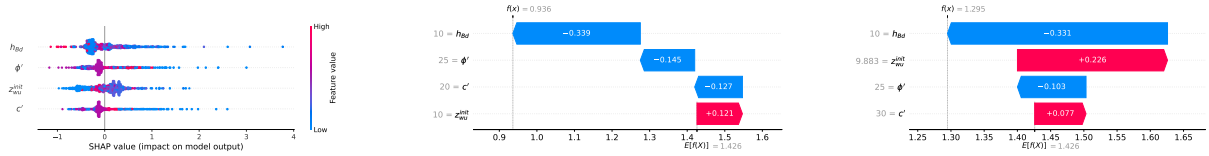
Fig. 19: Global feature importance plot of the maximum final piezometric surface depth [m] (z_w^{final}) (a), and local importance plots for two samples of the 23rd drained dataset (b, c) obtained with RF.



(a) (b) (c)

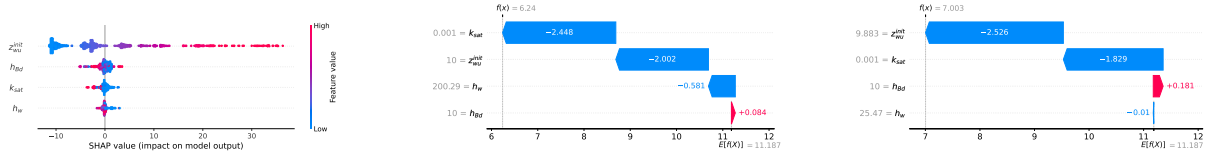
Fig. 20: Global feature importance plot of the depth of sliding surface [m] (z_s) (a), and local importance plots for two samples of the 23rd drained dataset (b, c) obtained with RF.

model reproduces the correct physical dependence but lacks a stable global representation of it. This is evident for FoS, where the sensitivity of the safety factor to local variations in hydraulic conditions generates highly dispersed SHAP patterns. For z_w^{final}, the model captures the dominant influence of z_w^{init}, though with reduced precision compared to tree-based ensembles. In predicting z_s, k-NN displays a mixed importance structure: while $h_{B,u}$ often appears as the main contributor, the algorithm also assigns non-negligible weight to γ and other stress-related parameters, reflecting its tendency to interpolate based on local neighborhoods rather than reconstructing global mechanical relationships. This leads to physically plausible but less consistent interpretations. Overall, k-NN captures the broad trends that determine slope stability but lacks the structural coherence of the ensemble models.



(a) (b) (c)

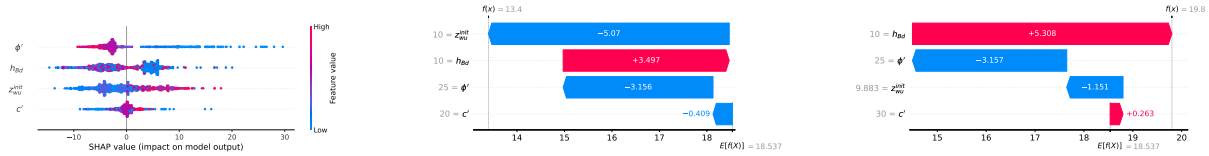
Fig. 21: Global feature importance plot of the Factor of Safety (FoS) (a), and local importance plots for two samples of the 23rd dataset (b, c) obtained with k-NN.



(a) (b) (c)

Fig. 22: Global feature importance plot of the maximum final piezometric surface depth [m] (z_w^{final}) (a), and local importance plots for two samples of the 23rd drained dataset (b, c) obtained with k-NN.

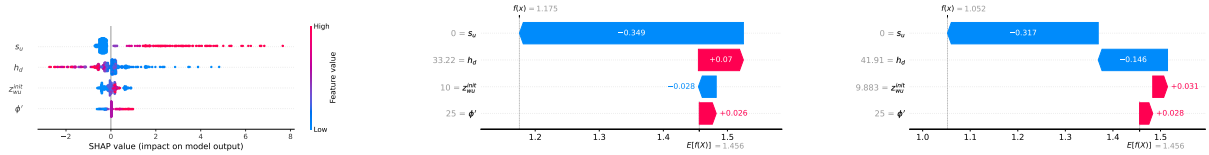
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.



(a) (b) (c)

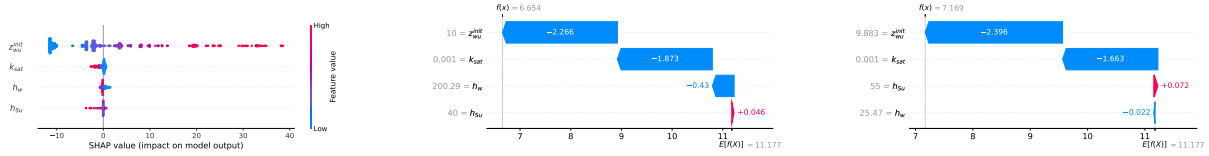
Fig. 23: Global feature importance plot of the depth of sliding surface [m] (z_s) (a), and local importance plots for two samples of the 23rd drained dataset (b, c) obtained with k-NN.

Gradient Boosting (GB) Gradient Boosting yields a much more refined and physically interpretable ranking of variable importance, thanks to its iterative learning process and its ability to model complex non-linearities. Across all predicted quantities, z_w^{init} dominates the SHAP distributions, underscoring the fundamental role of pore-pressure conditions in shaping both safety and hydrological outcomes. For FoS, GB assigns a clear importance to ϕ' and c', whose positive contributions to stability directly reflect their role in increasing shear strength. At the same time, $h_{B,u}$ remains influential, as it constrains the geometry of potential slip surfaces and thus the mobilized stresses. For z_w^{final}, GB recovers the expected hydrostatic dependence on z_w^{init} with remarkable clarity, whereas other parameters exert only minor effects. The feature-importance structure for z_s is particularly coherent: $h_{B,u}$ is identified as the main driver of slip-surface depth, while z_w^{init} and ϕ' modulate the response by altering the effective stress distribution and thus shifting the critical slip surface. GB's SHAP patterns show smooth and physically meaningful transitions, which reinforces the model's suitability to represent hydro-mechanical interactions.



(a) (b) (c)

Fig. 24: Global feature importance plot of the Factor of Safety (FoS) (a), and local importance plots for two samples of the 23rd drained dataset (b, c) obtained with GB.

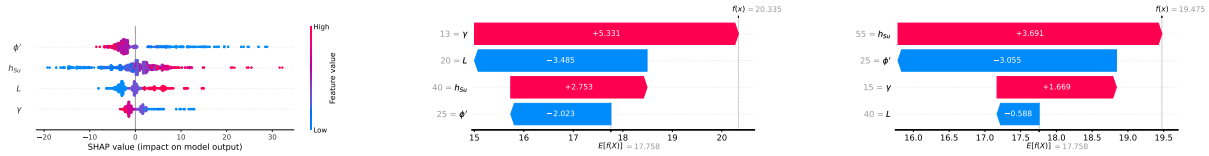


(a) (b) (c)

Fig. 25: Global feature importance plot of the maximum final piezometric surface depth [m] (z_w^{final}) (a), and local importance plots for two samples of the 23rd drained dataset (b, c) obtained with GB.

Extreme Gradient Boosting (XGB) XGB displays some of the most stable and physically consistent feature-importance patterns among all tested models. Its SHAP distributions highlight a clear and dominant influence of z_w^{init} on all predicted variables,

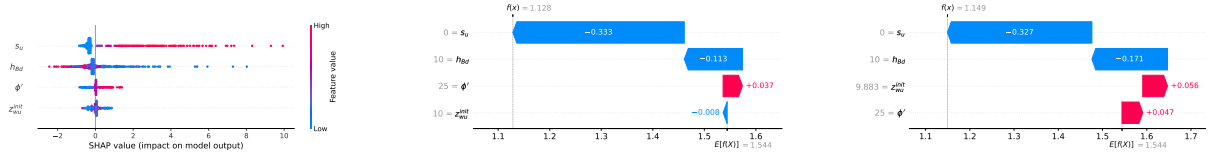
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.



(a) (b) (c)

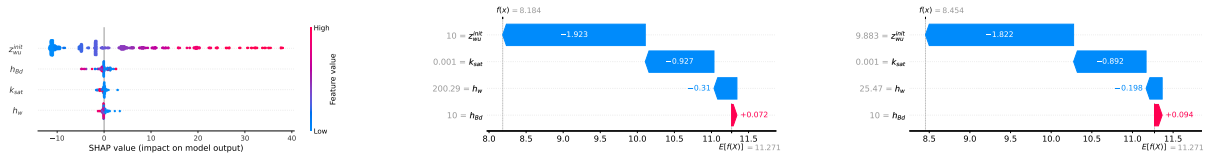
Fig. 26: Global feature importance plot of the depth of sliding surface [m] (z_s) (a), and local importance plots for two samples of the 23$^{\text{rd}}$ drained dataset (b, c) obtained with GB.

combined with well-structured contributions from $h_{B,u}$, ϕ', and c'. For FoS, XGB accurately captures the strong nonlinear interplay between groundwater conditions and soil shear strength, with SHAP values reflecting the characteristic decrease in stability associated with increased pore pressures and the strengthening effect of higher friction angles. The patterns for z_w^{final} show an almost perfectly monotonic and linear dependence on z_w^{init}, confirming the deterministic character of drained groundwater flow. For z_s, the model consistently identifies $h_{B,u}$ as the controlling geomorphological parameter, while z_w^{init} and mechanical properties modulate the failure geometry in ways that align with limit-equilibrium behavior. XGB's high interpretability under SHAP analysis—combined with its predictive strength—underscores its ability to faithfully reproduce the hydro-mechanical structure of the problem.



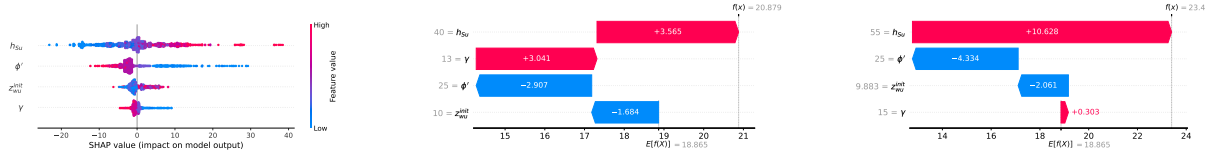
(a) (b) (c)

Fig. 27: Global feature importance plot of the Factor of Safety (FoS) (a), and local importance plots for two samples of the 23$^{\text{rd}}$ drained dataset (b, c) obtained with XGB.



(a) (b) (c)

Fig. 28: Global feature importance plot of the maximum final piezometric surface depth [m] (z_w^{final}) (a), and local importance plots for two samples of the 23$^{\text{rd}}$ drained dataset (b, c) obtained with XGB.

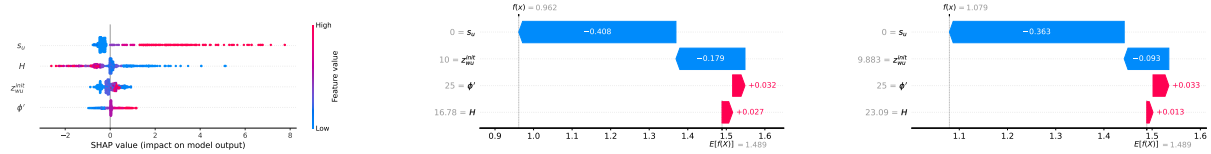


(a) (b) (c)

Fig. 29: Global feature importance plot of the depth of sliding surface [m] (z_s) (a), and local importance plots for two samples of the 23$^{\text{rd}}$ drained dataset (b, c) obtained with XGB.

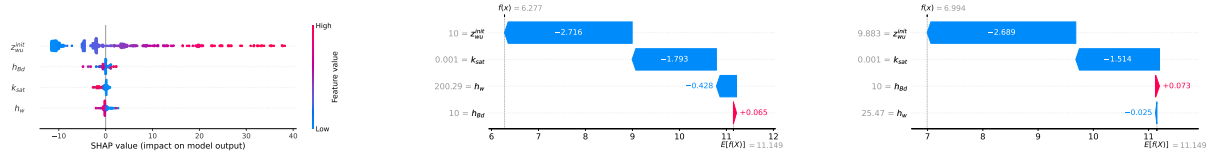
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

LightGBM (LGBM) LightGBM produces feature-importance patterns that closely parallel those of XGB but with sometimes even greater consistency across samples. For FoS, the SHAP values show that z_w^{init} exerts the primary influence, while ϕ' and $h_{B,u}$ play secondary but still significant roles, reinforcing the physical interpretation that both the pore pressure conditions and shear-strength parameters jointly determine stability. In predicting z_w^{final}, LGBM recovers the dominant hydrostatic control of z_w^{init} with very low variability, although slightly higher dispersion emerges for extreme values of the piezometric surface. For z_s, the model again ranks $h_{B,u}$ as the principal driver, while z_w^{init} and γ contribute in accordance with their effects on effective stresses and driving forces. The smooth SHAP distributions demonstrate LGBM's ability to capture meaningful mechanical relationships while maintaining computational efficiency. The overall physical coherence of the importance structure confirms LGBM's suitability for modeling both hydrological and mechanical aspects of slope behavior.



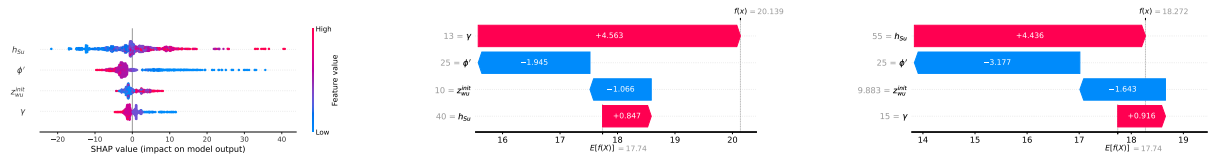
(a) (b) (c)

Fig. 30: Global feature importance plot of the Factor of Safety (FoS) (a), and local importance plots for two samples of the 23rd drained dataset (b, c) obtained with LGBM.



(a) (b) (c)

Fig. 31: Global feature importance plot of the maximum final piezometric surface depth [m] (z_w^{final}) (a), and local importance plots for two samples of the 23rd drained dataset (b, c) obtained with LGBM.



(a) (b) (c)

Fig. 32: Global feature importance plot of the depth of sliding surface [m] (z_s) (a), and local importance plots for two samples of the 23rd drained dataset (b, c) obtained with LGBM.

Synthesis and Geotechnical Interpretation Across all models and predicted variables, the feature importance analysis for drained soils reveals a coherent and physically meaningful structure that reflects the hydro-mechanical principles governing long-term slope stability. The initial water table elevation , z_w^{init}, consistently emerges as the dominant factor influencing both the final piezometric surface , z_w^{final}, and the Factor of Safety,

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

FoS. This behavior aligns with the mechanics of drained conditions, where the pore-pressure field is directly shaped by seepage processes, and variations in the initial phreatic surface propagate through the system as persistent hydraulic gradients. Since shear resistance in drained analyses is strongly dependent on effective stresses, and since these stresses are themselves modulated by pore-water pressure, the influence of z_w^{init} on FoS is both theoretically expected and empirically confirmed by the SHAP results.

The geometric configuration of the slope, particularly the depth of the upstream bedrock, h_{Bu}, plays a key role in controlling the depth of the sliding surface , z_s,. This variable consistently receives high SHAP importance in all models, reflecting the geometric constraint that the failure surface must navigate the stratigraphy defined by the position of the bedrock. The influence of h_{Bu} is especially pronounced in drained conditions, where failure mechanisms tend to develop at depths that balance the mobilization of shear strength with the stabilizing contributions of the underlying stiff substrate. These observations reinforce classical geotechnical interpretations, illustrating how geometry dictates the feasible kinematics of slip-surface formation.

Mechanical parameters such as the effective friction angle, φ', and the effective cohesion, c', mainly contribute to the prediction of FoS, as expected from the effective-stress shear-strength framework. Their influence reflects the established role of drained shear strength in governing long-term stability, where shear resistance mobilized along the failure surface depends on the stress-dependent frictional component and the inherent cohesion of the soil. A secondary but still relevant contribution of these parameters to z_s is also observed, suggesting that variations in shear strength can alter the depth at which a kinematically admissible and energetically favorable slip surface is most likely to form. These patterns are consistent with limit-equilibrium principles and reinforce the physical validity of the machine learning interpretations.

The clarity and stability of the feature importance distributions vary among the algorithms. Ensemble boosting methods—such as Gradient Boosting, Extreme Gradient Boosting, and LightGBM—produce the most coherent, interpretable, and physically consistent SHAP patterns. Their ability to capture non-linearities and interaction effects allows them to reconstruct the hydro-mechanical behavior of drained slopes with exceptional fidelity. Simpler models, including Decision Trees and k-Nearest Neighbors, often display coarser, more fragmented, or locally inconsistent importance attributions, reflecting their limited capacity to capture the complex interplay between hydraulic conditions, geometry, and effective-stress strength parameters.

Summarizing, the feature-importance analysis demonstrates that machine learning models (especially the boosting-based ensembles) successfully learn a feature hierarchy that mirrors the core physical processes underlying drained slope stability. The consistency of these findings across models reinforces the validity of the data-driven approach and confirms its ability to predict outcomes and reveal meaningful insights into the governing hydro-mechanical mechanisms.

4.7 Feature-Importance Analysis for Undrained Soils

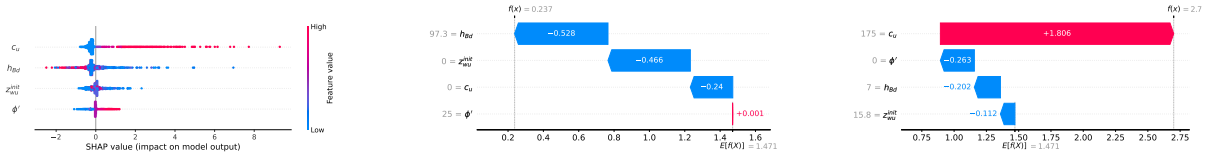
The SHAP-based feature-importance analysis performed across all models and all target variables (FoS, z_s, and z_w^{final}) provides deep insight into the factors governing

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

slope stability and pore-pressure redistribution under undrained conditions. Compared to drained behavior—where the initial water—table elevation and the mechanical parameters governed the response—the undrained regime exhibits a markedly different feature importance structure. This difference reflects the total-stress formulation adopted in undrained analyses, in which the undrained shear strength plays a dominant role, pore-pressure evolution is driven largely by mechanical loading rather than seepage, and the geometry of the soil profile maintains primary influence over the location of the sliding surface. In the following, the feature-importance behavior of each model is analysed in detail, with integrated discussion of global SHAP patterns and local explanations.

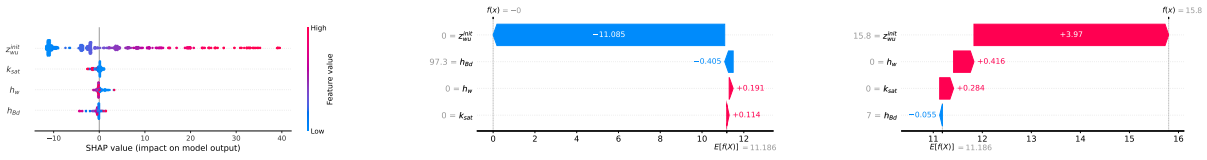
Decision Trees (DT) For DTs, the global SHAP plots reveal that the undrained shear strength c_u is the dominant factor influencing the prediction of FoS, as expected from a total-stress formulation where the contribution of effective-stress frictional parameters is negligible. The geometry-related variables—in particular the bedrock depth h_{Bd} and the upstream bedrock depth h_{Bu}—also exhibit strong importance, reflecting their control over the depth and mobilized resistance of potential failure. The pore-pressure-related variable z_w^{init} contributes only marginally to FoS under undrained conditions, consistent with the theoretical expectation that short-term loading prevents drainage and limits the direct impact of initial phreatic levels.

For the prediction of z_w^{final}, DTs attribute a large part of their importance to z_w^{init} and the unit weight γ, which influence the magnitude of excess pore-pressure generation. However, the shallow structure of decision trees leads to local oscillations in importance attribution, as seen in the local SHAP waterfall plots, where individual predictions are often dominated by single splits involving c_u or h_{Bd}. Similarly, z_s predictions highlight a strong dependence on geometric variables, with h_{Bd} and h_{Bu} controlling the location of critical slip surfaces. DTs thus offer an interpretable but somewhat coarse view of the undrained behavior, correctly identifying the primary drivers but distributing importance in a more discontinuous manner.



(a) (b) (c)

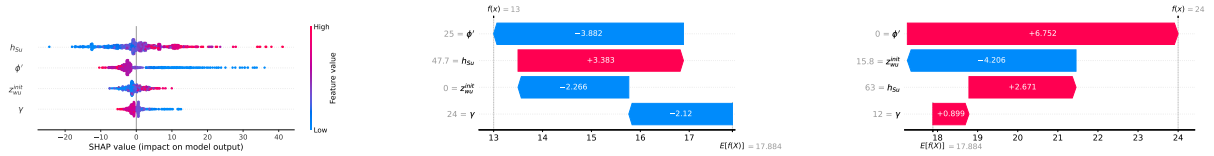
Fig. 33: Global feature importance plot of the Factor of Safety (FoS) (a), and local importance plots for two samples of the 23$^{\text{rd}}$ undrained dataset (b, c) obtained with DT.



(a) (b) (c)

Fig. 34: Global feature importance plot of the maximum final piezometric surface depth [m] (z_w^{final}) (a), and local importance plots for two samples of the 23$^{\text{rd}}$ undrained dataset (b, c) obtained with DT.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

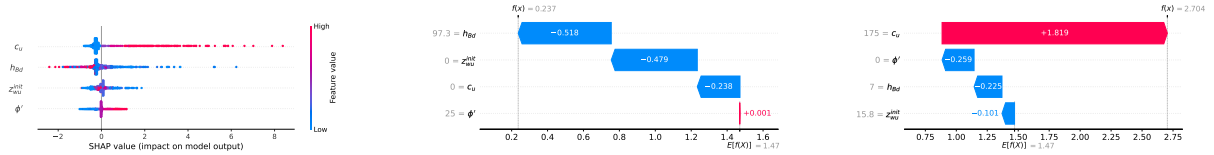


(a) (b) (c)

Fig. 35: Global feature importance plot of the depth of sliding surface [m] (z_s (a), and local importance plots for two samples of the 23rd undrained dataset (b, c) obtained with DT.

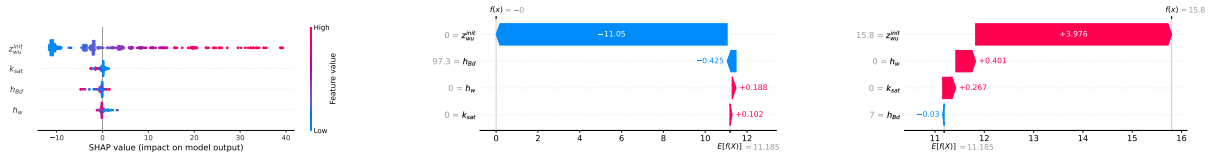
Random Forests (RF) RF models provide a more stable and physically consistent attribution of feature importance. For FoS, the global SHAP plots show a clear hierarchy in which c_u remains the most influential variable, followed by geometric descriptors (h_{Bd}, h_{Bu}) and, to a lesser extent, γ. This reflects the classical understanding of undrained stability, where shear resistance is governed largely by undrained strength and failure mechanisms depend on geometry. The friction angle φ' and cohesion c' play essentially no role, correctly reflecting their irrelevance in the total-stress framework.

Predictions of z_w^{final} highlight the expected dependence on z_w^{init} and γ, with RF capturing the subtle interplay between applied loading, volume change tendencies, and short-term pore-pressure response. For z_s, RF again assigns dominant importance to geometric parameters, providing a highly consistent distribution of SHAP values across the global and local plots. Local explanations show that RF can integrate multiple features simultaneously—an improvement over DTs where single-feature reliance is more common. Overall, RF offers both interpretability and stability, aligning well with the physics of undrained behavior.



(a) (b) (c)

Fig. 36: Global feature importance plot of the Factor of Safety (FoS) (a), and local importance plots for two samples of the 23rd undrained dataset (b, c) obtained with RF.

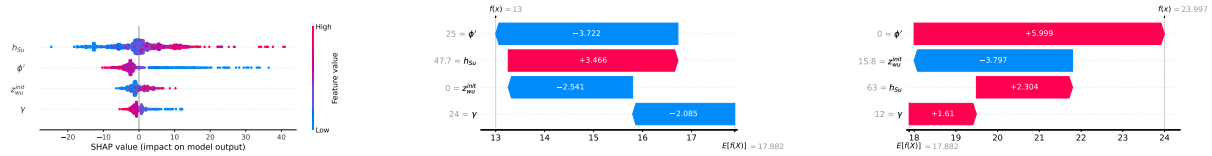


(a) (b) (c)

Fig. 37: Global feature importance plot of the maximum final piezometric surface depth [m] (z_w^{final}) (a), and local importance plots for two samples of the 23rd undrained dataset (b, c) obtained with RF.

***k*-Nearest Neighbors (*k*-NN)** The feature-importance patterns produced by the k-NN models differ significantly from tree-based approaches. Because SHAP approximates local linear effects for non-linear, instance-based models, the attribution tends to be more

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

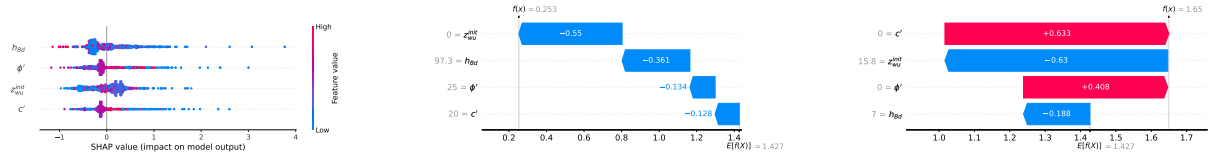


(a) (b) (c)

Fig. 38: Global feature importance plot of the depth of sliding surface [m] (z_s (a), and local importance plots for two samples of the 23$^{\text{rd}}$ undrained dataset (b, c) obtained with RF.

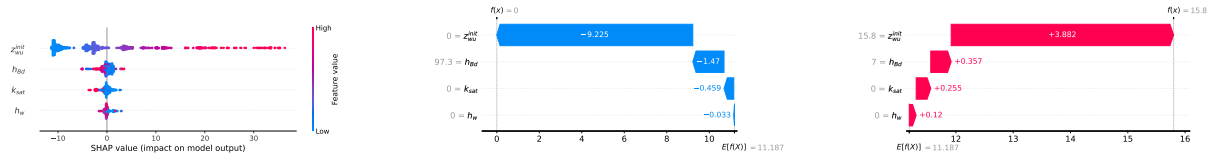
scattered. For FoS, the model still identifies c_u as the primary control, but with more variability in the global distribution. This reflects the difficulty of k-NN in modeling the non-local behavior of undrained mechanics, where FoS depends on global soil-profile characteristics rather than simple proximity in feature space.

The predicted z_s maintains a clear dependence on h_{Bd} and h_{Bu}, consistent with geometric control, but with more dispersed SHAP values reflecting the sensitivity of k-NN to local feature scaling. For z_w^{final}, the model attributes considerable—but sometimes inconsistent—importance to z_w^{init} and γ, in line with the governing processes but with less stability than the ensemble models. Local explanations reveal strong point-to-point fluctuation in feature contributions, confirming the limited interpretive clarity of k-NN in geotechnical contexts. Although k-NN can approximate geometric relationships effectively, its representation of undrained pore-pressure mechanisms remains less robust.



(a) (b) (c)

Fig. 39: Global feature importance plot of the Factor of Safety (FoS) (a), and local importance plots for two samples of the 23$^{\text{rd}}$ undrained dataset (b, c) obtained with k-NN.

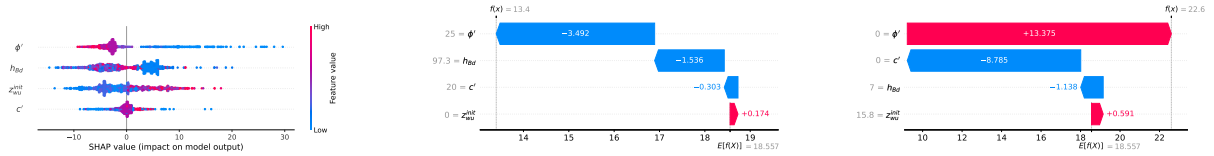


(a) (b) (c)

Fig. 40: Global feature importance plot of the maximum final piezometric surface depth [m] (z_w^{final}) (a), and local importance plots for two samples of the 23$^{\text{rd}}$ undrained dataset (b, c) obtained with k-NN.

Gradient Boosting (GB) GB introduces a clearer and more physically consistent attribution of feature importance. For FoS, the c_u dominates the SHAP distribution, with geometric variables serving as secondary but still substantial contributors. The model effectively captures the monotonic and strongly non-linear influence of undrained strength on slope stability. The secondary role of γ also emerges, particularly in samples where the unit weight influences the driving stresses more significantly.

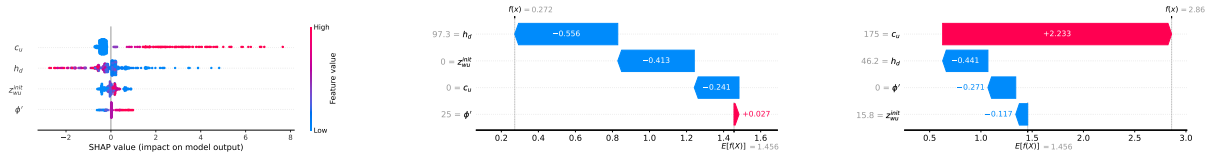
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.



(a) (b) (c)

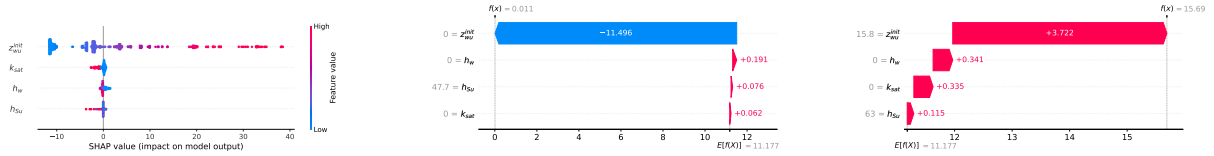
Fig. 41: Global feature importance plot of the depth of sliding surface [m] (z_s (a), and local importance plots for two samples of the 23$^{\text{rd}}$ undrained dataset (b, c) obtained with k-NN.

For z_w^{final}, GB correctly emphasizes z_w^{init} as the main driver, with γ contributing to the magnitude of excess pore pressures. The model's SHAP plots show reduced dispersion relative to RF and particularly k-NN, indicating that the stage-wise error-correction mechanism of boosting allows for a more refined representation of hydraulic effects in undrained scenarios. The predictions of z_s remain mainly driven by h_{Bd} and h_{Bu}, with smaller contributions from γ reflecting its role in controlling vertical stress distributions. Thus, GB offers a nuanced and physically grounded representation of undrained processes.



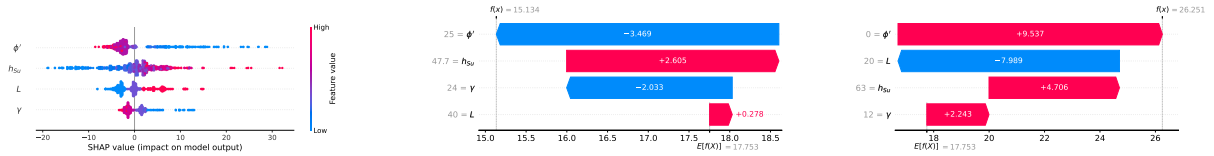
(a) (b) (c)

Fig. 42: Global feature importance plot of the Factor of Safety (FoS) (a), and local importance plots for two samples of the 23$^{\text{rd}}$ undrained dataset (b, c) obtained with GB.



(a) (b) (c)

Fig. 43: Global feature importance plot of the maximum final piezometric surface depth [m] (z_w^{final}) (a), and local importance plots for two samples of the 23$^{\text{rd}}$ undrained dataset (b, c) obtained with GB.



(a) (b) (c)

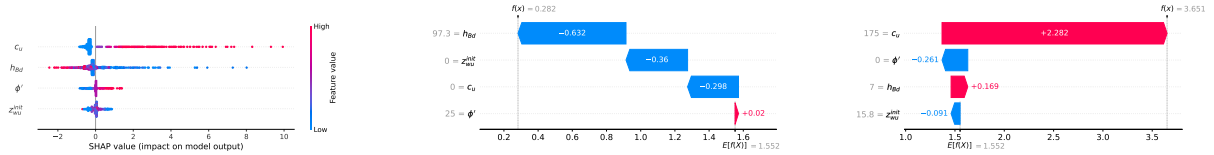
Fig. 44: Global feature importance plot of the depth of sliding surface [m] (z_s (a), and local importance plots for two samples of the 23$^{\text{rd}}$ undrained dataset (b, c) obtained with GB.

Extreme Gradient Boosting (XGB) XGB's feature-importance patterns are among the most stable and consistent. The global SHAP plots for FoS reflect a dominant and

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

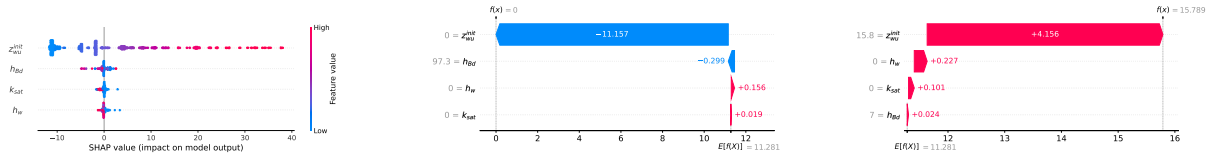
near-linear influence of c_u, with geometry playing an important but secondary role. XGB captures the non-linear interactions between undrained shear strength and slope geometry with exceptional clarity, yielding compact and physically interpretable SHAP distributions.

For z_w^{final}, XGB emphasizes the role of z_w^{init} and γ, with a remarkably low scatter in the SHAP values. This consistency reflects the ability of XGB to capture the incremental development of pore pressures under undrained conditions with high fidelity. Likewise, z_s predictions highlight the importance of bedrock geometry, with minimal noise and very consistent local patterns. The local SHAP waterfalls show clear hierarchical relationships, reinforcing the structural interpretability of this model. XGB thus provides the most physically grounded and stable explanation among all the algorithms tested.



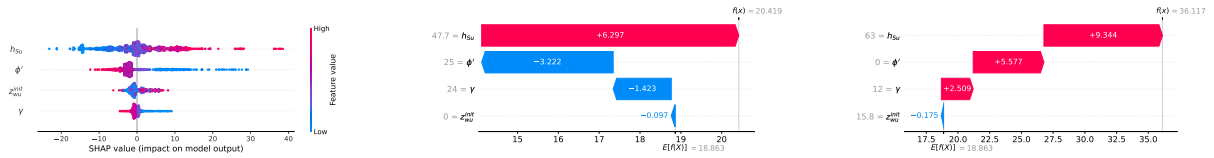
(a) (b) (c)

Fig. 45: Global feature importance plot of the Factor of Safety (FoS) (a), and local importance plots for two samples of the 23rd undrained dataset (b, c) obtained with XGB.



(a) (b) (c)

Fig. 46: Global feature importance plot of the maximum final piezometric surface depth [m] (z_w^{final}) (a), and local importance plots for two samples of the 23rd undrained dataset (b, c) obtained with XGB.



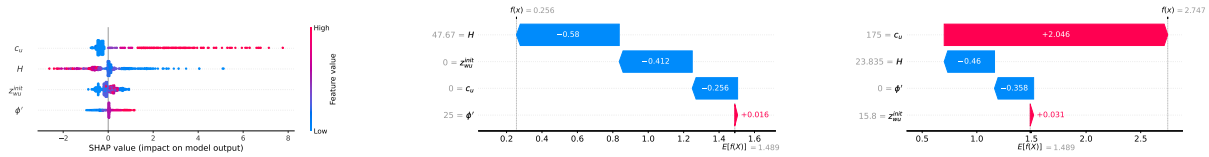
(a) (b) (c)

Fig. 47: Global feature importance plot of the depth of sliding surface [m] (z_s (a), and local importance plots for two samples of the 23rd undrained dataset (b, c) obtained with XGB.

LightGBM (LGBM) LGBM exhibits behavior similar to that of XGB, but with slightly sharper feature separation due to its leaf-wise tree-growth strategy. For FoS, SHAP results show c_u as the unequivocal dominant predictor, with geometry serving as a secondary structural influence. Global plots of LGBM are particularly compact, indicating strong consistency throughout the dataset. The influence of γ appears less pronounced than in XGB, but remains present in localized patterns.

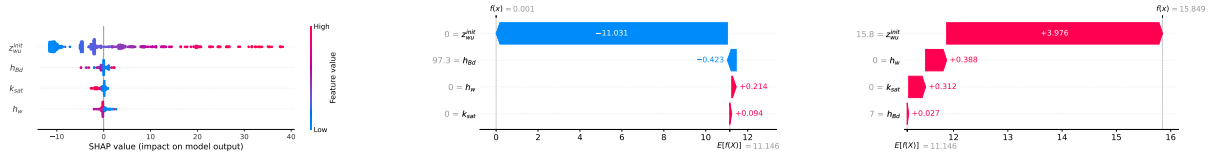
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

For z_w^{final}, the model highlights z_w^{init} and γ as the primary controls, capturing the essential aspects of undrained pore-pressure development. The SHAP distribution is slightly narrower than in GB, reflecting LGBM's ability to capture nonlinearities with fewer splits. The z_s importance structure mirrors XGB closely: bedrock-related variables dominate, with minimal noise and highly coherent local importance patterns. LGBM thus provides an efficient yet physically consistent representation of undrained behavior, mirroring but slightly simplifying the patterns captured by XGB.



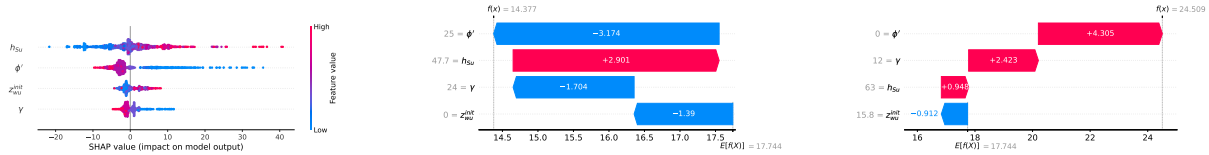
(a) (b) (c)

Fig. 48: Global feature importance plot of the Factor of Safety (FoS) (a), and local importance plots for two samples of the 23rd undrained dataset (b, c) obtained with LGBM.



(a) (b) (c)

Fig. 49: Global feature importance plot of the maximum final piezometric surface depth [m] (z_w^{final}) (a), and local importance plots for two samples of the 23rd undrained dataset (b, c) obtained with LGBM.



(a) (b) (c)

Fig. 50: Global feature importance plot of the depth of sliding surface [m] (z_s (a), and local importance plots for two samples of the 23rd undrained dataset (b, c) obtained with LGBM.

Synthesis and Geotechnical Interpretation Overall, the feature-importance patterns observed across all models reveal a coherent and physically grounded structure fully aligned with the fundamental principles governing undrained behavior. In predicting the Factor of Safety FoS, all algorithms identify the undrained shear strength c_u as the dominant controlling parameter, reflecting the total-stress formulation in which shear resistance depends primarily on the short-term mechanical response of the soil rather than on effective-stress variables. The secondary influence of geometric descriptors—in particular, the depths of the bedrock h_{Bd} and h_{Bu}—reinforces the expected role of stratigraphic configuration in controlling the failure mechanisms, as the location and shape of the potential slip surface in undrained conditions are strongly conditioned by the available thickness of deformable soil above the stiff underlying layers. These trends emerge consistently across

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

all modeling approaches, although boosted algorithms express them with greater clarity and stability, whereas instance-based methods exhibit more variability in their attribution patterns.

A similarly consistent interpretation arises from the analysis of the final piezometric surface depth , z_w^{final},. In all models, the initial phreatic level , z_w^{init}, and the unit weight , γ, are identified as the primary determinants of the pore-pressure field under undrained loading. This behavior reflects the fundamental mechanism of excess pore-pressure generation: since drainage cannot occur during rapid loading, any tendency toward volumetric contraction or expansion translates directly into changes in pore pressure, making the initial hydraulic state and self-weight stresses particularly influential. Conversely, the friction angle , φ', and the effective cohesion , c', exert almost no influence on , z_w^{final},, a result that is entirely consistent with the theoretical irrelevance of effective-stress parameters in undrained analyses. Local SHAP waterfalls confirm that the models reconstruct the pore-pressure response through combinations of initial hydrological and mechanical parameters, with boosting methods again providing the most coherent representation.

The prediction of the sliding-surface depth z_s further reinforces the central role of soil profile geometry. In all models, the variables h_{Bd} and h_{Bu} consistently dominate the SHAP distributions, reflecting the deterministic nature of slip-surface formation in undrained conditions. Since the mobilized undrained shear strength is nearly uniform with depth, the position of the critical failure surface is governed primarily by geometric and stratigraphic features, rather than by hydraulic or strength parameters. The remarkable stability with which boosting algorithms recover this behavior strengthens the interpretation that z_s is predominantly a geometry-driven output.

The feature-importance analysis thus demonstrates that machine learning models achieve accurate predictions, but also maps to the physical mechanisms that govern undrained slope stability. The convergence of SHAP patterns across different algorithms confirms the robustness of the identified feature hierarchy, while differences among models relate mainly to the sharpness and stability of the attributions rather than to the underlying geotechnical interpretation. Taken together, these results illustrate how data-driven approaches, when carefully trained and interpreted, can faithfully capture the essential hydro-mechanical processes of undrained behavior, offering both reliable predictive capabilities and meaningful insight into the governing physics.

5. Hydraulic Simulations and Dataset

To build a representative dataset for the river element, several hydrological and morphological conditions were simulated by combining different values of key parameters. The main variables considered were: rainfall return period, drainage basin area, mean slope of the drainage basin, mean slope of the river, and mean slope of the floodplain.

The mean slope of the drainage basin was classified into three categories: low (L), medium (M), and high (H). Similarly, both the river and floodplain slopes were divided into three levels: L, M, H for rivers, and LL, MM, HH for floodplains, the latter referring to two possible inclination directions on the horizontal plane (see Deliverable 3.2).

A triangular hydrograph was adopted to describe the temporal distribution of discharge, while typical and constant values were used for the run-off coefficient and the

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

roughness coefficient. Using these inputs, the time of concentration (Tc) was estimated for each scenario, from which the corresponding rainfall intensity and peak discharge (Qmax) were derived through the rational method, including a correction factor based on the drainage area.

Each scenario was then simulated using the HEC-RAS hydraulic model. For each configuration, a specific breach point was defined on the floodplain—representing the location where overflow begins from the river channel. The model produced spatial outputs in geoTIFF format, representing the maximum water depth (m), maximum flow velocity (m/s), and water arrival time (hours) over the floodplain.

The floodplain was modeled as a 5×5 km^2 square domain, characterized by three slope configurations: flat (LL), 0.1% inclined (MM), and 1% inclined (HH) in both horizontal directions. The combination of all these parameters resulted in a dataset that captures a sufficient range of hydraulic and topographic conditions suitable for the training and validation of predictive models. The coding system used to describe river and floodplain slopes follows a simple and consistent logic. For example, the code MMM indicates a scenario with a medium river slope (M) and a medium floodplain slope (MM). Similarly, an intuitive naming convention was applied to identify each simulated scenario. For example, A2-MMM-TR010 refers to a configuration with a drainage basin area of 2 km^2 (A2), a medium river slope (M), a medium floodplain slope (MM), and a rainfall return period of 10 years (TR010).

5.1 Data Preparation

The hydraulic dataset comprises various spatial fields produced for different return times (TR) and for multiple hydrodynamic scenarios. Each scenario includes three raster variables—water depth, arrival time, and flow velocity—each represented as a two-dimensional grid. For a given hydraulic variable $k \in \{\text{Depth, Arrival, Velocity}\}$, the corresponding field at a specific return time TR_i is indicated by $X_k(\mathbf{x}, TR_i)$.

This notation expresses the value of the hydraulic quantity k at the grid location $\mathbf{x}$, encapsulating the spatial variability of the phenomenon. However, because the return times available in the original dataset differ among the scenarios, direct comparison or joint processing of the fields is not straightforward. Some simulations include dense temporal sampling, whereas others provide isolated TR values. To guarantee uniformity and enable consistent multi-scenario analysis, all fields are projected onto a common temporal grid with constant increment $\Delta TR = 10$. This produces a standardized and homogeneous set of temporal snapshots for all scenarios.

5.1.1 Interpolation between consecutive return times

To reconstruct missing fields on this standardized grid, spatial interpolation is performed between pairs of existing return times. Consider two consecutive return times TR_a and TR_B available for the same scenario. Any intermediate instant TR_{new} lying between them is associated with the normalized temporal weight $w = \frac{TR_{\text{new}} - TR_a}{TR_b - TR_a}$.

This weight determines the relative influence of the earlier and later states on the interpolated field. However, hydraulic rasters contain the intensity information and a

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

spatially evolving inundation pattern, thereby requiring interpolation of both geometry and magnitude.

To ensure a coherent evolution of the flooded area, each binary flood mask is converted into a signed distance representation $\phi(\mathbf{x}, TR_i)$. This function takes positive values inside the flooded region, negative values outside, and vanishes along the inundation boundary. The geometry of the flood area at the intermediate return time is obtained by a linear combination of the signed distance fields: $\phi_{\text{new}} = (1-w)\,\phi(TR_a) + w\,\phi(TR_b)$.

The resulting field ϕ_{new} defines an interpolated flood extent whose boundary evolves smoothly between the two known return times, preserving the overall shape while avoiding discontinuities that would arise from direct pixel-wise interpolation.

Once the spatial support of the inundation is defined, the hydraulic intensity can be interpolated within it. For each variable, the intermediate field is computed as $X_k(TR_{\text{new}}) = (1-w)\,X_k(TR_a) + w\,X_k(TR_b)$, meaning that the physical quantity increases or decreases smoothly over time between the two states. Cells outside the interpolated flood region are assigned a value of zero to maintain physical consistency with the inundation pattern.

5.1.2 Interpolation of scenario descriptors

Each simulation is also associated with a set of continuous parameters describing geometric, hydraulic, or structural conditions. These descriptors are represented by a parameter vector $\mathbf{p}(TR_i)$. To provide each scenario with a fully consistent description on the standardized temporal grid, the parameter vectors are interpolated according to $\mathbf{p}(TR_{\text{new}}) = (1-w)\,\mathbf{p}(TR_a) + w\,\mathbf{p}(TR_b)$.

This procedure ensures that each interpolated raster is accompanied by a corresponding set of physically meaningful metadata. Categorical descriptors, which cannot be interpolated numerically, are assigned based on the nearest available return time, thus preserving their interpretative coherence.

5.1.3 Resulting standardized dataset

By applying this interpolation framework to all return time intervals and all scenarios, the dataset is transformed into a temporally uniform collection of hydraulic maps $X_k(\mathbf{x}, TR_j)$ and parameter vectors $\mathbf{p}(TR_j)$, defined at the standardized return times $TR_j = TR_{min} + j\Delta TR$.

The resulting dataset exhibits a continuous and physically plausible temporal evolution of both the flooded area and the associated hydraulic variables. This standardized structure facilitates comparative analysis between scenarios, enables consistent parameter extraction, and provides a reliable basis for the spatial modeling procedures described in the following sections.

5.2 Spatial Modeling Workflow

After constructing a temporally standardized dataset, the analysis proceeds through three spatial modeling stages. Each stage transforms the raster fields into a complementary representation that captures geometric, hydraulic, or structural features of the

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

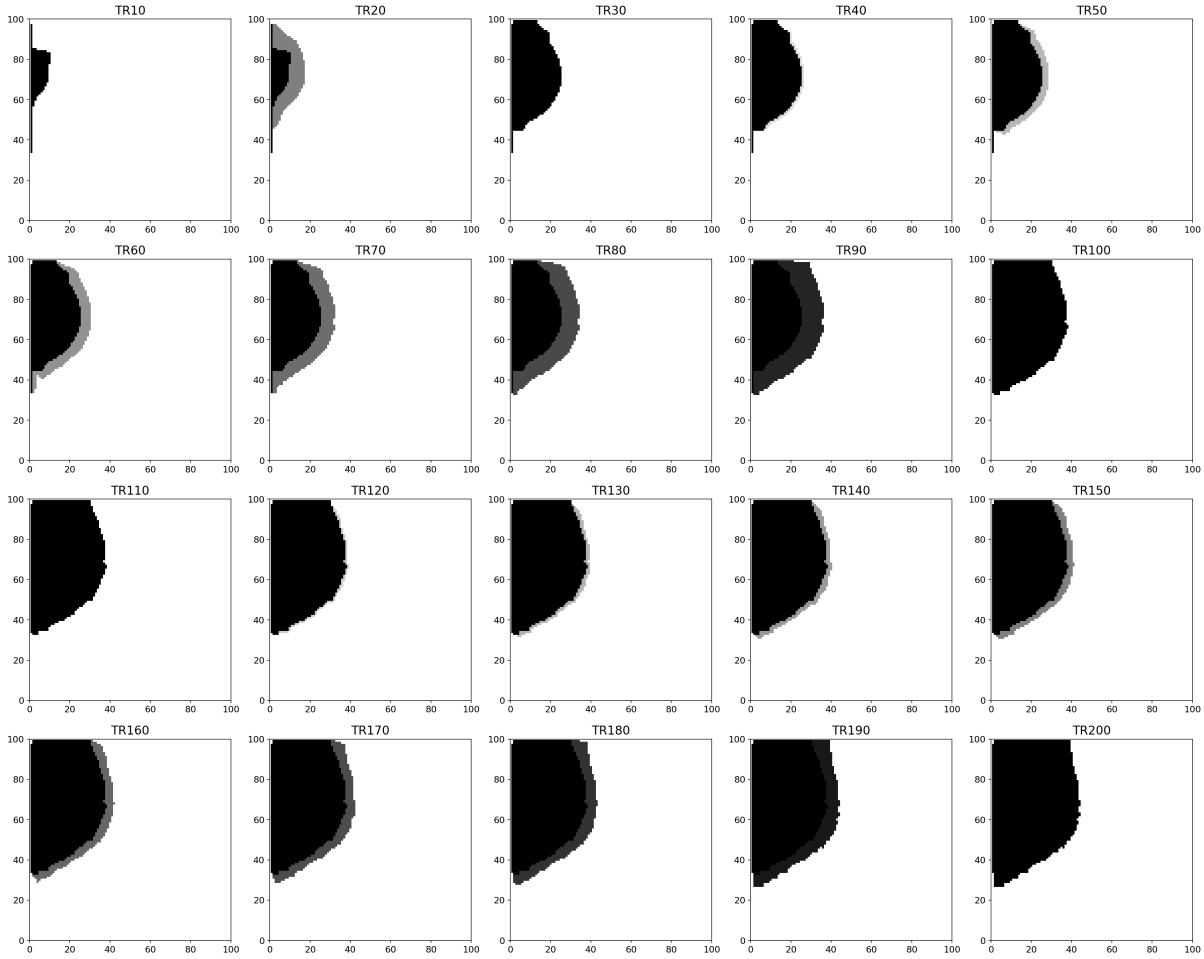


Fig. 51: Interpolated flooded area for the 20 return times (TR10–TR200). TR10, TR30, TR100, and TR200 represent the original flooded area, while all intermediate flooded area are generated through signed distance function (SDF)–based interpolation. The SDF interpolation preserves boundary coherence and captures gradual morphological changes.

system. Together, these stages provide a structured framework from which qualitative and quantitative assessments can be performed.

5.2.1 Terrain slope field

The first stage involves the construction of a synthetic slope field representing the large-scale topographic inclination of the study domain. A numerical matrix is generated to reproduce a planar elevation surface whose intensity progressively decreases along a diagonal direction, corresponding to a uniform downslope profile. Each cell in the grid, located at the coordinates $\mathbf{x} = (i, j)$, is assigned a value determined by a linear function of its relative position along this diagonal, typically expressed as $Z(i, j) = Z_0 - s, \frac{i+j}{2}$, resulting in a continuous gradient extending from the highest to the lowest point of the domain.

The steepness of the simulated terrain is controlled by a dimensionless slope coefficient s, which governs the rate at which the elevation decreases in the domain.

Three representative configurations are adopted to span a range of possible and rep-

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

resentative topographic conditions:

- a *low-slope* condition with coefficient $s = 0.0$, representing an essentially horizontal floodplain,
- a *medium-slope* condition with coefficient $s = 0.001$,
- a *high-slope* condition with coefficient $s = 0.01$, generating a significantly inclined terrain surface.

The resulting raster provides a two-dimensional approximation of the ground elevation and serves both as a contextual base layer for interpreting hydraulic patterns and as a simplified geometric model for downstream analyses involving flow direction or failure initiation.

5.2.2 Breach initiation zone

The second stage identifies the spatial location where the breach or overtopping event initiates along the left boundary of the domain. This operation is carried out by examining the depth field associated with each scenario. The algorithm inspects the first ten columns of the depth matrix and searches for the first non-zero entry, which indicates the presence of a meaningful hydraulic or topographic value.

The direction of the search depends on the prescribed breach configuration. For a "top-left" configuration, the scan begins at the upper-left corner and proceeds downward. For a "bottom-left" configuration, the scan begins at the lower-left corner and proceeds upward. Once the first valid cell is detected, a square window of fixed size (8×8 cells) is centered on this location. All cells inside the window are set to one in a new binary mask, while all remaining cells are set to zero.

This mask represents the breach initiation zone and isolates the region associated with the onset of structural failure. In addition, the local depth value at the breach point is recorded, providing a quantitative indication of the elevation or water level at the moment of breach initiation. This information is cucial to evaluate the hydraulic forcing and geometric conditions that can contribute to failure.

5.2.3 Segmentation of hydraulic variables

The third stage involves segmenting the depth, velocity, and arrival-time fields into discrete categories. This conversion transforms continuous hydraulic values into integer classes that delineate regions of similar intensity, facilitating both visualization and subsequent machine-learning workflows.

Each variable is divided into four classes based on explicit physical thresholds. For the depth field (in meters), the segmentation is defined as follows:

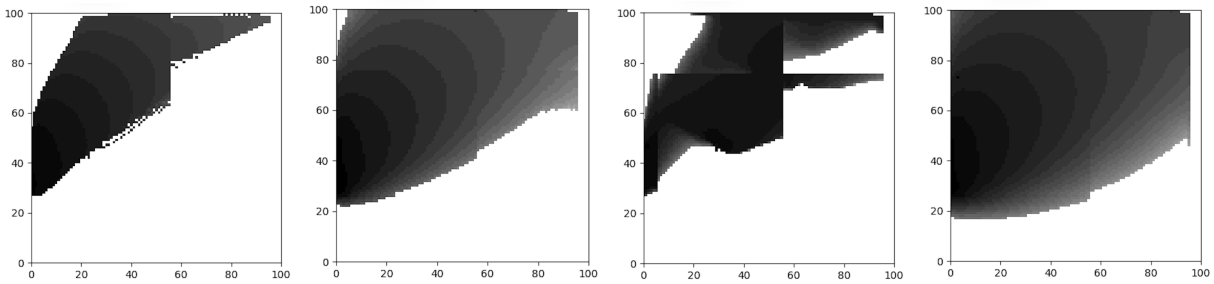
- class 1: values < 1 m,
- class 2: values between 1 m and 2 m,
- class 3: values between 2 m and 3 m,

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

- class 4: values > 3 m.

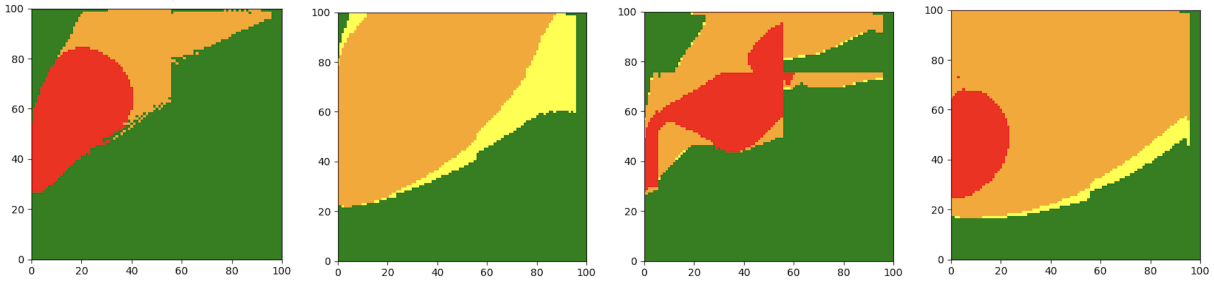
The velocity field (in m/s) adopts the same threshold structure:

- class 1: values < 1 m/s,
- class 2: values between 1 and 2 m/s,
- class 3: values between 2 and 3 m/s,
- class 4: values > 3 m/s.



(a) (b) (c) (d)

Fig. 52: Flooded-area fields for four representative cases. Each plot illustrates the spatial distribution of depth water across the flooded area, highlighting variations in extent and morphology driven by differences in the hydro-mechanical input conditions.



(a) (b) (c) (d)

Fig. 53: Semantic segmentation of the flooded-area fields shown. The maps display the corresponding categorical hazard classes, emphasizing how the segmentation model partitions the domain into discrete flood-severity zones.

For the arrival-time field (in seconds), the classification is intentionally inverted to reflect the physical interpretation that faster propagation corresponds to higher intensity. Thus, the segmentation is as follows:

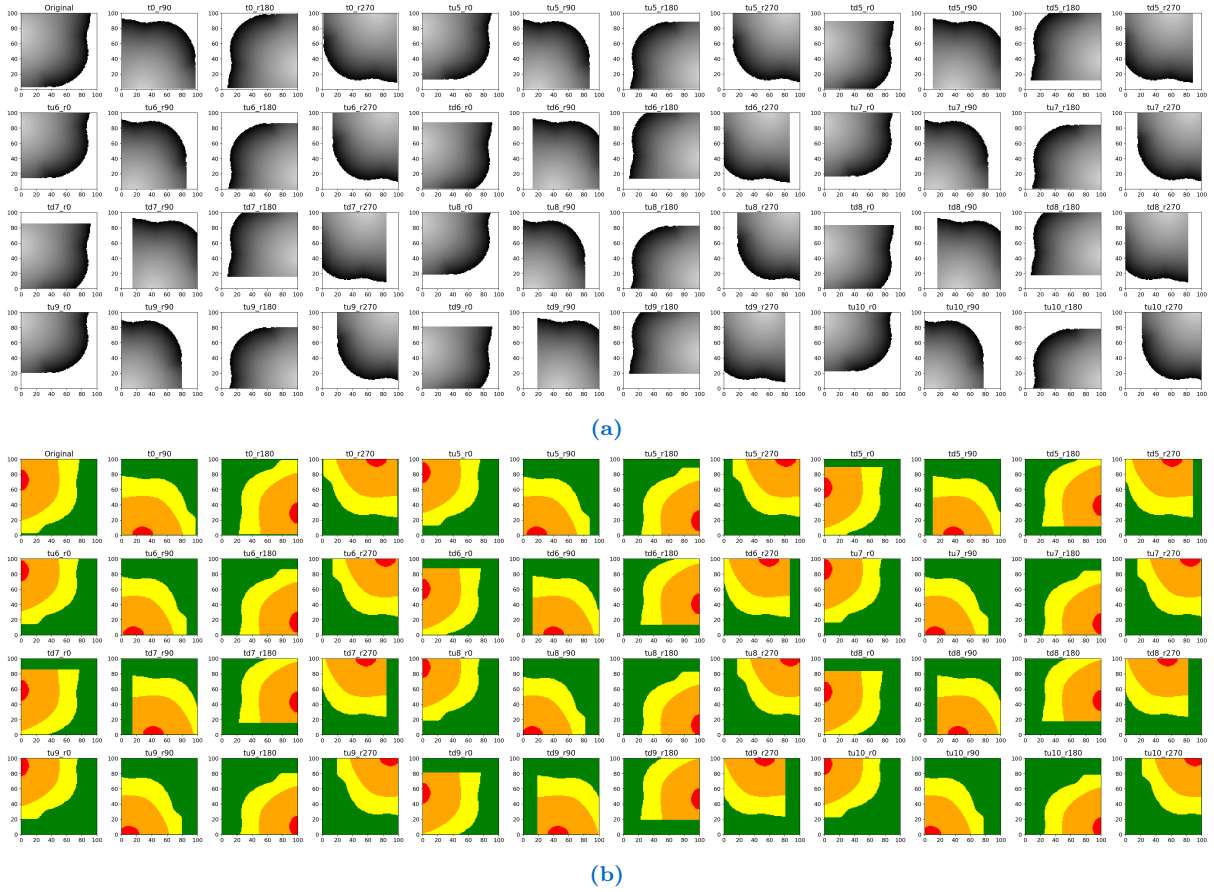
- class 4: values < 10 s,
- class 3: values between 10 s and 20 s,
- class 2: values between 20 s and 30 s,
- class 1: values > 30 s.

The output of this process is an integer-valued raster with values ranging from 1 to 4, where an increase in class number corresponds to an increase in hydraulic intensity. A dedicated colormap—typically using white, green, yellow, orange, and red—is applied to visually distinguish the classes and highlight spatial gradients.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

5.2.4 Integrated spatial framework

The three components generated in this workflo—the slope field, the breach initiation mask, and the segmented hydraulic variables—constitute a coherent spatial representation of the hydrodynamic system. The slope field encodes the geometric background, the breach mask identifies the triggering location of structural failure, and the segmented rasters classify hydraulic intensity across the domain. Because all scenarios are processed using the same thresholds and criteria, the resulting spatial layers are fully comparable across case studies. This consistent and interpretable structure supports both qualitative inspection and quantitative modeling and provides a robust foundation for subsequent numerical simulations or machine-learning analyses.



(a)

(b)

Fig. 54: Augmented flooded-area maps generated through systematic spatial transformations, including horizontal and vertical shifts as well as rotations at multiple angles (a). Corresponding augmented semantic segmentation maps derived from the transformed flooded-area fields (b). These augmentations expand the variability of the training dataset while preserving the underlying hydrodynamic structure. The segmentation retains class consistency across all augmentations, ensuring that the expanded dataset remains coherent for supervised learning tasks.

5.3 Data Augmentation of Hydraulic Maps

To enhance the robustness and generalization capability of the predictive models, a systematic data augmentation procedure was applied to all geoTIFF matrices produced by the hydraulic simulations. This procedure expands the dataset by generating multiple

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

geometric variants of each spatial field, while preserving the underlying physical patterns associated with flood propagation, flow velocity, and water depth. The objective is to expose the learning model to a wider set of spatial configurations that emulate slight perturbations or alternative orientations of the hydraulic processes.

The implemented augmentation strategy consists of two main groups of transformations: *vertical shifts* and *planar rotations*.

5.3.1 Vertical Shifts

Vertical shifts simulate small displacements of the hydraulic fields along the longitudinal direction of the floodplain. Each raster matrix is shifted upward or downward by a fixed number of rows. For every original geoTIFF, the following displacements were applied:

$$\Delta r \in \{15,\ 16,\ 17,\ 18,\ 19,\ 20\} \text{ rows.}$$

For each value of Δr, two complementary operations were generated:

- **Upward shift** (tuΔr): the top Δr rows are removed and the bottom of the matrix is padded with zero-valued rows, simulating an upstream translation of the hydraulic pattern.
- **Downward shift** (tdΔr): the bottom Δr rows are removed and replaced with zero-valued rows at the top, emulating a downstream displacement.

These operations maintain the internal spatial structure of hydraulic features (e.g., deep-water regions, high-velocity channels), while allowing the learning model to acquire invariance to small positional offsets.

5.3.2 Planar Rotations

To incorporate orientation invariance—particularly useful given that hydraulic patterns do not possess a unique preferential direction—each original and vertically shifted matrix was then rotated by the four canonical angles:

$$\theta \in \{0^\circ,\ 90^\circ,\ 180^\circ,\ 270^\circ\}.$$

Rotations were applied after shifting, thus producing a complete set of geometric variants for each transformation type. This significantly increases the diversity of spatial configurations while ensuring that the physical consistency of the hydraulic fields is preserved.

5.3.3 Output Structure and Metadata

Each augmented raster is saved as a separate `.tif` file, following a systematic naming convention:

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

`<name>_<shift>_<angle>.tif`

where `<shift>` identifies the applied shift transformation (`t0`, `tu15`, `td20`, etc.), and θ denotes the rotation angle. For each augmented image, a corresponding entry is appended to the metadata table (`input.csv`).

For each original raster, the augmentation pipeline produces: 1 (original) + 6 upward shifts + 6 downward shifts, each further combined with 4 rotation angles, resulting in a comprehensive and continuously populated dataset. The resulting hydraulic maps strengthen the ability of predictive models to generalize across positional, orientational, and morphological variability inherent in hydrodynamic processes.

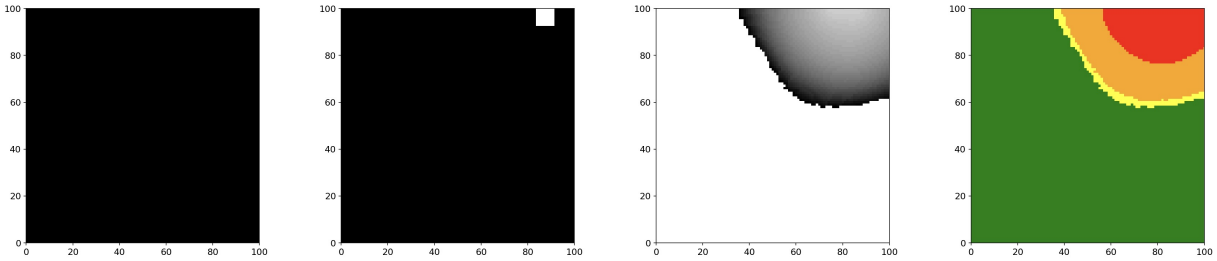


Fig. 55: Inputs and output for one representative scenario: (a) the slope plane, (b) the breach-location mask, (c) the flooded-area map, and (d) the corresponding segmented hazard map.

6. Semantic Segmentation Models and Evaluation Framework

In this section, we present the architecture of the four state-of-the-art semantic segmentation models employed in this study, describe the quantitative metrics used to evaluate their performance under a 10-fold cross-validation scheme, and provide an extended discussion of the obtained results. The focus is to understand the numerical differences among the models and the underlying reasons for their distinct behavior when segmenting hydrodynamic raster fields such as arrival time, water depth, and flow velocity.

6.1 Model Architectures

Two architectures were developed to investigate how hydrological information can be exploited in different input formulations. Both models underwent an extensive hyperparameter search, exploring broad ranges of architectural configurations and training parameters to identify the best-performing settings.

6.1.1 Feature-Based Generator

The first model is a fully connected generator that maps a vector of hydrological and geomorphological descriptors into a categorical raster of size 100 × 100 100×100. This formulation assesses how much spatial structure can be inferred solely from aggregated hydrological features.

The hyperparameter exploration covered variations in depth, width, regularization, and training dynamics. The number of hidden layers varied between *2* and *4*, while the

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

width of each layer ranged from *128* to *1024* units. Activation functions were tested across *ReLU*, *GELU*, and *LeakyReLU*. Dropout rates were explored in the interval *0.0–0.5*. The weight decay values were evaluated in *0*, $1e^{-5}$, and $1e^{-4}$, and the learning rate was swept over the continuous range *5e-4–5e-3*. Batch sizes of 16, 32, and 64 were compared.

The optimal model consists of a three-layer multilayer perceptron with 256–512 units per layer, trained with a learning rate $\eta = 10^{-3}$. Despite lacking spatial inductive biases, the model manages to reproduce broad hydrodynamic patterns, showing that global descriptors capture part of the underlying spatial distribution of flood intensity.

6.1.2 Multimodal Generator

The second architecture extends the previous formulation by incorporating two synthetic spatial priors: a slope field describing the dominant topographic inclination and a breach-location mask identifying the likely point of structural failure. Both matrices are flattened and fused with the tabular descriptors to create a joint representation that leverages both global hydrological information and simplified spatial cues.

The hyperparameter search explored several fusion mechanisms, including direct concatenation, linear projection onto a shared embedding space, and late fusion through parallel MLP branches. The embedding depth varied between 1 and 3 layers, whereas embedding width ranged from 256 to 2048. Spatial-channel weighting factors were evaluated in $\{0.5, 1.0, 1.5, 2\}$. Activation functions were tested across ReLU and GELU, dropout was explored in 0.0–0.3, and learning rates were sampled from $\{5e^{-4}, 2e^{-3}\}$. Batch sizes of 16 and 32 were also investigated.

The best configuration uses two embedding layers of 1024 units with ReLU activation, direct concatenation of spatial and tabular channels, and a learning rate $\eta = 10^{-3}$. The inclusion of spatial priors significantly improves the directional coherence and positional accuracy of the generated maps, reducing the ambiguity inherent in the feature-only formulation.

6.2 Evaluation Metrics

To evaluate the performance of the segmentation models, we employed Accuracy, Mean Intersection over Union (mIoU), Precision, Recall, and F1-score. Each metric captures a different property of the predictions, and together they provide a comprehensive evaluation of segmentation quality.

Accuracy Accuracy measures the proportion of correctly classified pixels over the entire image:

$$\text{Accuracy} = \frac{\sum_{c=1}^{N} TP_c}{\sum_{c=1}^{N} (TP_c + FP_c + FN_c)} \tag{3}$$

It provides a global correctness, but may be misleading in the case of imbalanced classes. In our case, this is particularly relevant because the green class (representing lower hazard zones) typically occupies a much larger spatial extent than higher-intensity classes.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Intersection over Union (IoU) and mIoU The IoU for a given class quantifies the spatial overlap between predicted and ground-truth regions:

$$\text{IoU}c = \frac{TPc}{TP_c + FP_c + FN_c} \tag{4}$$

The mean IoU (mIoU) is the average across all classes:

$$\text{mIoU} = \frac{1}{N}\sum_{c=1}^{N} \text{IoU}_c \tag{5}$$

mIoU is a stringent metric: even small misalignments along boundaries or slight over/under-estimations of affected areas can considerably lower the score. Because flood-related raster fields often contain thin gradients and sharp transition zones, mIoU serves as a robust indicator of spatial consistency of a model.

Precision Precision measures the proportion of predicted positive pixels that are correct:

$$\text{Precision}c = \frac{TPc}{TP_c + FP_c} \tag{6}$$

High precision indicates that the model avoids over-prediction. In flood segmentation, this is crucial: overestimating hazardous zones may unnecessarily bias risk assessments.

Recall Recall measures the proportion of true positive pixels that the model successfully detects:

$$\text{Recall}c = \frac{TPc}{TP_c + FN_c} \tag{7}$$

High recall indicates that the model reliably identifies hazardous regions. Underestimation is often more detrimental in hazard mapping, making recall a central metric.

F1-score The F1-score represents the harmonic mean of Precision and Recall:

$$\text{F1}c = \frac{2 \cdot \text{Precision}c \cdot \text{Recall}c}{\text{Precision}c + \text{Recall}_c} \tag{8}$$

Averaged across classes, it provides a balanced measure of both boundary accuracy and prediction reliability. Because many hydrodynamic structures have irregular shapes and thin extensions, the F1-score is particularly relevant for evaluating segmentation of arrival fronts and peak-velocity regions.

6.3 Discussion of Results

The comparison between the feature-based generator and the multimodal generator under the 10-fold cross-validation scheme reveals clear performance differences that can be directly linked to the availability of spatial information in the input representation.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Across all evaluation metrics, the multimodal architecture consistently outperforms the feature-only formulation, demonstrating the crucial role of even minimal spatial priors when segmenting hydrodynamic raster fields.

Tab. 20
Mean and standard deviation of segmentation metrics over 10-fold cross-validation for the optimized feature-based and multimodal models.

Model	Accuracy	mIoU	F1-score	Precision	Recall
Feature-Based Generator	0.865 ± 0.021	0.602 ± 0.027	0.654 ± 0.023	0.671 ± 0.026	0.640 ± 0.028
Multimodal Generator	$\mathbf{0.903 \pm 0.017}$	$\mathbf{0.654 \pm 0.024}$	$\mathbf{0.701 \pm 0.020}$	$\mathbf{0.718 \pm 0.022}$	$\mathbf{0.685 \pm 0.023}$

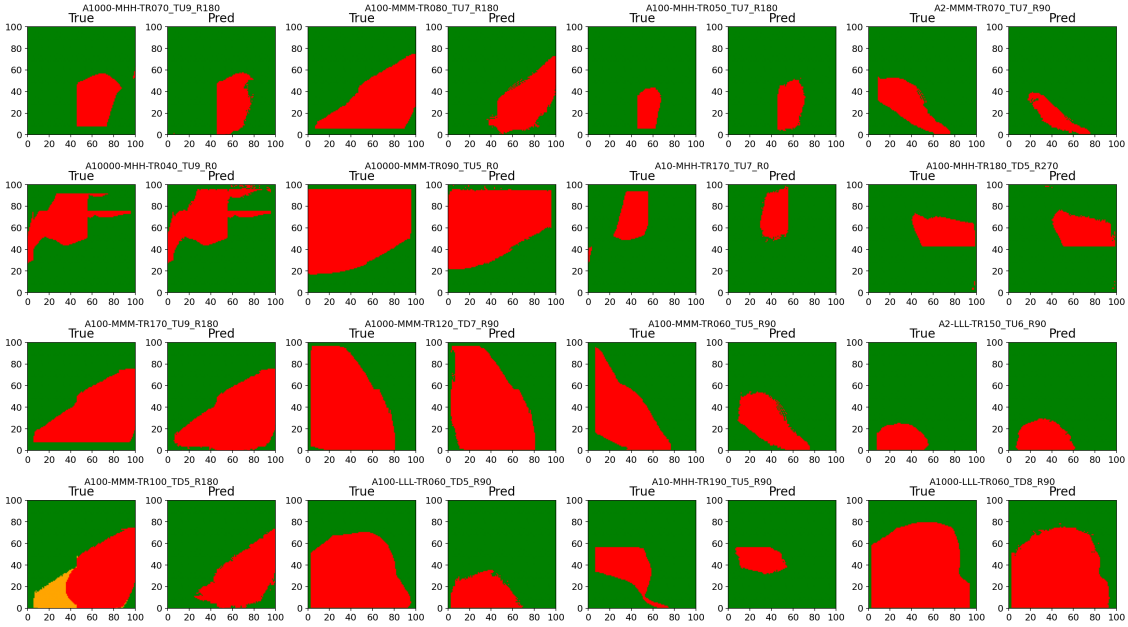


Fig. 56: Comparison between ground-truth segmentation maps ("True") and the corresponding predictions ("Pred") generated by the multimodal network across different samples. Each pair illustrates the model's ability to reproduce the spatial extent and morphology of the flooded areas under varying hydro-mechanical conditions and augmentations.

The feature-based generator, relying solely on aggregated hydrological and geomorphological descriptors, achieves competitive accuracy and produces coherent large-scale patterns. This indicates that global descriptors indeed encode part of the underlying physical structure governing flood propagation. However, the model exhibits systematic limitations when reconstructing fine-scale spatial details, especially along sharp transition zones such as arrival fronts or regions of rapidly varying water depth. These shortcomings are reflected in its lower mIoU and F1-scores, which highlight difficulties in aligning class boundaries and capturing thin or irregular hazardous areas. The observed variability across folds further suggests that, without explicit spatial cues, the model is more sensitive to the composition of the training subsets.

The multimodal generator addresses these limitations by incorporating simple but informative spatial priors—namely the slope field and breach-location mask. These additional inputs provide directional and positional context that the feature-only architecture

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

cannot infer from global descriptors alone. As a result, the model demonstrates substantial improvements in both spatial precision and class discrimination, reflected in higher mIoU, F1-score, and precision–recall balance. The inclusion of slope information helps guide the model toward physically plausible flow directions, while the breach mask anchors the predicted hazard patterns to realistic spatial origins. Together, these priors mitigate ambiguities in regions where hydrodynamic variables exhibit strong spatial gradients.

The lower standard deviations observed across folds further indicate that the multimodal generator generalizes more reliably than the feature-based variant. This suggests that even coarse spatial information stabilizes the learning process and reduces the model's dependence on the specific characteristics of each training split.

Overall, the results demonstrate that purely feature-based formulations, while capable of capturing broad hydrodynamic trends, cannot fully reproduce the spatial complexity of flood-related rasters. Incorporating lightweight spatial priors significantly enhances predictive accuracy and structural coherence, confirming that flood segmentation tasks benefit from architectures capable of leveraging both global descriptors and spatially grounded information. These findings align with recent evidence in environmental modeling, where hybrid representations frequently outperform strictly tabular or strictly spatial approaches, especially when fine-scale delineation is required.

7. Conclusions

This deliverable presented the results of the design, development and evaluation of AI models for hydrogeological risk assessment within the SAFE-LAND project. The implemented workflow—ranging from data preparation to model training and interpretation—confirmed the effectiveness of AI methods, with ensemble and deep learning models showing the most robust and accurate performance.

The use of explainable AI allowed us to identify the most influential predictors, improving the transparency and interpretability of the susceptibility estimates. The results provide a reliable basis for the next project activities.

References

[1] W. Fellenius, "Calculation of the stability of earth dams," in *Proc. of the second congress on large dams*, vol. 4, 1936, pp. 445–463.

[2] A. W. Bishop, "The use of the slip circle in the stability analysis of slopes," *Geotechnique*, vol. 5, no. 1, pp. 7–17, 1955.

[3] N. R. U. Morgenstern and V. E. Price, "The analysis of the stability of general slip surfaces," *Geotechnique*, vol. 15, no. 1, pp. 79–93, 1965.

[4] E. Spencer, "A method of analysis of the stability of embankments assuming parallel inter-slice forces," *Geotechnique*, vol. 17, no. 1, pp. 11–26, 1967.

[5] J. K. Mitchell and K. Soga, *Fundamentals of soil behavior*. John Wiley & Sons, 2005, vol. 3, p. 558.

[6] M. Budhu, *Soil mechanics fundamentals*. John Wiley & Sons, 2015.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

[7] E. E. Alonso et al., "A constitutive model for partially saturated soils," *Géotechnique*, vol. 40, no. 3, pp. 405–430, 1990.

[8] D. G. Fredlund and H. Rahardjo, *Soil mechanics for unsaturated soils*. John Wiley & Sons, 1993.

[9] S. K. Vanapalli et al., "Model for the prediction of shear strength with respect to soil suction," *Canadian Geotechnical Journal*, vol. 33, no. 3, pp. 379–392, 1996.

[10] M. T. V. Genuchten, "A closed-form equation for predicting the hydraulic conductivity of unsaturated soils," *Soil Science Society of America Journal*, vol. 44, no. 5, pp. 892–898, 1980.

[11] D. Q. Li et al., "Efficient and consistent reliability analysis of soil slope stability using both limit equilibrium analysis and finite element analysis," *Applied Mathematical Modelling*, vol. 40, no. 9-10, pp. 5216–5229, 2016.

[12] F. Vahedifard et al., "Effective stress-based limit-equilibrium analysis for homogeneous unsaturated slopes," *International Journal of Geomechanics*, vol. 16, no. 6, p. D4016003, 2016.

[13] P. Harrington, *Machine Learning in Action*. Manning Publications Co., 2012.

[14] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.