

# Optimized and Explainable Feature Selection for Soil Moisture Prediction Across Sites

Bamory Ahmed Toru Koné<sup>1,2</sup><sup>a</sup>, Rima Grati<sup>1,3</sup><sup>b</sup>, Bassem Bouaziz<sup>1,4</sup><sup>c</sup>, Khoulood Boukadi<sup>1,2</sup><sup>d</sup>  
and Massimo Mecella<sup>5</sup><sup>e</sup>

<sup>1</sup>University of Sfax, MIRACL Laboratory, Tunisia

<sup>2</sup>Faculty of Economics and Management of Sfax, Tunisia

<sup>3</sup>Computer Science Department, Zayed University, College of Technological Innovation, Abu Dhabi, U.A.E.

<sup>4</sup>Higher Institute of Computer Science and Multimedia of Sfax, Tunisia

<sup>5</sup>Dipartimento di Informatica e Sistemistica, Sapienza University, Rome, Italy

**Keywords:** Precision Agriculture, Soil Moisture Prediction, Feature Selection, Explainable AI, Model Generalizability.


**Abstract:** Accurate soil moisture prediction is critical for improving agricultural practices and managing water supplies. While feature selection techniques have proven useful in enhancing machine learning models' performance in predicting soil moisture, their adaptation to different soil conditions remains limited. To address this gap, this study presents a novel multisite feature selection framework that draws on meteorological and soil data from three distinct locations with mineral, calcareous, and organic soils. The framework identifies soil-specific features through targeted selection processes and then uses SHAP, an explainable AI technique, to assess their global importance and influence. Furthermore, cross-site validation is performed to assess the transferability and generalizability of selected features, giving insight into their resilience across different environments. The proposed approach, which combines explainable AI and cross-site validation, provides a complete approach to understanding and improving feature relevance for soil moisture prediction. Overall, this study establishes the foundation for building more generalizable and robust predictive models, which will improve their applicability in a variety of agricultural and environmental scenarios.


## 1 INTRODUCTION


Global water scarcity is a developing issue caused by various factors, including global warming and excessive water consumption. As global temperatures rise, the earth's climate changes, resulting in droughts, less rainfall, and higher evaporation rates. These changes could affect crop development as well as access to clean water, resulting in food insecurity and scarcity (Togneri et al., 2023). In addition to global warming, many regions use water inefficiently, resulting in water source depletion and increased competition for scarce water resources. Overirrigation, inefficient water distribution infrastructure, and water-intensive


agricultural techniques are all examples of wasteful water use. Numerous scenarios have investigated the incorporation of IoT technologies into agricultural activities (Fattouch et al., 2020), facilitating more intelligent and adaptable decision systems aimed at addressing such inefficiencies. To solve these issues, it is critical to enhance water use efficiency in irrigation systems and practices. Irrigated agriculture accounts for almost 40% of global food production despite accounting for only 17 %of cultivated land (Feres and García-Vila, 2018). Efficient water use in this setting may alleviate water shortages and increase food production.


Irrigated agriculture requires precise crop water requirements based on irrigation scheduling (IS) for optimal water usage (Jones, 2004), (Ben Abdallah et al., 2022), (Ben Abdallah et al., 2023a). Furthermore, predicting upcoming soil moisture is critical to accomplishing effective irrigation management since

<sup>a</sup> <https://orcid.org/0000-0002-5302-0406>

<sup>b</sup> <https://orcid.org/0000-0002-6995-465X>

<sup>c</sup> <https://orcid.org/0000-0002-3692-9482>

<sup>d</sup> <https://orcid.org/0000-0002-6744-711X>

<sup>e</sup> <https://orcid.org/0000-0002-9730-8882>

it allows for an accurate assessment of water stress or availability in the soil. Forecasting such a critical parameter is, therefore, crucial for proactive, long-term decisions in efficient IS (Prasad et al., 2018) and water management (Sanuade et al., 2018). This is because future soil moisture knowledge may capture previously difficult-to-record unexpected land-water fluctuations, avoiding ineffective irrigation decisions. As a result, researchers have extensively utilized machine learning (ML) to forecast such a vital parameter (Ben Abdallah et al., 2023b), (Koné et al., 2023).

Furthermore, feature selection is a crucial step in developing machine learning models because it selects the most relevant features for the prediction task, hence reducing model complexity and enhancing predictive performance. This process is especially significant for soil moisture prediction because numerous studies have investigated the impact of various environmental and climatological factors on model accuracy. Several studies, including (Togneri et al., 2023), (Yu et al., 2021), and (Adeyemi et al., 2018), have used varying climatological and soil-related features to forecast future soil moisture content. Despite the widely recognized advantages of feature selection, attempts to identify the most important features in soil moisture prediction have not been thoroughly evaluated in contrasting circumstances. More specifically, existing soil moisture prediction studies (Togneri et al., 2023), (Adeyemi et al., 2018), (Yu et al., 2021), (Dubois et al., 2021), (Cai et al., 2019), (He et al., 2022), while achieving high performance with fewer features, often lack comprehensive investigation of the influence of their selected input features under contrasting soil conditions. This lack of generalizability emphasizes the need for a more robust methodology capable of accounting for a wide range of soil and climate variables.

To address this limitation, we propose a novel feature selection approach to effectively identify relevant features for soil moisture prediction across distinct sites with contrasting soil conditions. Specifically, we examine three different soil types: mineral, calcareous, and organic. Overall, the main focuses of this paper are as follows:

1. **Development of a Multisite Feature Selection Framework:** This framework assesses feature relevance across distinct soil types, enabling a robust comparison of feature importance for mineral, calcareous, and organic soils.
2. **Cross-Site Validation of Feature Relevance:** We evaluate models trained on selected features from one site on other sites, providing insights into the cross-site transferability of feature sets. This validation highlights the robustness of the selected

features when applied to different soil conditions, helping to identify universally impactful predictors.

3. **Feature Analysis Using SHAP Explanations:** We use SHAP, a widely used explainable AI technique, to assess the input features. SHAP gives global rankings and insights into the contributions of specific features to soil moisture predictions, allowing for a better understanding of their impact across different soil types. By comparing SHAP explanations from different sites, we can see how the relevance and behavior of features change with soil and environmental variables.

It is crucial to note that this paper does not attempt to comprehensively address all challenges associated with soil moisture prediction. Instead, the emphasis in this study is on advancing the field by proposing a focused feature selection approach customized to soil-specific as well as cross-site scenarios. As a result, substantial training and evaluation of machine learning models, whether shallow or deep, are outside the scope of this paper as these aspects have been thoroughly documented in earlier studies. Therefore, the fundamental purpose of this research is to improve soil moisture prediction by developing a more effective and generalizable feature selection framework that is intended to supplement and reinforce existing predictive methods.

## 2 MATERIALS AND METHODS

The proposed approach in this study is illustrated in Figure 1. We begin by selecting meteorological and soil data from a multisite dataset collected at three different locations, each characterized by unique soil types: mineral, calcareous, or organic. For each soil type, the dataset is analyzed independently to perform feature selection, which results in identifying the most relevant soil-specific properties. To further understand the value and influence of these features, we use SHAP, a widely used explainable AI technique. SHAP allows us to evaluate selected features' global ranking and contribution to soil moisture prediction. Furthermore, the SHAP explanations developed for each site are then compared to study the variability in feature impact across the three soil types, revealing how soil-specific factors influence the prediction process. This framework not only highlights the relevance of soil-specific feature selection but also emphasizes the need to understand feature contributions under varying environmental conditions, laying the groundwork for developing robust and generalizable soil moisture prediction models.

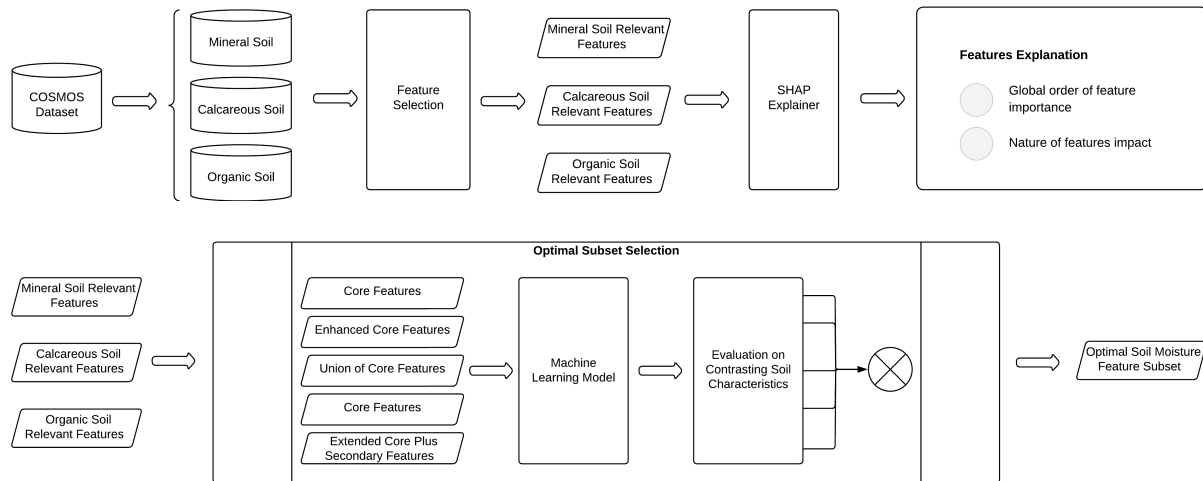


Figure 1: Overall Approach.

Table 1: Descriptive summary of Balruderry training site.

Feature full name	Abbreviation	Unit
Net radiation	RN	Watts per square meter ( $Wm^{-2}$ )
Precipitation	PRECIP	Millimetres (mm)
Air pressure	PA	HectoPascals (hPa)
Air temperature	Abbreviation	Degrees Celcius ( $^{\circ}C$ )
Wind speed	WS	Metres per second ( $ms^{-1}$ )
Wind direction	WD	Degrees ( $^{\circ}$ )
Absolute humidity	Q	Grams per cubic meter ( $gm^{-3}$ )
Relative humidity	RH	Percent (%)
Heat Flux	G1, G2	Watts per square meter ( $Wm^{-2}$ )
Soil temperature from $TDT_x$ sensor	TDTx.TSOIL	Degrees Celcius ( $^{\circ}C$ )
Soil moisture from $TDT_x$ sensor	TDTx_VWC	Percent (%)
Incoming longwave radiation	LWIN	Megajoule per square meter ( $MJm^{-2}$ )
Outgoing longwave radiation	LWOUT	Megajoule per square meter ( $MJm^{-2}$ )
Incoming shortwave radiation	SWIN	Megajoule per square meter ( $MJm^{-2}$ )
Outgoing shortwave radiation	SWOUT	Megajoule per square meter ( $MJm^{-2}$ )
Effective depth of CRNS (D86 at 75m)	D86.75m	Centimetres (cm)
Albedo	ALBEDO	

Furthermore, we combine the chosen site-specific features into a categorization scheme that has four subsets: core, enhanced core, union, and enhanced core + secondary features. The particulars of each subset are defined as follows:

1. **Core Consistent Features:** This subset contains the features that are shared among the core features across all three sites.
2. **Enhanced Core Features:** This subset includes the core consistent features and adds any secondary features that appear consistently across all three sites.
3. **Union of Core Features:** This subset consists of the union of core features across the three sites, capturing all primary predictors for each site.

4. **Extended Core Plus Secondary Features:** This subset builds on the union of core features by adding any secondary features that are important across all sites. In this case, only RH meets this criterion.

By examining these subsets, we seek to lay a solid foundation for a soil moisture prediction model that can effectively generalize over a wide range of environmental conditions. Detailed descriptions of each component of our approach are provided in the following sections of this research.

## 2.1 Study Site and Data Preprocessing

This study used the Cosmic-ray soil moisture monitoring (COSMOS) dataset (Stanley et al., 2023) to

train and evaluate models because of its availability and broad range coverage. The dataset is made up of data collected from 51 different sites across the United Kingdom, each with its own unique set of characteristics. This study used the most recent version of COSMOS and included daily hydrometeorological and soil measurements for approximately ten years, from October 2013 to December 2023. Meteorological data include radiation (shortwave, longwave, and net), precipitation, atmospheric pressure, air temperature, wind speed and direction, and humidity. Soil observations include measurements of soil heat flux, temperature, and moisture reported as volumetric water content (VWC) at different depths.

From the initial dataset of 51 sites, we chose nine different sites for the purpose of our study, with three sites for each soil characteristic. For effective modeling, we emphasized sites with few missing values while maintaining site diversity. The characteristics of the different sites are summarized in Table 2. Furthermore, missing values were interpolated to prevent the models from performing significantly worse. Interpolation is the process of calculating missing values for an observation based on its preceding values. The sequential nature of this interpolation technique fits the temporal nature of time-series data. A brief description of the main features of the dataset are summarized in Table 1.

Table 2: Characteristics of the training and test sites.

Site Name	Abbreviation	Soil type	Land cover
Balruddery	BALRD	Mineral soil	Arable and Horticulture
Alice Holt	ALIC1	Mineral soil	Broadleaf woodland
Bickley Hall	BICKL	Mineral soil	Grassland
Chimney Meadows	CHIMN	Calcareous soil	Improved grassland
Lullington Heath	LULLN	Calcareous soil	Calcareous grassland
Porton Down	PORTN	Calcareous soil	Improved grassland
Glensaugh	GLENS	Organic soil	Heather
Henfaes Farm	HENFS	Organic soil	Acid grassland
Plynlimon	PLYNL	Organic soil	Acid grassland

The dataset was then divided into training, validation, and test data. Because we are working with time-series data, the split was performed sequentially to preserve the data's temporal dynamics. As a result, we used data from 2014 to 2019 for training and 2020 and 2021 for validation. Additionally, data gathered in 2022 at each location was not included in the training phase and therefore constitute evaluation or test sets. Finally, we standardized the data as an additional preprocessing step to deal with the wide range of input parameters. This ensures that the mean value of each input parameter is 0 and the standard deviation is 1, allowing neural network-based models to learn more effectively from data.

## 2.2 Feature Selection Techniques

Feature selection is a critical preprocessing step in machine learning that identifies the most relevant features (variables) in a dataset for model training. The optimization technique analyzes alternative subsets of features to identify one that is optimal or near-optimal based on specific performance criteria, with the goal of minimizing a given measure (Karnan et al., 2011). This study focuses on two widely used feature selection techniques: recursive feature elimination (RFE) and Boruta.

Recursive Feature Elimination (Guyon et al., 2002) is a feature selection method that improves model performance, decreases overfitting and increases interpretability. The method iteratively removes the least significant features based on a ranking criterion until the target amount of features is obtained. The initial stage in RFE is to train a machine learning model on the original dataset and rank the features according to relevance. Feature weights in linear models, feature importance in tree-based models, and support vector machine coefficients all serve as ranking criteria. Several research studies have shown that RFE is useful in a range of domains, including soil quality assessment and decision support systems in agriculture (Ijadi Maghsoodi et al., 2023).

Boruta (Kursa et al., 2010) proposes a powerful feature selection approach for selecting key characteristics from huge datasets. It is specifically developed for all-relevant feature selection, as opposed to RFE, which concentrates on selecting a minimal sample of features. The approach compares the significance of the original features to random shadow features. These shadow features are created by individually rearranging the values of each feature. Boruta can determine the true relevance of each feature and successfully remove insignificant or noisy ones by comparing the importance scores of the original features to the shadow features.

## 3 RESULTS AND DISCUSSIONS

### 3.1 Site-Specific Feature Relevance Analysis

To determine the most important features for soil moisture prediction across different soil types, we used an iterative feature selection procedure with recursive feature elimination. This approach allowed us to identify variables that consistently improve forecast accuracy across soil types, as well as those that are context-dependent. By adapting feature selection



to each soil type, we seek to develop models that are more resilient in terms of assessment metrics for predicting soil moisture in a variety of environments. Furthermore, for each soil type, we examined three different locations, using RFE to determine the most essential subsets of features for each site. Table 3 shows the top features identified for each site and soil type, along with the  $R^2$  scores of fine tuned random forest models trained on these subsets. Each site's data is divided into three rows: the first represents the initial set of all features, the second shows selected features via RFE, and the third displays those chosen by Boruta. This structured method allows us to compare feature significance across feature selection techniques and environments, which helps us better understand cross-soil predictors and site-specific features.

Our findings show that RFE and Boruta effectively select relevant features, minimizing inputs and improving model performance across all sites and soil types. Regarding the selection techniques, Boruta outperformed RFE in most sites, demonstrating its robustness. As for the soil types, mineral soils have consistent features such as `LWIN`, `SWOUT`, `PRECIP`, and `TD1_VWC` (previous soil moisture values). These features were consistently ranked among the best selections in at least one selection technique across all sites, with `PRECIP` and `TD1_VWC` showing reliability across both RFE and Boruta. Other factors, such as `RN` and `RH`, were generally important but not as consistently impactful as the main features. Based on these results, we classified the features as follows:

1. Core Features: Features consistently appearing in the top selections of at least one method across all sites of the soil type.
2. Secondary Features: Features relevant in most sites of the soil type but less critical than the primary group.
3. Supplementary Features: Features of lesser relevance, appearing in fewer sites.

Table 4 summarizes all relevant features by soil type according to this classification. These insights lay the groundwork for the following subsection.

### 3.2 Interpreting Feature Importance and Impact with SHAP

Figure 2a illustrates the global explanation of the model by highlighting the importance of each feature as well as its effect on the model's outputs in mineral soil. Similar illustrations for calcareous and organic soils are depicted in Figures 2b and 2c.

Given these global explanations, we might infer that the model efficiently captures soil moisture dynamics by utilizing essential environmental interactions rather than relying on less interpretable or coincidental data patterns to improve accuracy. This is demonstrated by the model's identification that factors such as low precipitation and high levels of solar radiation are significant contributors to lower soil moisture. Such trends are consistent with accepted environmental principles: higher radiation typically reduces soil moisture, whereas increasing precipitation usually increases it. Thus, the model's capacity to reflect these well-known contextual interactions contributes to increased confidence in its predictions, strengthening its trustworthiness even in the context of typically opaque "black-box" models.

Based on the model's explanations for each type of soil, it is obvious that soil type has a major impact on how the model understands the correlations between features and soil moisture forecasts. In mineral and organic soils, `TD1_VWC` consistently improves forecasts. At the same time, its dominance is somewhat reduced in calcareous soils due to the additional influence of thermal and radiative features like `STP_TSOIL50` and `SWIN`. Furthermore, `PRECIP` has a strong positive effect in mineral soils but becomes less impactful in calcareous and organic soils, probably due to varying water retention and drainage capacities. Further investigation demonstrates that some features have conflicting effects depending on the soil type. For example, `SWOUT`, which significantly effects forecasts in mineral soils, has a less pronounced or even opposite effect in organic soils.

As a result, the plots show how environmental variables including radiation, precipitation, and soil temperature interact differently depending on the soil type. Mineral soils have more linear and clear interactions (for example, increased precipitation increases soil moisture). However, in organic soils with higher water retention, the impact of these drivers can be mitigated or exhibit nonlinear trends. Overall, changing soil type modifies the magnitude and direction of feature impacts. Universally significant features (e.g., `TD1_VWC`) remain important but may interact differently with secondary factors based on soil physical and hydrological properties. Some features, such as `SWOUT`, switch their impact polarity, reflecting how different soils respond to environmental variables. Given these findings, the limitations of prior research that rely solely on feature selection for a particular soil type become obvious, despite the great predictive performances achieved in those studies. A more generalized approach is to incorporate soil type data into the feature selection process.

Table 3: Model Performance by Soil Type, Site, and Feature Set.

Soil Type	Training Site	Features	R <sup>2</sup> Score
Mineral	BALRD	All 25 features	0.6713
		SWIN, RN, PRECIP, RH, TDT1_VWC, TDT2_VWC	0.7001
		LWIN, SWIN, SWOUT, RN, PRECIP, RH, TDT1_VWC, TDT2_VWC	0.7133
	ALIC1	All 25 features	0.9793
		LWIN, SWOUT, PRECIP, RH, TDT1_VWC, STP_TSOIL5	0.9898
		LWIN, SWOUT, PRECIP, Q, TDT1_VWC	0.9902
	BICKL	All 25 features	0.9004
		RN, PRECIP, TDT1_VWC, TDT2_VWC, D86_75M, ALBEDO	0.9062
		LWIN, SWOUT, PRECIP, TDT1_VWC, ALBEDO	0.9405
Calcareous	CHIMN	All 25 features	0.8407
		SWIN, PRECIP, PA, RH, G1, TDT1_VWC	0.8619
		SWIN, PRECIP, RH, TDT1_VWC	0.8718
	LULLN	All 25 features	0.9414
		SWIN, PRECIP, G1, TDT1_VWC, TDT2_VWC, STP_TSOIL20	0.9400
		SWIN, PRECIP, TDT1_VWC, TDT2_VWC	0.9440
	PORTN	All 25 features	0.8983
		PRECIP, PA, RH, TDT1_TSOIL, TDT1_VWC, TDT2_VWC	0.9001
		LWIN, PRECIP, TDT1_TSOIL, TDT1_VWC	0.9166
Organic	GLENS	All 25 features	0.8224
		PRECIP, RH, TDT1_VWC, TDT2_VWC, D86_75M, ALBEDO	0.8670
		RH, TDT1_VWC, TDT2_VWC, D86_75M	0.8784
	HENFS	All 25 features	0.7108
		PRECIP, PA, TDT1_VWC, TDT2_VWC, STP_TSOIL50, D86_75M	0.7197
		PRECIP, PA, TDT1_VWC, TDT2_VWC, STP_TSOIL50, D86_75M	0.7356
	PLYNL	All 25 features	0.6512
		LWIN, SWIN, PRECIP, TDT1_VWC, TDT2_VWC, STP_TSOIL50	0.6931
		LWIN, SWIN, SWOUT, PRECIP, RH, TDT1_VWC, TDT2_VWC, STP_TSOIL2, STP_TSOIL50	0.6839

Table 4: Feature Classification by Soil Type.

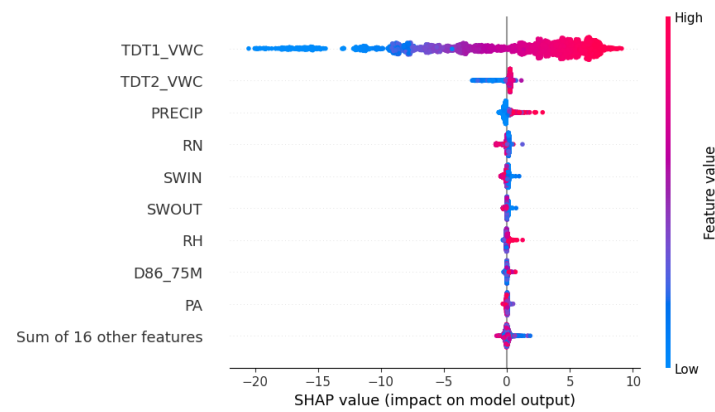
Soil Type	Classification	Features
Mineral	Core Features	LWIN, SWOUT, PRECIP, TDT1_VWC
	Secondary Features	RN, RH
	Supplementary Features	SWIN, TDT2_VWC, STP_TSOIL5, Q, D86_75M, ALBEDO
Calcareous	Core Features	PRECIP, TDT1_VWC
	Secondary Features	SWIN, PA, RH, G1, TDT2_VWC
	Supplementary Features	STP_TSOIL20, TDT1_TSOIL, LWIN
Organic	Core Features	PRECIP, TDT1_VWC, TDT2_VWC, STP_TSOIL50
	Secondary Features	RH, D86_75M
	Supplementary Features	PA, LWIN, SWIN, SWOUT, STP_TSOIL2

### 3.3 Evaluating Feature Subsets for Cross-Site Generalization

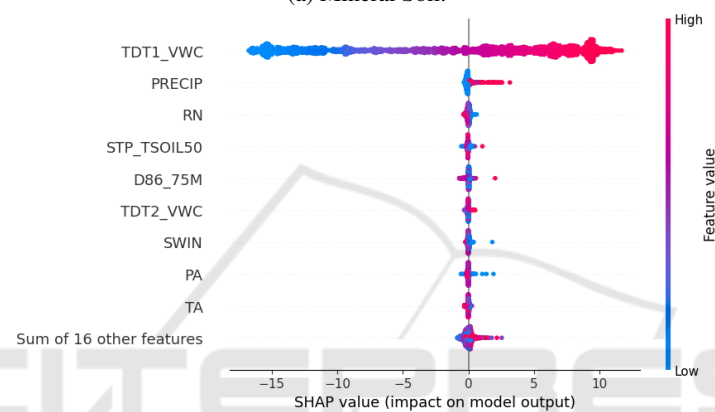
Building on the prior research, our goal here is to identify features that are consistent across sites rather than those that are unique to each one. These cross-site features are not site-specific, yet they are necessary for reliable soil moisture prediction across a wide range of soil types. To accomplish this, we evaluated four feature subsets: core, enhanced core, union, and enhanced core plus secondary features (specified in Section 2). These feature subsets are evaluated for their overall usefulness in predicting soil moisture across different locations and soil types (Table 5). The

table compares cross-site model performance using feature classification at three training sites (BICKL, CHIMN, and GLENS). Key conclusions can be drawn regarding the effectiveness of each feature subset in producing accurate and generalizable soil moisture predictions across different sites:

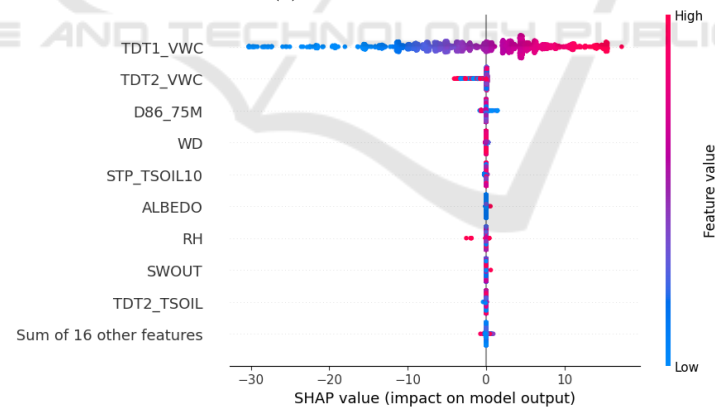
- Enhanced Core Features Perform Well Across Sites:** The enhanced core features subset, consisting of PRECIP, TDT1\_VWC, and RH, consistently yields strong  $R^2$  scores across all evaluation sites. For example, when BICKL is the training site, models evaluated at CHIMN and GLENS achieve  $R^2$  scores of 0.8391 and 0.8708, respectively. Similarly, training on CHIMN and evaluating on



(a) Mineral Soil.



(b) Calcareous Soil.



(c) Organic Soil.

Figure 2: SHAP beeswarm plot showing the impact of various features on the model's prediction for each type of soil.

BICKL and GLENS yields  $R^2$  scores of 0.9404 and 0.8720. This subset often performs comparably to, or better than, larger feature sets (such as the Union of Core Features and Extended Core Plus Secondary Features), suggesting that these three features offer both efficiency and predictive strength.

- **Core Consistent Features Provide a Simpler,**

**Yet Effective, Model:** The core consistent features subset (PRECIP and TDT1\_VWC) achieves relatively high  $R^2$  scores, but it is slightly outperformed by the enhanced core features subset in most cases. For example, training on CHIMN with this subset yields  $R^2$  scores of 0.9410 on BICKL and 0.8655 on GLENS, while adding RH (in the enhanced core features subset) improves

Table 5: Cross-Site Model Performance by Feature Classification and Training Site.

Training Site	Classification	Feature Names	Evaluation Site	$R^2$ Score
BICKL	Core Consistent Features	PRECIP, TDT1_VWC	CHIMN GLENS	0.8345 0.8564
	Enhanced Core Features	PRECIP, TDT1_VWC, RH	CHIMN GLENS	0.8391 0.8708
	Union of Core Features	LWIN, SWOUT, PRECIP, STP_TSOIL50, TDT1_VWC, TDT2_VWC	CHIMN GLENS	0.8472 0.8418
	Extended Core Plus Secondary Features	LWIN, SWOUT, PRECIP, STP_TSOIL50, TDT1_VWC, TDT2_VWC, RH	CHIMN GLENS	0.8561 0.8253
	Site-specific Features	LWIN, SWOUT, PRECIP, TDT1_VWC, ALBEDO	CHIMN GLENS	0.8185 0.8718
CHIMN	Core Consistent Features	PRECIP, TDT1_VWC	BICKL GLENS	0.9410 0.8655
	Enhanced Core Features	PRECIP, TDT1_VWC, RH	BICKL GLENS	0.9404 0.8720
	Union of Core Features	LWIN, SWOUT, PRECIP, STP_TSOIL50, TDT1_VWC, TDT2_VWC	BICKL GLENS	0.9325 0.8408
	Extended Core Plus Secondary Features	LWIN, SWOUT, PRECIP, STP_TSOIL50, TDT1_VWC, TDT2_VWC, RH	BICKL GLENS	0.9350 0.8439
	Site-specific Features	LWIN, SWOUT, PRECIP, TDT1_VWC, ALBEDO	BICKL GLENS	0.9353 0.8673
GLENS	Core Consistent Features	PRECIP, TDT1_VWC	BICKL CHIMN	0.9353 0.8236
	Enhanced Core Features	PRECIP, TDT1_VWC, RH	BICKL CHIMN	0.9474 0.8573
	Union of Core Features	LWIN, SWOUT, PRECIP, STP_TSOIL50, TDT1_VWC, TDT2_VWC	BICKL CHIMN	0.9339 0.8398
	Extended Core Plus Secondary Features	LWIN, SWOUT, PRECIP, STP_TSOIL50, TDT1_VWC, TDT2_VWC, RH	BICKL CHIMN	0.9223 0.8404
	Site-specific Features	LWIN, SWOUT, PRECIP, TDT1_VWC, ALBEDO	BICKL CHIMN	0.9045 0.8495

the  $R^2$  score on GLENS to 0.8720. Although core consistent features perform well, the slight increase in performance from adding RH indicates that RH is an important factor for enhanced generalizability.

- **Union of Core Features Shows Inconsistent Results:** The union of core features subset, which includes six features (LWIN, SWOUT, PRECIP, STP\_TSOIL50, TDT1\_VWC, and TDT2\_VWC), performs inconsistently across sites. While it achieves relatively high scores, it sometimes falls below the simpler enhanced core features subset. For example, training on BICKL and evaluating GLENS with this subset results in an  $R^2$  score of 0.8418, which is lower than the 0.8708 achieved with the enhanced core features. This suggests that adding more features does not necessarily improve performance and may even introduce noise or overfitting.
- **Extended Core Plus Secondary Features Do Not Consistently Improve Performance:** The extended core plus secondary features subset, which adds RH to the union of core features, does

not consistently outperform simpler feature subsets. In some cases, it achieves similar or even lower  $R^2$  scores compared to the core consistent or enhanced core features subsets. For instance, training on GLENS and evaluating on BICKL gives an  $R^2$  score of 0.9223 with this subset, which is lower than the 0.9474 obtained with enhanced core features. This indicates that the inclusion of all core and secondary features may not provide additional predictive power and suggests that smaller, well-selected feature sets can be more effective.

- **Site-Specific Features Provide Comparable Performance but Lack Generalizability:** The site-specific features subset achieves competitive  $R^2$  scores, sometimes matching or exceeding those of other subsets. For example, when trained on CHIMN and evaluated on GLENS, the site-specific subset achieves an  $R^2$  of 0.8673, close to the enhanced core subset's 0.8720. However, models trained on site-specific features are less generalizable, as they are tailored to particular site characteristics that may not transfer well across different environments. This indicates that while



site-specific features may be useful for local predictions, they lack the consistency required for cross-site applications.

Overall, the enhanced core features subset (PRECIP, TDT1\_VWC, and RH) consistently delivers strong performance across all sites, providing an ideal mix of simplicity and predictive accuracy. In most situations, it outperforms or matches bigger feature sets, demonstrating its ability to predict soil moisture across many sites. This shows that a small collection of highly predictive variables is desirable for generalizable soil moisture models in a variety of situations. These findings show that including site-specific variables might reduce model generalizability, which is commonly missed in studies that focus on single-site data. By selecting universally relevant features, we improve the model's adaptability to a variety of environmental situations while maintaining high performance. As a result, our research proposes an ideal feature subset for soil moisture prediction that maximizes accuracy while keeping model simplicity, thereby enabling the development of robust, generalizable prediction models.

## 4 CONCLUSIONS

This study proposed a multisite feature selection framework to address the limitations of single-site feature selection methods in soil moisture prediction. While single-site feature selection can effectively reduce data dimensionality and improve model performance for a specific site or sites with similar soil characteristics, its relevance is much reduced when applied to sites with varied soil properties. The current literature has not adequately investigated the performance of soil moisture prediction models across diverse soil types, resulting in a crucial gap in understanding their generalizability. To address this gap, the present article examined three separate soil types and identified the most relevant inputs specific to each, as well as those that are universally significant across all three types. The feature selection procedure involves creating numerous probable subsets of features and meticulously examining their relevance, both within and between sites. Two advanced techniques, RFE and the Boruta algorithm, were used to systematically determine and validate these features.

The results of this research highlight a number of major contributions. First, the identified site-specific features provide a solid foundation for studies concentrating on specific soil types, allowing them to quickly identify the most relevant indicators for enhanced model performance. Second, our study es-

tablishes a robust and generalizable feature selection framework by validating the transferability of important predictive characteristics across diverse sites (TDT1\_VWC, PRECIP, and RN). This approach delivers great predicted accuracy while keeping model complexity low and assuring cross-site generalizability. As a result, our findings demonstrate that models trained on a core subset of globally significant characteristics can effectively generalize across different soil types, indicating the potential for multisite feature selection to improve environmental modeling tasks. This framework not only enhances prediction accuracy, but it also increases model efficiency and adaptability under a variety of environmental situations.

## REFERENCES

- Adeyemi, O., Grove, I., Peets, S., Domun, Y., and Norton, T. (2018). Dynamic Neural Network Modelling of Soil Moisture Content for Predictive Irrigation Scheduling. *Sensors*, 18(10):3408. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.
- Ben Abdallah, E., Grati, R., and Boukadi, K. (2022). A machine learning-based approach for smart agriculture via stacking-based ensemble learning and feature selection methods. In *2022 18th International Conference on Intelligent Environments (IE)*, pages 1–8. IEEE.
- Ben Abdallah, E., Grati, R., and Boukadi, K. (2023a). Towards an explainable irrigation scheduling approach by predicting soil moisture and evapotranspiration via multi-target regression. *Journal of Ambient Intelligence and Smart Environments*, 15(1):89–110.
- Ben Abdallah, E., Grati, R., and Boukadi, K. (2023b). Towards an explainable irrigation scheduling approach by predicting soil moisture and evapotranspiration via multi-target regression. *Journal of Ambient Intelligence and Smart Environments*, 15(1):89–110. Publisher: IOS Press.
- Cai, Y., Zheng, W., Zhang, X., Zhangzhong, L., and Xue, X. (2019). Research on soil moisture prediction model based on deep learning. *PLOS ONE*, 14(4):e0214508. Publisher: Public Library of Science.
- Dubois, A., Teytaud, F., and Verel, S. (2021). Short term soil moisture forecasts for potato crop farming: A machine learning approach. *Computers and Electronics in Agriculture*, 180:105902.
- Fattouch, N., Lahmar, I. B., and Boukadi, K. (2020). Iot-aware business process: comprehensive survey, discussion and challenges. In *2020 IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pages 100–105. IEEE.
- Fereres, E. and García-Vila, M. (2018). Irrigation Management for Efficient Crop Production. In Meyers,

- R. A., editor, *Encyclopedia of Sustainability Science and Technology*, pages 1–17. Springer, New York, NY.
- Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1):389–422.
- He, B., Jia, B., Zhao, Y., Wang, X., Wei, M., and Dietzel, R. (2022). Estimate soil moisture of maize by combining support vector machine and chaotic whale optimization algorithm. *Agricultural Water Management*, 267:107618.
- Ijadi Maghsoodi, A., Torkayesh, A. E., Wood, L. C., Herrera-Viedma, E., and Govindan, K. (2023). A machine learning driven multiple criteria decision analysis using ls-svm feature elimination: Sustainability performance assessment with incomplete data. *Engineering Applications of Artificial Intelligence*, 119:105785.
- Jones, H. G. (2004). Irrigation scheduling: advantages and pitfalls of plant-based methods. *Journal of Experimental Botany*, 55(407):2427–2436.
- Karnan, M., Akila, M., and Krishnaraj, N. (2011). Biometric personal authentication using keystroke dynamics: A review. *Applied Soft Computing*, 11(2):1565–1573.
- Koné, B. A. T., Grati, R., Bouaziz, B., and Boukadi, K. (2023). A new long short-term memory based approach for soil moisture prediction. *Journal of Ambient Intelligence and Smart Environments*, 15(3):255–268.
- Kursa, M. B., Jankowski, A., and Rudnicki, W. R. (2010). Boruta – A System for Feature Selection. *Fundamenta Informaticae*, 101(4):271–285. Publisher: IOS Press.
- Prasad, R., Deo, R. C., Li, Y., and Maraseni, T. (2018). Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition. *Geoderma*, 330:136–161.
- Sanuade, O. A., Adetokunbo, P., Oladunjoye, M. A., and Olajojo, A. A. (2018). Predicting moisture content of soil from thermal properties using artificial neural network. *Arabian Journal of Geosciences*, 11(18):566.
- Stanley, S., Antoniou, V., Askquith-Ellis, A., Ball, L., Bennett, E., Blake, J., Boorman, D., Brooks, M., Clarke, M., Cooper, H., Cowan, N., Cumming, A., Evans, J., Farrand, P., Fry, M., Hitt, O., Lord, W., Morrison, R., Nash, G., Rylett, D., Scarlett, P., Swain, O., Szczukulska, M., Thornton, J., Trill, E., Warwick, A., and Winterbourn, B. (2023). Daily and sub-daily hydrometeorological and soil data (2013-2022) [cosmos-uk].
- Togneri, R., Prati, R., Nagano, H., and Kamienski, C. (2023). Data-driven water need estimation for iot-based smart irrigation: A survey. *Expert Systems with Applications*, 225:120194.
- Yu, J., Zhang, X., Xu, L., Dong, J., and Zhangzhong, L. (2021). A hybrid CNN-GRU model for predicting soil moisture in maize root zone. *Agricultural Water Management*, 245:106649.