

DATA-DRIVEN MODELING OF SOIL WATER CONTENT USING DIFFERENT RAINFALL PREDICTORS

AMIN GHEZAVATINEZHAD^(*), MASSIMILIANO BORDONI^(*), ALESSIA GIAROLA^(*),
VALERIO VIVALDI^(*), ANTONIO GAMBARANI^(*) & CLAUDIA MEISINA^(*)

^(*)University of Pavia - Department of Earth and Environmental Sciences - Pavia, Italy
Corresponding author: amin.ghezavatinezha01@universitadipavia.it

EXTENDED ABSTRACT

L'umidità del terreno è una componente cruciale dei processi idrogeologici e geotecnici. In particolare, è molto importante per la stabilità dei versanti e per l'innescio delle frane, specialmente per le frane superficiali che sono molto influenzate dalla quantità di acqua presente nel suolo prima di un forte evento piovoso.

La valutazione corretta delle condizioni che causano le frane superficiali e la possibile creazione di un sistema di allerta rapida per la loro previsione dipendono molto da una stima del Contenuto d'Acqua nel Terreno (SWC), specialmente quando questo è vicino alla saturazione. Per affrontare queste sfide, i metodi statistici basati sui dati, come gli algoritmi di *Machine Learning* (ML), sono diventati soluzioni efficaci per stimare e mappare l'umidità del suolo.

Lo scopo di questo studio è prevedere la quantità di acqua nel suolo usando solo indicatori basati sulla pioggia, senza usare i dati della temperatura dell'aria o dell'evaporazione, che spesso non sono disponibili nel normale monitoraggio geotecnico. L'obiettivo è valutare quanto bene i modelli *Random Forest* (RF) possono prevedere SWC in due diversi tipi di gestione del suolo a profondità di 10, 50 e 90 cm, concentrandosi in particolare sulla previsione dei valori vicini alla saturazione completa.

Il sito monitorato (SMV) si trova nella zona collinare centrale dell'Oltrepò Pavese. Questo versante presenta un terreno prevalentemente argilloso. Un monitoraggio idrologico continuo è stato fatto dal 2021 al 2024 su tutto il suolo del sito, per un totale di 4 punti di monitoraggio. Tra questi, i punti esaminati hanno riguardato i punti con gestioni "Control" e "Rolling".

Le serie temporali di umidità del suolo analizzate sono state misurate a 10, 50 e 90 cm dal piano campagna in entrambi i punti di monitoraggio, per testare la capacità del modello di ricostruire le dinamiche dell'umidità del suolo in diversi strati. È stato applicato un metodo di somma a finestra mobile per ottenere un gruppo di Indici di Pioggia Cumulativa (CRI), che riflettono il fatto che il contenuto d'acqua nel suolo in qualsiasi momento dipende dalla pioggia precedentemente caduta. Per considerare la variazione stagionale dell'evapotraspirazione (che è più alta in estate che in inverno), il Mese dell'Anno è stato incluso come indicatore categorico nei dati. Una volta scelti gli indicatori, un algoritmo *Random Forest* (RF) è stato usato per prevedere il contenuto d'acqua nel suolo. L'intero gruppo di dati è stato diviso in un set di training per sviluppare il modello e un set di test per la sua validazione.

Il modello ha ottenuto un $R^2 > 0.98$ a tutte le profondità studiate (10, 50 e 90 cm), confermando la capacità dell'algoritmo di riprodurre con precisione le tendenze del contenuto d'acqua nel suolo. Tuttavia, le prestazioni sono diminuite quando il modello è stato applicato ai nuovi dati del 2024, mostrando chiari limiti di previsione. Per entrambi i siti *Control* e *Rolling*, la variabile 'Mese' ha avuto l'impatto maggiore sul modello a tutte le profondità (10, 50 e 90 cm). Nel sito *Rolling*, anche a profondità minori (10 e 50 cm), gli indicatori di pioggia a medio e lungo termine come 'Pioggia in 30 giorni' e 'Pioggia in 90 giorni' hanno mostrato un'importanza maggiore rispetto all'indice della pioggia immediata 'Pioggia in 1 giorno'. Al contrario, per il sito *Control*, gli indici a lungo termine ('Pioggia in 90 giorni' e 'Pioggia in 60 giorni') hanno mostrato un'importanza maggiore rispetto all'indicatore a breve termine 'Pioggia in 1 giorno', anche alla profondità minore (10 cm).

In generale, i risultati della validazione esterna sottolineano l'importanza fondamentale di allenare i modelli basati sui dati con serie temporali abbastanza lunghe da includere tutte le possibili condizioni meteorologiche nel sito di studio. La ricerca dimostra che il modello identifica le condizioni critiche vicine alla saturazione (il 10% dei giorni più umidi) con un basso margine di errore, testimoniato da un indice RMSE che va da 0.020 a 0.057 m³/m³. Questi risultati suggeriscono che il modello *Random Forest* è affidabile per prevedere le condizioni che potrebbero causare frane superficiali, anche senza includere la temperatura dell'aria o altre variabili meteorologiche più difficilmente misurabili.

ABSTRACT

Soil moisture is a crucial component of hydrological and geotechnical processes, especially for shallow slope failures. When adopting a physically-based modeling approach for the prediction of hydrological variables, it can however be challenging to gather the necessary data and find what parameters to use on a large scale. The goal of this study is therefore to adopt a data-driven approach to predict the amount of water in the soil using only rainfall based predictors.

The study used a Random Forest algorithm that was trained using only cumulative rainfall indices and time markers. The analysis was performed independently for the two distinct soil management practices (Control and Rolling) at depths of 10, 50, and 90 cm. A randomized validation test was done and the results show that the prediction of water content has very high internal consistency ($R^2 > 0.98$).

However, the predictions for the year 2024, which had not been used for the training of the model, showed clear physical limits. The reason of this poor performance is related to meteorological conditions on January-April 2024 period, which differed from the ones present in training time series.

Consequently, the results highlight the importance of reconstructing data-driven models with a training time-series long enough to consider all the possible meteorological conditions. Nevertheless, the results show that the model can find important near-saturation conditions with a small margin of error. This means that this Random Forest model can be reliable in predicting conditions that could lead to shallow landslide triggering, even while neglecting air temperature data.

KEYWORDS: soil water content, machine learning, random forest, shallow landslides, Oltrepò Pavese, hydrological monitoring

INTRODUCTION

Soil moisture is a crucial component of hydrological and geotechnical processes. In particular, it is crucial for slope stability and the triggering of landslides (ZHANG *et alii*, 2024; WILSON *et alii*, 2005), especially for shallow slope failures that are strongly influenced by the amount of water present in the soil before a strong rainfall event (HALTER *et alii*, 2025). The successful assessment of shallow landslides triggering conditions and the possible creation of an early warning system tool for their prediction depend strongly on a correct estimation of Soil Water Content (SWC), especially when the soil is close to saturation.

However, it is still challenging to model soil moisture in high spatial and temporal resolutions. It can also be complex to gather the necessary inputs to apply physically-based models for the prediction of hydrological variables, especially at a large scale (BAKHSHIAN *et alii*, 2025). To tackle these challenges, data driven methods, such as Machine Learning (ML) algorithms,

have emerged as effective solutions for soil moisture estimation and mapping (KHOSE & MAILAPALLI, 2024; HUSSAIN *et alii*, 2025). For instance, Random Forest (RF) algorithms have demonstrated their capability to effectively manage non linear relationships between environmental variables and soil hydraulic response under different meteorological conditions (ZHANG *et alii*, 2024). It is also becoming more important for practical applications, to develop reliable solution to predict soil moisture using only easy-to-be-obtained data, such as rainfall (NGUYEN *et alii*, 2022).

The aim of this study is to predict the amount of water in the soil using only rainfall-based predictors without relying on air temperature or evaporation data, which are frequently unavailable in standard geotechnical monitoring.

The goal is to evaluate how effectively Random Forest (RF) models can predict SWC in two different types of soil management at depths of 10, 50, and 90 cm, focusing especially on the prediction of saturation degree values close to complete saturation.

MATERIALS AND METHODS

Test site

The monitored test site (SMV site) is located in central Oltrepò Pavese hilly zone (Figure 1, Table 1). The hillslope is characterized by slope angle in the range 5–15°, with a northwest to west aspect. The land use is represented by a vineyard of approximately 14 years old, trained to unilateral Guyot with vertical shoot positioning (VSP), and spaced at 2.2×1.0 m for a planting density of about 4545 vines ha⁻¹. From the geological point of view, the bedrock is composed by turbiditic sandstones, marls, and marly limestones. Above the bedrock, a colluvial soil, with thickness between 70 and 100 cm, is present. This soil is classified as a Calcaric Cambisols, with a predominantly clayey texture throughout the entire soil profile (8% sand, 31% silt, 61% clay; CUNIAL *et alii*, 2025).

At the test-site, soil located in the interrow between vineyards is managed in four different ways:

- Control: spontaneous grass cover managed by mowing;
- Rolling: seeded cover crop based on cereals, flattened without tillage;
- Green Manure of seeded cover crop based on cereals;
- Sub-row Mulching of seed cover crop based on cereals.

Continuous hydrological monitoring was carried out from 2021 to 2024 on all the soil of the test site, for a total of 4 monitoring points (CUNIAL *et alii*, 2025). However, for this study, the examined monitoring trends pertained only to Control and Rolling sites, facilitating a comparison of soil management practices devoid of tillage operations.

Soil water content was measured between 0 and 90 cm in depth, every 10 cm, through a Sentek Drill & Drop soil moisture probe. These sensors acquired data every 10 minutes, which were then averaged to a daily temporal resolution. The analyzed time series of water content measurements were acquired at 10, 50 and

Slope elevation (m a.s.l.)	257 - 287
Slope angle (°)	5 - 15
Slope aspect	NW - W
Bedrock lithology	Turbiditic sandstones, marls, and marly limestones
Soil depth (m)	0.6 - 1.0
Soil texture	8% of sand, 31% of silt, 61% of clay

Tab. 1 - Main features of the test site

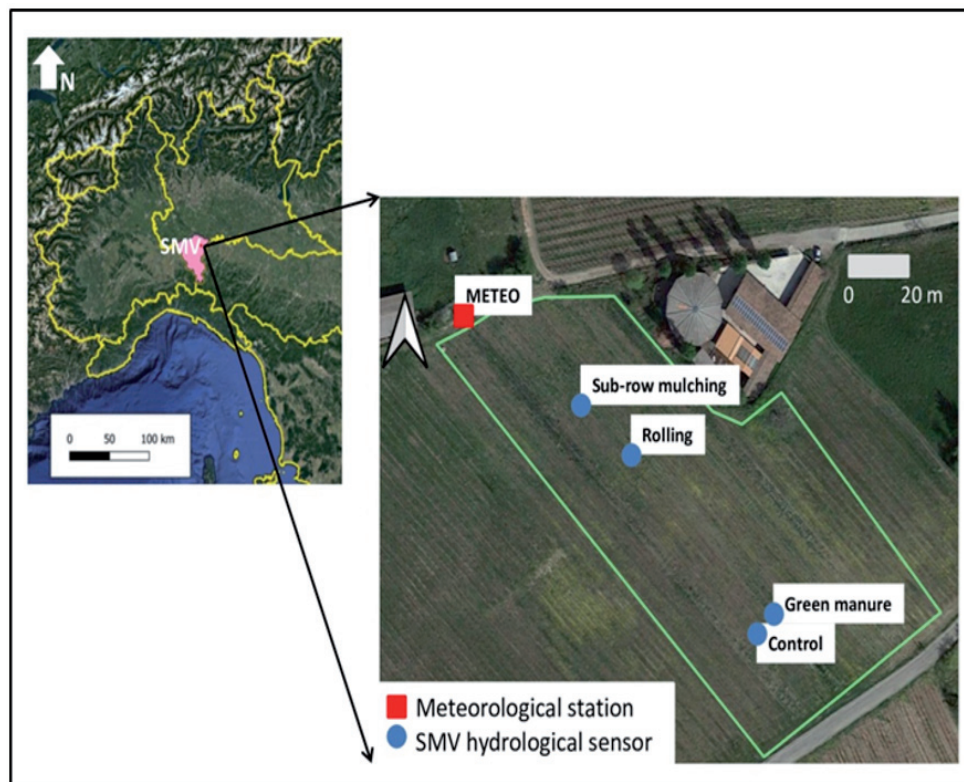


Fig. 1 - Location of the test-site and soil management practices

90 cm at both monitoring points, in order to test the capability of the model in reconstructing soil moisture dynamics across different soil horizons. Daily meteorological data were provided by a meteorological station (NetsensMeteoSense 4.0) located at the test site, which recorded, rainfall, air temperature and humidity, wind speed, atmospheric pressure and total solar radiation.

RF model for the prediction of soil water content

The dataset consists of daily time series of volumetric soil water content (SWC, m^3/m^3) gathered from the automated monitoring stations installed at the test-site. Measurements were collected from April 2021 to April 2024, recording soil water content at three different depths: 10 cm, 50 cm, and 90 cm. The analysis have been done in this study was performed independently for the two distinct soil management practices (Control and Rolling).

The developed data-driven method was built considering only rainfall data (NGUYEN *et alii*, 2022). This study deliberately omitted air temperature, relative humidity, and solar radiation inputs to evaluate the constraints of rainfall-only monitoring. This stands in contrast to complex physically-based models that typically require such variables to calculate Penman-Monteith evapotranspiration.

The model predictors consists of daily rainfall measurements along with a set of cumulative rainfall amounts for the days preceding a given date. A moving window summation method was applied to derive a set of Cumulative Rainfall Indices (CRI) reflecting the fact that soil water content at any given time is a function of antecedent precipitation. Consequently, these indices were categorized to reflect different duration of hydrological memory:

- Short-term memory (1–3 days): Represents the immediate response of the topsoil layers to infiltration events.

- Medium-term memory (7–30 days): Serves as a proxy for water percolation and storage dynamics within the intermediate soil horizons over time
- Long-term memory (60–90 days): Accounts for deep percolation lags and the base saturation level of the slope, both of which are critical factors for deep seated instability.

To account for seasonal variation in evapotranspiration (which is higher in summer than in winter), the Month of the Year was included as a categorical predictor in the feature space.

Once the predictors were selected, a Random Forest (RF) algorithm was used for soil water content prediction. This data-driven method is an ensemble learning method that employs bootstrap aggregating (bagging) of decision trees and model to do the regression analysis. We chose RF for this study because it has been shown to be robust in dealing with nonlinear hydrological relationships and less prone to overfitting, a common problem with single decision tree models (ZHANG *et alii*, 2024; HUSSAIN *et alii*, 2025).

The model follows the procedure proposed by BORDONI *et alii* (2021), which consists of the following steps:

1. Data Partitioning: We trained separate RF models for each target variable (depths of 10, 50, and 90 cm) and each type of soil management (Control and Rolling), resulting in six different sub-models.
2. Hyperparameter Tuning: The number of trees (n_trees) was selected to ensure the convergence of the Out-of-Bag (OOB) error. The number of variables randomly chosen as candidates at each split ($mtry$) was set to the standard regression default ($p/3$, where p is the number of predictors).
3. Model implementation: The entire dataset was split into a training set for model development and a test set for validation. The training set consists of 75% of the data from the 2021–2023 period, while the remaining 25% serves as the test set. This approach provides a realistic assessment of the model's ability to predict future soil moisture trends. Furthermore, an

Training Performance (2022–2023)

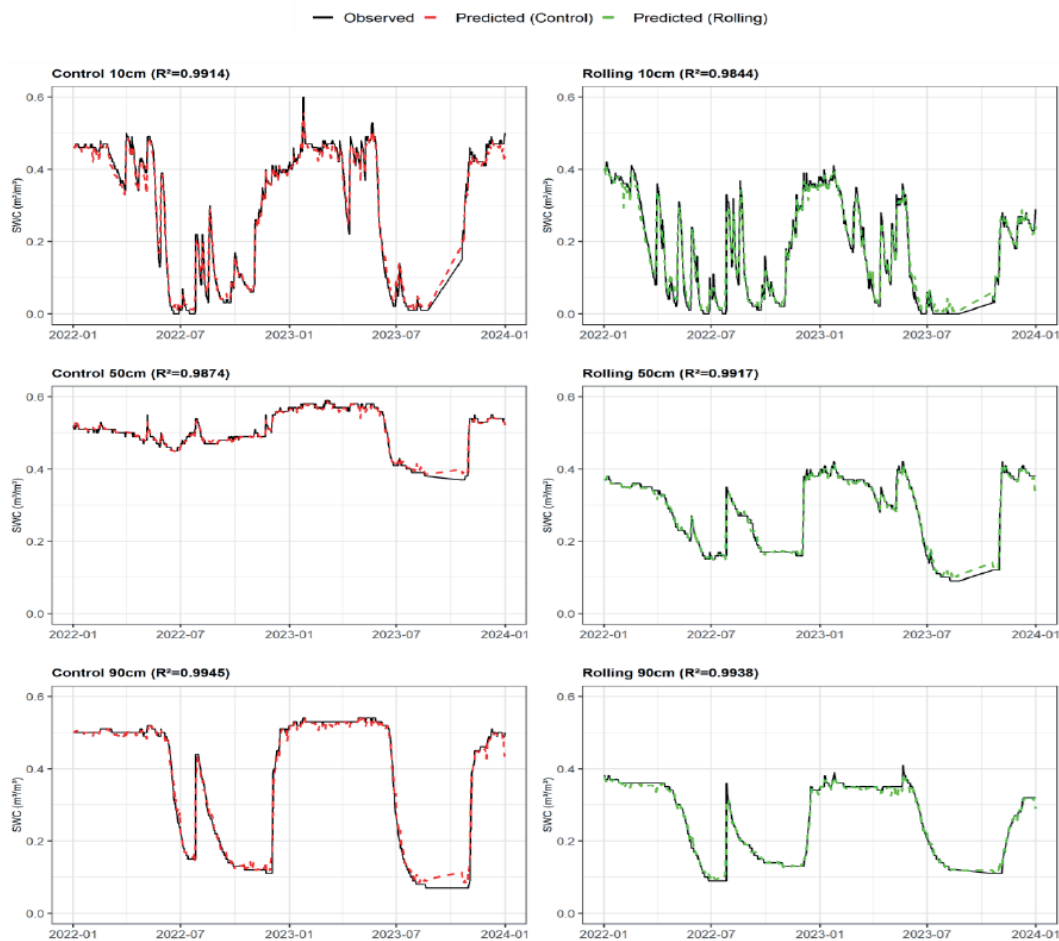


Fig. 2 - Training_2022–2023 period at different depths for both test site

additional validation was performed by predicting 2024 trends, applying the previously developed model to a novel dataset.

4. Performance of the model: The reliability of the model was evaluated by comparing daily observed and modeled soil water content. The Coefficient of Determination (R^2) for both the training and test phases was used as the primary statistical indicator of the model performance.
5. Variable Importance: The algorithm quantified the importance of each predictor by measuring the increase in Mean Squared Error (MSE) when a variable's values were changed. This analysis provides insight into an otherwise "black-box" model by identifying which time lags have the most significant impact at different soil depths.

Furthermore, we evaluated the model's capability to predict the specific conditions that trigger shallow landslides triggering. Such landslides usually occur when the soil is near saturation, causing a significant drop in matric suction (HALTER *et alii*, 2025). Consequently, relying solely on average performance metrics is insufficient to assess the capability the model's accuracy under near-saturation conditions. For these reasons, we specifically analyzed the model's performance during the wettest 10% of days (defined as periods when the SWC exceeded the 90th percentile of the soil water content distribution).

RESULTS AND DISCUSSIONS

Model reliability

A randomized validation was performed to verify the mathematical soundness of the Random Forest algorithm prior to evaluating its predictive performance on unseen data. During this phase, the dataset was randomly split into 75% for training and 25% for testing.

The results (Figure 2) show that the water content predictions for the Rolling management exhibit very high internal consistency, with R^2 values exceeding 0.98 across all analyzed depths. This scores indicate, that the model effectively captures the relationship between rainfall patterns and soil moisture dynamics when the training and testing sets share the same climatic distribution.

Similarly, the Control management showed a comparable level of robustness. The model achieved an $R^2 > 0.98$ at all studied depths (10, 50, and 90 cm), confirming the algorithm's ability to accurately reproduce soil water content trends during the internal validation phase.

Prediction of the external time series

The second phase of the analysis regarded soil water content for an external time series (January–April 2024) which was independent of the original training and test sets (Table 2; Figure 3).

In the Control management, the model struggled to predict soil moisture variations, especially at the 50 cm depth, where the R^2 dropped from 0.56 at 10 cm to 0.06 at 50 cm. This poor

Depth	Control Soil (R^2)	Rolling Soil (R^2)
10 cm	0.56	0.09
50 cm	0.06	0.50
90 cm	0.48	0.12

Tab. 2 - Predictions of the external time series

performance is partly due to the exceptional rainfall recorded in 2024 (509.6 mm between January and April, compared to an average of 112.2 mm for the same period in previous years). These anomalous weather pattern induced near-saturation conditions across both soil managements and all depths. Since such conditions were absent from the training dataset, the model was unable to accurately predict the resulting soil moisture dynamics.

In the Rolling management, R^2 values were similarly low for the 2024 validation set, particularly at the 10 and 90 cm depth. Overall, the external validation results emphasize the critical importance of training data-driven models with time series long enough to encompass the full range of potential meteorological conditions at the study site. By including a broader spectrum of climate variability, it is possible to accurately estimate soil hydrological responses during both extreme drying and near-saturation periods.

Importance of Variables

We analyzed the Variable Importance ranking (Figure 4) to assess the impact of the different predictors on model performance.

For both the Control and Rolling sites, the variable "Month" had the highest impact on model across all depths (10, 50, and 90 cm). This importance is related to a pronounced seasonal trends in soil water content, influenced by varying meteorological conditions throughout the year. The model reliably captured these patterns by utilizing the temporal information associated with each measurement date.

Regarding the rainfall predictors, a clear pattern emerged for both soils managements. In the Rolling site, even at shallower depths (10 and 50 cm), medium- to long-term rainfall predictors such as "Rain in 30 days" and "Rain in 90 days" exhibited higher importance than the immediate rainfall "Rain in 1 day" index. This indicates that total soil water content is more accurately predicted based on cumulative rainfall over extended periods rather than on the preceding day's precipitation. At a depth of 90 cm, the long-term 'memory effect' of "Rain in 60 days" remained the strongest predictor, highlighting the time lag required for water to infiltrate into deeper layers. This behavior is also related to soil textural features; the clayey soils, characteristic of the study area, require more time for deep infiltration due to their low hydraulic conductivity and limited permeability.

In contrast, for the Control site, the long-term indices ("Rain in 90 day" and "Rain in 60 day") showed higher importance than the short-term "Rain in 1 day" predictor, even at the shallowest

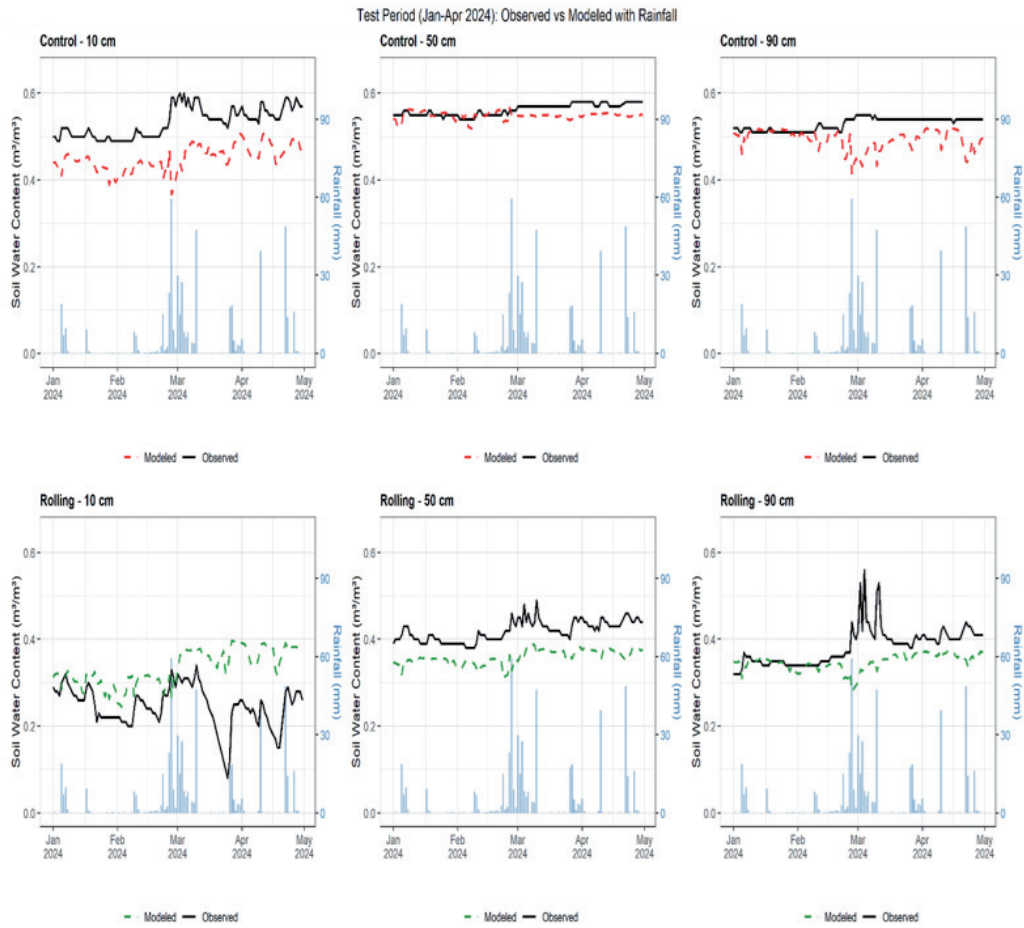


Fig. 3 - Time series plots comparing Control vs. Rolling predictions for 2024



Fig. 4 - Variable Importance Plot showing “Month” as the top predictor

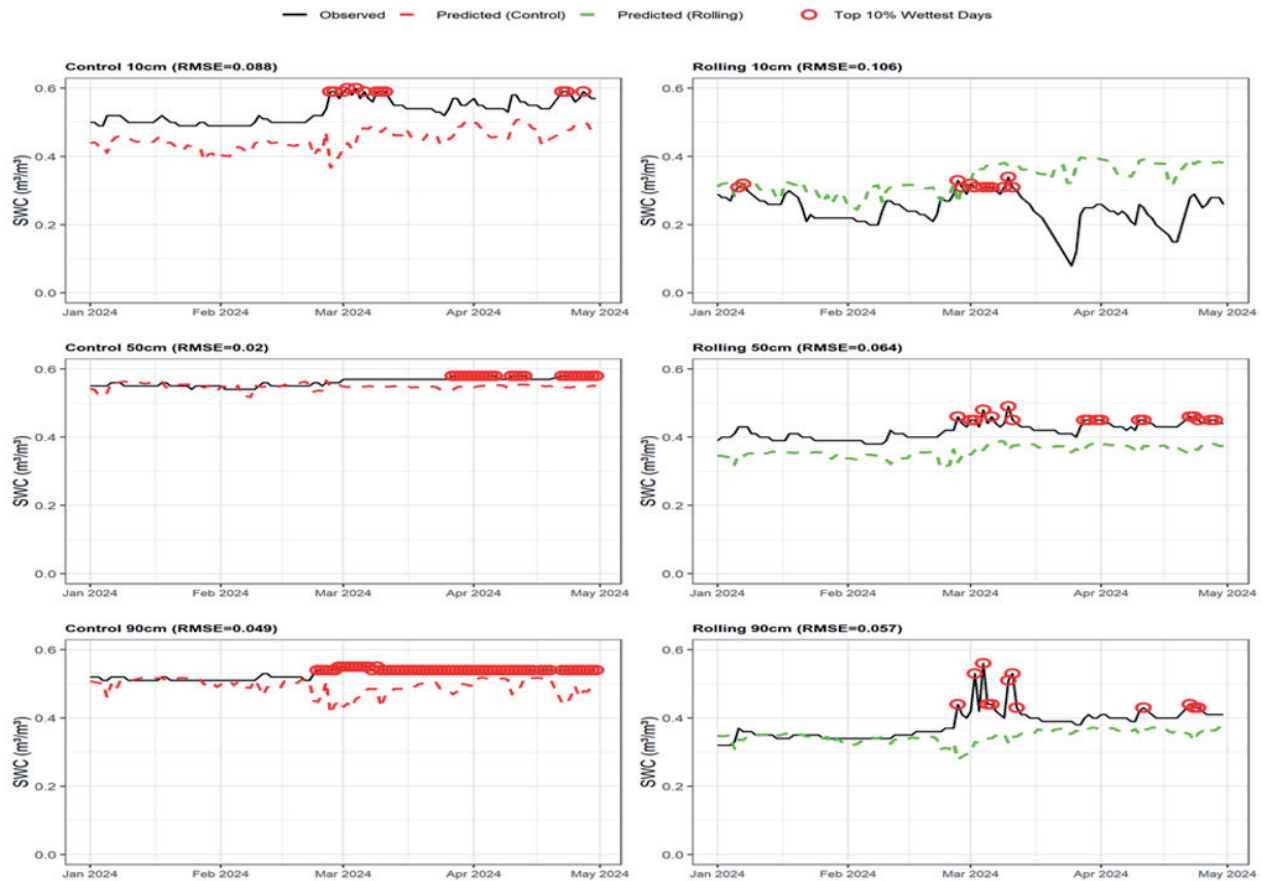


Fig. 5 - The model's performance during the wettest 10% of days in 2024

depth (10 cm). This suggests that in natural soils, the root zone and vegetation cover enhance moisture retention, making the soil water content more sensitive to cumulative seasonal rainfall (3 months) than individual daily precipitation events.

Ability of the model in conditions close to saturation

Predicting instability triggers is essential for the potential application of this model within landslide early warning systems (LEWS).

The model's performance during the wettest 10% of days in 2024 (highlighted with red circles in Figure 5) varied significantly with depth:

1. Quantitative Performance (RMSE): The model's reliability improved with depth. Error margins were higher in the top layers (10 cm), with an RMSE of 0.088 for Control and 0.106 for Rolling. Conversely, deeper layers exhibited greater stability; the Control site at 50 cm achieved the highest accuracy (RMSE = 0.020), followed closely by the 90 cm layers (Control: 0.049, Rolling: 0.057).
2. Pattern Recognition: The results indicate excellent temporal

synchronization: the predicted signal (green dashed line) consistently rises in correspondence with observed peaks (black line). However, the model struggles with peak magnitude. While observed soil water content reached values as high as $0.55 \text{ m}^3/\text{m}^3$ the model predictions typically leveled off (plateaued) around $0.40 \text{ m}^3/\text{m}^3$.

CONCLUSIONS

This study tested the feasibility of a data-driven model for predicting soil water content on slopes susceptible to shallow landslides. The study employed a Random Forest algorithm trained exclusively on cumulative rainfall indices and temporal features.

The results demonstrated that model performance varied significantly depending on soil depth and management technique. The model proved to be mathematically robust ($R^2 > 0.98$) when applied to the 2021-2023 calibration period. However, performance decreased when the model was applied to the novel 2024 dataset, revealing clear predictive limitations. This decline was attributed to the exceptional meteorological conditions of early 2024, which differed significantly from those in the training years. Consequently,

these external validation results highlight the critical importance of utilizing training time-series that are sufficiently long to encompass the full range of potential meteorological variability at the study site.

Regarding the importance of the predictors, the parameters that most significantly affect SWC trends vary according to soil management and depth. At the site with natural grass cover (Control), the month of the year and long-term cumulative rainfall were the primary drivers of soil water contents trends. Similarly, in innovative Rolling management, the model provided accurate predictions, especially at deeper levels, with high importance again related to seasonality and long-term cumulative rainfall. This is consistent with the clayey texture of the soil in the study area.

The research demonstrates that the model identifies critical near-saturation conditions (the top 10% of wettest days) with a low margin of error with RMSE ranging from 0.020 to 0.057 m³/m³. These

findings suggest that Random Forest model is reliable for predicting conditions that could lead to shallow landslides, even without incorporating air temperature or other complex meteorological variables. This reliability is particularly evident when identifying deep-seated saturation within the Rolling management plots.

Future developments of the modeling procedure could incorporate air temperature measurements as proxies for evapotranspiration, which may further enhance the model's predictive accuracy. Additionally, more advanced architectures, such as Long Short Term Memory (LSTM) neural networks, could more effectively capture the soil's "hydrological memory" and complex temporal relationships compared to conventional ensemble-based regression models. Finally, applying the model to diverse soil types and geological settings will be essential to evaluate its reliability and generalizability across different environments.

REFERENCES

- BAKHSHIAN S., ZAREPAKZAD N., NEVERMANN H., HOHENEGGER C., OR D. & SHOKRI N. (2025) - *Field-scale soil moisture dynamics predicted by deep learning*. *Advances in Water Resources*, **201**: 104976. <https://doi.org/10.1016/j.advwatres.2025.104976>
- BORDONI M., INZAGHI F., VIVALDI V., VALENTINO R., BITTELLI M. & MEISINA C. (2021) - *A data-driven method for the temporal estimation of soil water potential and its application for shallow landslides prediction*. *Water*, **13**(9): 1208. <https://doi.org/10.3390/w13091208>
- CUNIAL L., DITI I., BONINI P., PATELLI R., GATTI M., COLA G., BORDONI M., NGUYEN T., MEISINA C., CONFALONIERI R., PALEARI L. & PONI S. (2025) - *Novel termination techniques 919 of winter cover crops in the vineyard: Effects on physiology and performance of Pinot Noir and Malvasia di Candia aromatica grapevines*. *European Journal of Agronomy*, **164**: 127514. <https://doi.org/10.1016/j.eja.2025.127514>
- HALTER T., LEHMANN P., WICKI A., AARON J. & STÄHLI M. (2025) - *Optimising landslide initiation modelling with high-resolution saturation prediction based on soil moisture monitoring data*. *Landslides*, **22**: 2897-2914. <https://doi.org/10.1007/s10346-024-02304-x>
- HUSSAIN S., ARSHAD M., CHEEMA M.J. M., QAMAR M. U., WAJID S. A. & DACCACHE A. (2025) - *Advancing soil moisture prediction using satellite and uav-based imagery with machine learning models*. *Earth Systems and Environment*. (Online first). <https://doi.org/10.1007/s41748-025-00761-5>
- KHOSE S. B. & MAILAPALLI D. R. (2024) - *Spatial mapping of soil moisture content using very-high resolution UAV-based multispectral image analytics*. *Smart Agricultural Technology*, **8**: 100467. <https://doi.org/10.1016/j.atech.2024.100467>
- NGUYEN T. T., NGO H. H., GUO W., CHANG S. W., NGUYEN D. D., NGUYEN C. T., ZHANG J., LIANG S., BUI X.T. & HOANG N. B. (2022) - *A low-cost approach for soil moisture prediction using multi-sensor data and machine learning algorithm*. *Science of The Total Environment*, **833**: 155066. <https://doi.org/10.1016/j.scitotenv.2022.155066>
- WILSON D. J., WESTERN A. W., GRAYSON R. B. (2005) - *A terrain and data-based method for generating the spatial distribution of soil moisture*. *Advances in Water Resources*, **28**(1): 43-54. <https://doi.org/10.1016/j.advwatres.2004.09.007>
- ZHANG M., GE Y., HEUVELINK G. B. M., MA Y. (2024) - *High resolution soil moisture mapping in 3D space and time using machine learning and depth functions*. *Geoderma*, **452**: 117117. <https://doi.org/10.1016/j.geoderma.2024.117117>

Received January 2026 - Accepted April 2026