Journal Pre-proof

Enhanced Root Zone Soil Moisture Monitoring Using Multitemporal Remote Sensing Data and Machine Learning Techniques

Atefeh Nouraki, Mona Golabi, Mohammad Albaji, Abd Ali Naseri, Saeid Homayouni

PII: S2352-9385(24)00218-0

DOI: https://doi.org/10.1016/j.rsase.2024.101354

Reference: RSASE 101354

To appear in: *Remote Sensing Applications: Society and Environment*

Received Date: 4 May 2024

Revised Date: 5 September 2024

Accepted Date: 10 September 2024

Please cite this article as: Nouraki, A., Golabi, M., Albaji, M., Naseri, A.A., Homayouni, S., Enhanced Root Zone Soil Moisture Monitoring Using Multitemporal Remote Sensing Data and Machine Learning Techniques, *Remote Sensing Applications: Society and Environment*, https://doi.org/10.1016/j.rsase.2024.101354.

Enhanced Root Zone Soil Moisture Monitoring Using Multitemporal Remote Sensing Data and Machine Learning Techniques

Atefeh Nouraki[a], Mona Golabi[a,*], Mohammad Albaji[a], Abd Ali Naseri[a], Saeid Homayouni[b]

[a] Department of Irrigation and Drainage, Faculty of Water and Environmental Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran

[b] Centre Eau Terre Environnement, Institut national de la recherche scientifique (INRS), 490 Couronne St, Quebec, QC G1K 9A9, Canada.

[*]Corresponding author, E-mail address: mona_golabi@yahoo.com.

Abstract

Accurate root zone soil moisture (RZSM) estimation using remote sensing (RS) in areas with dense vegetation is essential for real-time field monitoring and precise irrigation scheduling. Traditional methods often face challenges due to the dense crop cover and the complexity of soil and climate interactions. These challenges include the coarse spatial resolution of available soil moisture products, the influence of vegetation and surface roughness, and the difficulty of estimating RZSM from surface data. Aiming to overcome these limitations, two RZSM estimation methods were developed by combining synthetic aperture radar (SAR) data from Sentinel-1 (VV and VH polarizations) and optical and thermal RS data from Landsat-8. These data sources were used in conjunction with various machine learning (ML) models such as M5-pruned (M5P), support vector regression (SVR), extreme gradient boosting (XGBoost), and random forest regression (RFR) to improve the accuracy of soil moisture estimation. In addition to RS data, soil physical and hydraulic properties, meteorological variables, and topographical parameters were selected as inputs to the ML models for estimating the RZSM of sugarcane crops in Khuzestan, Iran. This study identified the temperature vegetation dryness index (TVDI) as a critical parameter for estimating RZSM in combination with the Sentinel-1 SAR data under high vegetation conditions. In both methods, the RFR algorithm outperformed, with similar performance, the XGBoost, SVR, and M5P algorithms in estimating soil surface moisture (R^2=0.89, RMSE=0.04 cm^3cm^{-3}). However, the accuracy of the RFR algorithm decreased with increasing depth for both the optical-thermal and combined SAR and optical-thermal RS data. This decrease was more pronounced in the combined approach, particularly for the root zone, where the RMSE reached approximately 0.073 cm^3cm^{-3}. Accordingly, the key findings demonstrated that the optical-thermal RS data outperformed the SAR RS data for retrieving RZSM in high-vegetated areas. However, combining TVDI with SAR data is a substantial improvement that opens a new path in radar-based RZSM estimation methods under high vegetation conditions.

Keywords: Soil moisture retrieval; Landsat-8; Sentinel-1 SAR; Machine learning algorithms; Agricultural areas.

1. Introduction

Soil moisture, or soil water content, is a crucial variable influencing agricultural management, biophysical and hydrological cycles, environmental and ecological activities, land surface energy partition into sensible and latent heat flux, and land surface-atmosphere interactions (Toth et al., 2019). Moreover, the spatio-temporal monitoring of soil moisture changes can help characterize soil fertility, irrigation scheduling, and yield prediction (Gill et al.,

2006). In the past decade, soil moisture mapping and monitoring at various levels, from field to global scales, have evolved considerably (Baghdadi et al., 2017; Wang et al., 2022). Nonetheless, locally measuring soil moisture at fine spatial scales across a vast domain is challenging because it necessitates additional field and laboratory work, as well as the problem of extrapolating measured values due to inhomogeneities, soil texture, and micro-topography, which will inevitably result in substantial inaccuracies (Gill et al., 2006). These limitations can be overcome by remote sensing (RS) techniques and technologies, which provide greater spatial and temporal coverages and have the advantages of being fast, economical, and non-destructive (Adab et al., 2020; Wang et al., 2022).

The main RS methods for monitoring soil moisture are based on optical, thermal, and active and passive microwave Earth observations. Several satellite missions, including Soil Moisture and Ocean Salinity (SMOS), Advanced Scatterometer (ASCAT), and Soil Moisture Active Passive (SMAP), have been utilized to estimate soil moisture successfully (Brocca et al. 2011; El Hajj et al., 2018; Min et al., 2022). These soil moisture products have limited usability at the farm and field scales due to their large spatial resolutions of several kilometers (Zhang et al., 2021). Active microwave sensors in Band-C, such as Radarsat-2 and Sentinel-1 synthetic aperture radar (SAR), have effectively addressed the challenges of monitoring near-surface soil moisture (SSM) over field areas by directly correlating the microwave backscattering to the soil water content (Hosseini and McNairn, 2017; Nguyen et al., 2022). The main features that highlight microwaves are their low sensitivity to clouds, sunlight, and all atmospheric conditions. However, achieving a satisfactory estimation can be challenging because measurements are affected significantly by surface roughness and water content, especially in the presence of moderate to heavy vegetation. In contrast, optical and thermal RS have the benefit of generating high-resolution maps. Due to their physical relationship with soil moisture in vegetation conditions, they are frequently used to predict soil moisture (Guo et al., 2022; Ryu et al., 2021). However, the accuracy of this approach is easily affected by weather conditions. Notably, active microwaves, optical, and thermal RS can only detect soil moisture changes in the upper surface layer. Therefore, it is necessary to determine an appropriate approach to estimate root zone soil moisture (RZSM) with a high spatial resolution at farm and field scales.

Based on the physical nature of electromagnetic waves, using RS alone to retrieve RZSM in dense agricultural areas has significant flaws in the theory that the effect of vegetation cover must be eliminated for more precise prediction. Also, soil properties and topography are crucial parameters that affect the spatial-temporal changes in soil moisture (Nguyen et al., 2022). So, these properties can be used along with RS methods to increase soil moisture estimation accuracy. However, the high volume of required data in estimating soil moisture using different RS approaches and soil properties for linear regression analysis may limit the applicability of some statistical hypotheses, such as further data, nonlinearity, heterogeneity, and several others (Yeh and Lien, 2009). This indicates a fundamental requirement to improve and develop RS techniques for a more satisfactory RZSM determination at farm and field scales.

In recent years, several researchers have developed methods to avoid complicated physical relationships and effectively address nonlinear problems to meet application needs better (Carranza et al., 2021; Toth et al., 2019; Wang et al., 2022). Among these methods, deep learning models, such as convolutional neural networks (CNN),

recurrent neural networks (RNN), and deep neural networks (DNN), have demonstrated higher accuracy in soil moisture estimation. However, these models have a significant drawback: their requirement for large datasets poses a severe challenge in regions with limited data availability (Guo et al., 2022). This limitation restricts their practical application in many real-world scenarios. On the other hand, machine learning (ML) models such as artificial neural networks (ANN), decision trees (DT), support vector machines (SVM), random forests (RF), and extreme gradient boosting (XGBoost) are widely and successfully employed for soil moisture prediction due to their simplicity and greater interpretability (Fathololoumi et al., 2020; Babaeian et al., 2021). These models can work effectively with smaller datasets, making them more suitable for regions where data collection is limited.

In this context, Acharya et al. (2021) employed tree-based models and ANN to model field soil moisture in the Red River Valley of the North. They demonstrated that tree-based models outperform ANN in training speed, ease of parameter adjustment, and overall predictive performance. Similarly, Adab et al. (2020) found that the RF model outperformed both SVR and ANN in estimating soil moisture content (SMC). Their findings highlighted that ANN is prone to overfitting, significantly reducing its generalizability. Furthermore, ANN's performance is heavily dependent on its network architecture and the complexity of the sample data, making it less stable in varying conditions. On the other hand, tree-based models offer an interpretable structure, allowing for a deeper understanding of the importance of different features and their interactions, even when dealing with limited data (Al-Aizari et al., 2024). This interpretability is crucial for applications like soil moisture estimation, where understanding the contributing factors is essential. In contrast, due to their complex architectures, ANN models are often perceived as black boxes, where the internal decision-making processes are opaque and less transparent to users. Moreover, ANNs typically require a large volume of data to achieve optimal performance, and their predictive power diminishes, especially when validation data falls outside the training data range or when using small datasets (Melesse et al., 2020). Considering these factors and aiming for higher accuracy and better interpretability, tree-based models were selected for this study.

Carranza et al. (2021) estimated RZSM using RF and meteorological data, leaf area index, and hydraulic soil data. They showed that RF could obtain the spatiotemporal variability of soil moisture and estimate RZSM with reasonable accuracy, while RF predictions for extreme dry and wet conditions were less accurate. However, it is difficult to generalize their approach on the scale of agricultural areas due to the limitations of interpolation and extrapolation of the auxiliary variables and the spatial heterogeneity of soil texture. To address these limitations, this study utilized vegetation, soil, and moisture indices derived from Landsat-8 satellite images to obtain RZSM data with better spatial-temporal consistency.

In microwave RS approaches, several investigations have demonstrated that using the normalized difference vegetation index (NDVI) as the sole vegetation descriptor enables the computation of vegetation effects on the total backscattered coefficients with adequate precision (Baghdadi et al., 2017; Liu et al., 2022). NDVI is widely employed due to its simple estimation, straightforward availability at various spatial and temporal resolutions, and ability to eliminate noise induced by changing sun angles, topographic illumination, clouds, or shadow (Kumari et al., 2021). Using a DT model, Bazzi et al. (2022) investigated the potential of the Sentinel-1 SAR data to detect

irrigation events in corn, soybean, sorghum, and potato fields. The results showed that developed vegetation cover was a significant constraint in detecting irrigation events. Accordingly, NDVI values higher than 0.70 reduced the accuracy of irrigation detection to less than 40%, while moderate vegetation cover (NDVI<0.60) yielded approximately 60% accuracy. On the other hand, Ryu et al. (2021) suggested a relationship between surface soil moisture derived from the global land data assimilation system (GLDAS) and the temperature vegetation dryness index (TVDI) using linear regression. This approach estimates SSM by analyzing the triangular/trapezoidal feature space, incorporating temperature variability and its association with vegetation density. Building on this foundation, this study's key innovation is the use of TVDI to enhance the prediction of RZSM at the field scale. This approach is particularly effective in addressing the limitations associated with NDVI in microwave RS under dense vegetation conditions. As per our review, this is the first study to specifically investigate the effectiveness of combining TVDI with Sentinel-1 SAR data over agricultural lands. This innovative approach offers a more accurate and robust solution for RZSM estimation in environments where more commonly used vegetation indices (VIs), like NDVI, may not perform optimally.

Most previous studies on soil moisture estimation using RS data, whether optical, thermal, or SAR, have focused on SSM in areas with low vegetation cover. However, none of these studies has comprehensively addressed the estimation of RZSM in regions with dense vegetation cover. Therefore, this study aims to address this gap by employing two approaches: the first using optical-thermal information from Landsat-8, and the second combining VIs derived from Landsat-8 data with SAR data from Sentinel-1 in advanced ML algorithms for accurate RZSM estimation. The proposed approach, which also incorporates climatic data and the physical and hydraulic properties of the soil, enhances the accuracy and comprehensiveness of RZSM estimation compared to earlier studies, which primarily relied on coarse resolution data and more straightforward modeling techniques. Additionally, this study provides a more detailed assessment of RZSM estimation at the field scale, which can assist farmers and water resource managers in improving irrigation decision-making and enhancing agricultural productivity. To achieve the primary objectives of this study, the following research questions were posed: (i) Can integrating meteorological parameters and soil physical and hydraulic properties with Landsat-8 data using ML methods improve the accuracy of field-scale RZSM estimation? (ii) Is retrieving RZSM using Sentinel-1 SAR data and ML algorithms in the areas with dominant vegetation cover possible? (iii) Which vegetation descriptor, NDVI or TVDI, has shown superior efficacy when combined with the Sentinel-1 SAR RS approach? (iv) Can ML models (M5P, SVR, XGBoost, and RFR) effectively monitor RZSM and identify critical variables? (v) Can including SSM data points enhance RZSM retrieval using the ML methods?

2. Materials and Methods

2.1. Study area

Field experiments were conducted at Hakim Farabi Sugarcane Agro-Industrial Company (HFSAIC) in the Khuzestan province of Iran (latitude: 30° 54′ 1″ N to 31° 3′ 34″ N, longitude: 48° 31′ 5″ E to 48° 39′ 4″ E Fig. 1). The area exhibits sparse precipitation and intense evaporation. The climate is arid and semi-arid with an average

yearly temperature of 22.92 ℃, annual rainfall of 203 mm, and 2270.24 mm yr^{-1} of yearly evaporation from open pans. Its average altitude is 6 m above the mean sea level, and the slope of the lands of this unit is between 0.1 to 0.2 per thousand. This region's primary agricultural crop is sugarcane. The study area, with a total area of 14000 hectares, was evenly partitioned into rectangular units (about 480 units) with 25 hectares (250 m × 1000 m), of which nine farms were selected for this study. The soil texture in the region ranges from moderate to very heavy. The sugarcane farms had identical planting, harvesting, and agricultural operations. The field irrigation system was surface irrigation, and the total irrigation water consumption was 3000 mm. Overall, the information presented in this section was obtained from the HFSAIC Research Center and is based on local meteorological reports, soil and agronomy studies, and direct field observations conducted in the study area.

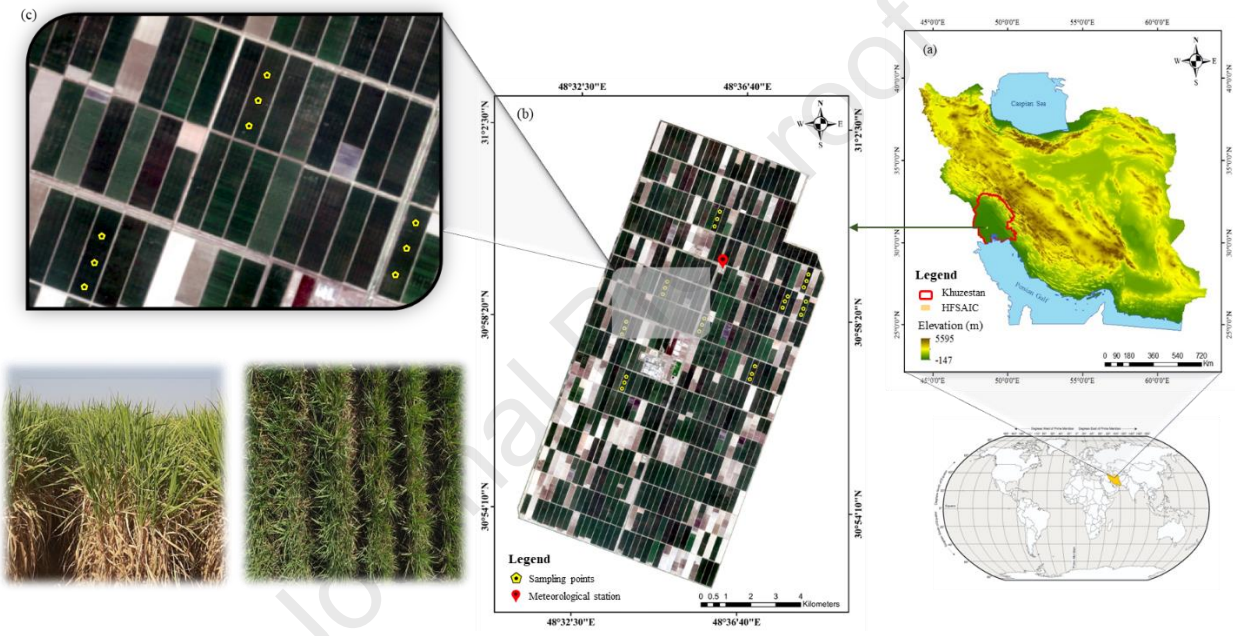


Fig. 1. Study area. (a) The location of the case study, (b) Overview of the HFSAIC with sampling locations, and (c) Detailed view of a subset of the sampling distribution

2.2. Data collection and preparation

During the sugarcane crop growth seasons, soil moisture measurements were taken from twenty-seven different locations of the specified farms at five depths: 0-10, 10-30, 30-50, 50-70, and 70-90 cm. Several samples were taken from different plots with various crop ages inside the research area to measure the SMC from November 2019 to October 2020 crop years. Also, to eliminate inadvertent measurement error, each sampling location's SMC was measured three times, and the ultimate mean value was used to determine the sampling point's actual SMC. Over 4400 samples were gathered from the study area during the sugarcane season.

The soil samples were air-dried, crushed, and sieved through a 2 mm mesh at the HFSAIC laboratory. Soil properties, e.g., textural characterization (clay, silt, and sand %), dry bulk density (BD), water content at field

capacity (θ_{FC}), water content at the permanent wilting point (θ_{PWP}), and porosity (ϕ) were measured (Klute, 1988). The ϕ was derived using the measured BD and considering a mean value of soil particle density of 2.65 g cm^{-3}, which is usual in croplands containing silica-rich soils and devoid of a substantial quantity of organic matter (Flint and Flint, 2002). ROSETTA pedotransfer functions (PTFs) were used to derive soil hydraulic parameters (Schaap et al., 2001). The PTFs provide statistical estimates of hard-to-obtain hydraulic soil parameters based on easily measured fundamental properties such as BD and soil texture. For this research, the van Genuchten (1980) model was used to estimate saturated hydraulic conductivity (Ks), residual volumetric moisture (θ_r), saturated volumetric moisture (θ_s), scaling parameter(α), and shape parameter (n). Some physical and hydraulic soil properties are listed in Table 1.

Table 1. Basic statistics of soil physical and hydraulic properties in the root zone of the study area.

Soil Property	Min					Max					Mean				
	0-10	**10-30**	**30-50**	**50-70**	**70-90**	**0-10**	**10-30**	**30-50**	**50-70**	**70-90**	**0-10**	**10-30**	**30-50**	**50-70**	**70-90**
Sand (%)	7	8	2.2	3	2.2	20	22	21	20	19	12.8	13.2	12.77	13.5	11.4
Silt (%)	45	43	45	46	44	56	57	59	57	58	51.1	52.3	51.47	50.2	50.1
Clay (%)	33	33	32	33	34	44	45	44	45	46	36.1	36.7	35.75	36.7	38.5
BD ($g\ cm^{-3}$)	1.3	1.33	1.35	1.3	1.3	1.5	1.52	1.6	1.5	1.48	1.4	1.41	1.44	1.4	1.39
θ_{FC} ($cm^3\ cm^{-3}$)	0.33	0.33	0.34	0.36	0.37	0.41	0.41	0.4	0.42	0.44	0.38	0.38	0.39	0.38	0.39
θ_{PWP} ($cm^3\ cm^{-3}$)	0.14	0.14	0.13	0.15	0.15	0.19	0.19	0.19	0.18	0.20	0.17	0.17	0.18	0.16	0.17
Φ ($cm^3\ cm^{-3}$)	0.43	0.42	0.41	0.43	0.44	0.50	0.50	0.49	0.51	0.51	0.47	0.46	0.45	0.47	0.48
Ks ($cm\ d^{-1}$)	11.6	12.14	12.25	11.59	11.77	12.8	14.2	14.3	14.7	15.55	12.3	12.6	12.65	12.7	12.93

The meteorological station located in HFSAIC provided the weather data for this study. The climatic station collected air temperature (T), relative humidity (RH), wind speed (W), sunshine hours (S_h), solar radiation (R_n), and class A evaporation pan (E).

2.3. Satellite data and image analysis

2.3.1. Data acquisition

This study aimed to estimate soil moisture from two satellite Earth observations, i.e., level-1 Landsat-8 multispectral data and Sentinel-1 C-band dual polarimetric SAR imagery. The Landsat-8 data were acquired from the specific path and row coordinates (path 165, row 39) covering the study area. These images are freely accessible to global users through the United States Geological Survey's data distribution website (https://glovis.usgs.gov/) (USGS). In addition, high-resolution Level-1 ground range detected (GRD) Sentinel-1 images with 10m×10m pixels and dual-polarization (vertical-vertical (VV) polarization and vertical-horizontal (VH) polarization) acquired in interferometric wide swath (IW) mode were obtained from the ESA Copernicus data hub

(https://scihub.copernicus.eu). Table 2 provides the acquisition dates from Sentinel-1 and cloud-free Landsat-8 satellites during the specified period.

Table 2. Details of satellite data used in the study.

Satellite	Date of acquisition	Resolution
Landsat-8	11-Nov, 27-Nov, 29-Dec. (2019); 30-Jan, 15-Feb, 2-Mar, 3-Apr, 21-May, 6-Jun, 22-Jun, 24-Jul, 9-Agu, 25-Agu, 10-Sep, 26-Sep, 12-Oct. (2020)	OLI (30 m) TIRS (100 m)
Sentinel-1	10-Nov, 27-Nov, 28-Dec. (2019); 31-Jan, 14-Feb, 2-Mar, 20-May, 6-Jun, 23-Jun, 24-Jul, 10-Agu, 24-Agu, 10-Sep, 27-Sep, 11-Oct. (2020)	10 m

2.3.2. Data preprocessing

The Landsat-8 images were processed via two main steps presented in Fig. 2. The environment for visualizing images (ENVI) 5.3.1 software preprocessed the Landsat-8 images (ITT Systems, ITT Exelis, Herndon, VA, USA). The fast line-of-sight atmospheric analysis of spectral hypercubes (FLAASH) module in ENVI 5.3.1 was applied for atmospheric correction to retrieve surface reflectance accurately. The study area was then extracted from the corrected scenes, and the spatial resolution of the images was resampled to 10 meters using the nearest neighbor method. Subsequently, standard and efficient spectral indices were used to develop an effective combination for constructing the retrieval model.

In this study, in addition to spectral indices, the TVDI based on the land surface temperature (LST) - NDVI space was utilized to capture spatial and temporal variations in soil moisture. As described in Table 3, LST_{Max} and LST_{Min} represent the surface temperature under dry and wet soil conditions derived from the LST-NDVI trapezoid for a specific location (satellite scene). These values are calculated using the dry and wet edges of the trapezoid for each NDVI value:

$$LST_{Max} = i_d + S_d\,NDVI \quad (1)$$

$$LST_{Min} = i_w + S_w\,NDVI \quad (2)$$

In these equations, i_d and S_d denote the intercept and slope of the dry edge, respectively, while i_w and S_w represent the intercept and slope of the wet edge. For more details, refer to Moran et al.'s (1994) and Sadeghi et al. (2017) research. TVDI and other vegetation, soil, and moisture indices used in this study have been identified as sensitive indices of soil moisture conditions. This makes it particularly valuable for monitoring areas with heterogeneous soil moisture levels, such as arid and semi-arid regions. Table 3 lists the spectral indices derived from Landsat-8 used in this study.

Table 3. Vegetation, soil, and moisture indices from optical and thermal RS observations.

Variable	Acronym	Equation	Reference
Normalized Difference Vegetation Index	NDVI	$\frac{\rho_{NIR}-\rho_{Red}}{\rho_{NIR}+\rho_{Red}}$	Rouse et al. (1974)
Fraction Normalized Difference Vegetation Index	fNDVI	$(1-NDVI)^{0.625}$	Agam et al. (2007)
Normalized Multi-Band Drought Index	NMDI	$\frac{\rho_{NIR}-(\rho_{SWIR1}-\rho_{SWIR2})}{\rho_{NIR}+(\rho_{SWIR1}-\rho_{SWIR2})}$	Wang & Qu (2007)
Normalized Difference Water Index	NDWI	$\frac{\rho_{NIR}-\rho_{SWIR1}}{\rho_{NIR}+\rho_{SWIR1}}$	Gao (1996)
Modified Normalized Difference Water Index	MNDWI	$\frac{\rho_{Green}-\rho_{SWIR1}}{\rho_{Green}+\rho_{SWIR1}}$	Xu (2006)
Normalized Soil Moisture Index	NSMI	$\frac{\rho_{SWIR1}-\rho_{SWIR2}}{\rho_{SWIR1}+\rho_{SWIR2}}$	Haubrock et al. (2008)
Visible and Shortwave Infrared Drought Index	VSDI	$1-[(\rho_{SWIR1}-\rho_{Blue})+(\rho_{Red}-\rho_{Blue})]$	Zhang et al. (2013)
Global Vegetation Moisture Index	GVMI	$\frac{(\rho_{NIR}+0.1)-(\rho_{SWIR1}+0.02)}{(\rho_{NIR}+0.1)+(\rho_{SWIR1}+0.02)}$	Ceccato et al. (2002)
Moisture Stress Index	MSI	$\frac{\rho_{SWIR1}}{\rho_{NIR}}$	Hunt. & Rock (1989)
Enhanced Vegetation Index	EVI	$\frac{2.5(\rho_{NIR}-\rho_{Red})}{\rho_{NIR}+6\,\rho_{Red}-7.5\rho_{Blue}+1}$	Huete et al. (2002)
Soil Adjusted Vegetation Index	SAVI	$1.25\frac{(\rho_{NIR}-\rho_{Red})}{(\rho_{NIR}+\rho_{Red}+0.25)}$	Huete (1988)
Ratio Vegetation Index	RVI	$\frac{\rho_{NIR}}{\rho_{Red}}$	Jordan (1969)
NIR Transformed Reflectance	NTR	$\frac{(1-\rho_{NIR})^2}{2\,\rho_{NIR}}$	Kubelka & Munk (1931) and Babaeian et al. (2021)
Temperature Vegetation Dryness Index	TVDI	$\frac{LST-LST_{Min}}{LST_{Max}-LST_{Min}}$, $LST=\gamma\times\left[\frac{1}{g}\times(\psi_1\times L_{sensor}+\psi_2)+\psi_3\right]+\delta$	Moran et al. (1994), Jiménez-Muñoz et al. (2008)
Albedo	–	$0.356\,\rho_{Blue}+0.130\,\rho_{Green}+0.373\rho_{Red}+0.085\rho_{NIR}+0.072\rho_{SWIR1}+0.072\rho_{SWIR2}-0.0018$	Liang (2001)
Brightness	–	$0.3029\rho_{Blue}+0.2786\rho_{Green}+0.4733\rho_{Red}+0.5599\rho_{NIR}+0.508\rho_{SWIR1}+0.1872\rho_{SWIR2}$	Baig et al. (2014)
Greenness	–	$-0.2941\rho_{Blue}-0.243\rho_{Green}-0.5424\rho_{Red}+0.7276\rho_{NIR}+0.0713\rho_{SWIR1}-0.1608\rho_{SWIR2}$	
Wetness	–	$0.1511\rho_{Blue}+0.1973\rho_{Green}+0.3283\rho_{Red}+0.3407\rho_{NIR}-0.7117\rho_{SWIR1}-0.4559\rho_{SWIR2}$	

The Sentinel-1 images were chosen based on their acquisition times to ensure that the SAR and optical data were approximately synchronous in pairs. Afterward, the SAR images underwent preprocessing with the Sentinel Application Platform (SNAP) version 9.0.0 of ESA with the Sentinel-1 Toolbox version 8.0.5. The preprocessing procedures include applying an orbit file, removing the thermal noise, radiometric calibration, speckle filter to reduce the speckle noise, terrain correction, and conversion to dB using a logarithmic transformation (Foumelis et al., 2018). Terrain correction was performed with the Range-Doppler terrain correction module using the Shuttle Radar Topography Mission (SRTM) 1-sec digital elevation model (DEM). From it, the slope and aspect were computed with ArcGIS 10.8. It has been proven that vegetation indices in the form of the backscatter intensity ratio are efficient for characterizing vegetation morphology and estimating soil moisture using Sentinel-1 images (Bhogapurapu et al., 2022). Therefore, in addition to extracting the two principal polarizations (VV, VH), the vegetation index (VI) obtained from the dual SAR dataset was used to reduce the vegetation's effect on soil backscatter. The DpRVIc index can be calculated using equations 3 and 4, where q, the ratio parameter, is determined by dividing the backscattering coefficients of VH and VV.

$$DpRVI_C=\frac{q(q+3)}{(q+1)^2} \quad (3)$$

$$q = \frac{\sigma VH}{\sigma VV} \quad (4)$$

228

229 **2.4. Methodology**

230 **2.4.1. Overall retrieval framework**

231 This study used diverse input data, including RS information, meteorological data, soil properties, and topographic
232 features, to accurately estimate RZSM using four ML algorithms (M5P, SVR, XGBoost, and RFR). Preparing these
233 input data meticulously and appropriately before applying them to the algorithms is essential for improving the
234 accuracy of the results. The following will detail the process and strategies to ensure reliable estimations. A critical
235 aspect of data preparation involved addressing the study area's spatial and temporal dimensions. Meteorological
236 variables were considered uniform across the study area and only varied by measurement days. In contrast, the
237 Kriging interpolation method in ArcMap 10.8.2 generated spatially continuous maps based on field sampling data
238 for soil physical and hydraulic properties.

239 Following the meticulous preparation of the input data, the initially vital step in the modeling process was to
240 determine the effective parameters for estimating soil moisture. Conducting this preliminary step before modeling
241 can enhance result quality, mitigate the risk of overfitting, and significantly reduce training time. To this end, the
242 information gain ratio (IGR) technique was employed to identify the most influential parameters for predicting
243 RZSM. Subsequently, the top-ranked parameters exhibiting the highest IGR values were integrated into soil
244 moisture estimation algorithms, and their accuracy was evaluated (Gibson, 2020).

245 The predictor variables were divided into two datasets: 70% of the data (comprising six fields and 293 points) were
246 allocated for the training phase, while the remaining 30% (3 fields and 125 points) were reserved for the testing
247 phase. The training dataset was employed to develop the ML algorithms, and the testing dataset was used for
248 validation. Notably, the normalization was performed before modeling to improve the data's integrity and minimize
249 redundancy.

250 In the subsequent step, recognizing that the performance of ML models is highly contingent on properly tuning
251 hyperparameters, this study employed a combination of 5-fold cross-validation (CV) and grid search (GS) to
252 optimize these parameters. The 5-fold CV method involves partitioning the training data into five subsets,
253 sequentially excluding one subset to serve as the validation subset and using the remaining data to fit the model.
254 This process is repeated five times to ensure a robust estimate of the test error rate. The models were implemented in
255 Python, leveraging essential libraries such as Pandas, scikit-learn, NumPy, m5py, and XGBoost for data processing,
256 model development, and evaluation. Fig. 2 depicts the framework for estimating RZSM in this paper.

257 To address the primary research questions, this study employed various ML algorithms to estimate RZSM in three
258 distinct scenarios: (1) Estimating RZSM using optical and thermal RS methods, meteorological data, and soil
259 physical-hydraulic properties. (2) Estimating RZSM using Sentinel-1 SAR data combined with NDVI, TVDI, and

topographic parameters. (3) Incorporation of SSM as an auxiliary input parameter into scenarios 1 and 2 to assess potential improvements in RZSM estimation.

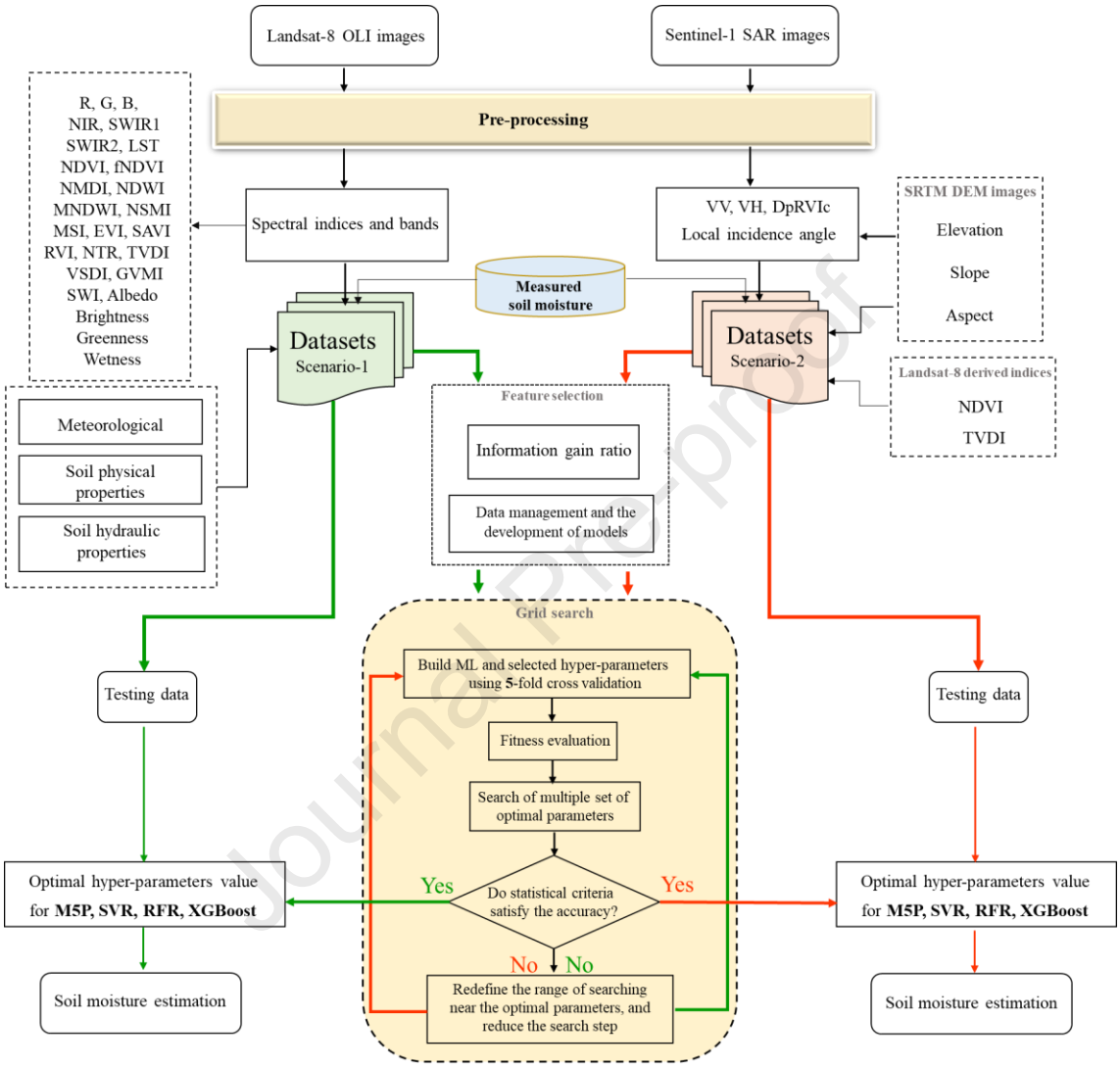


Fig. 2. Flowchart of the study procedure

2.4.2. Machine learning algorithms

2.4.2.1. M5- pruned (M5P) model tree

Wang and Witten (1996) rebuilt and proposed the M5P algorithm from the M5 algorithm, a binary DT using a linear regression function at the leaf (terminal node) to predict continuous numerical properties. The main advantage of M5P is that it can perform better in the data set than M5 by reducing tree size. For more explanation, refer to Quinlan's study (1992). This model analyzes the data's implicit patterns and relations based on several principles,

rules, and regression equations. However, this part of the operation in the other artificial intelligent models, such as ANN, is hidden, and just the analysis outcomes are presented (Wang and Witten, 1996). M5P tree development consists of three primary steps: tree construction, tree pruning, and tree smoothing. The M5P model tree construction procedure aims to maximize a metric known as the standard deviation reduction. The standard deviation reduction (SDR) is determined using the following formula:

$$\mathrm{SDR} = \mathrm{sd}(\mathrm{S}) - \sum \frac{\mathrm{S_i}}{|\mathrm{S}|} \times \mathrm{sd}(\mathrm{S_i}) \tag{5}$$

$$\mathrm{sd}(\mathrm{S_i}) = \sqrt{\sum_{1}^{\mathrm{N}} \frac{(\mathrm{S_i} - \bar{\mathrm{S}})^2}{\mathrm{N} - 1}} \tag{6}$$

$$\bar{\mathrm{S}} = \sum_{1}^{\mathrm{N}} \frac{\mathrm{S_i}}{\mathrm{N}} \tag{7}$$

where S_i is the set resulting from splitting the node based on a specified attribute, S is the set of the data records reaching the node, and sd is the standard deviation (Wang and Witten, 1996).

2.4.2.2. Support vector regression (SVR)

The support vector machines were developed by Cortes and Vapnik (1995), and they can deal with classification (SVM) or regression (SVR) problems. SVR offers the advantages of a simple structure, robust adaptability, and potent capacity for addressing difficulties with few samples and nonlinear and high-dimensional data. The objective of backup vector regression is to identify the function f(x) for training patterns x_i. So that it has the largest possible margin of training values y_i. The SVR model's regression function can be represented as:

$$\mathrm{f(x)} = \mathrm{w} \times \phi(\mathrm{x}) + \mathrm{b} \tag{8}$$

where f represents the regression function, w and b stand for the weight and the bias, respectively, and ϕ is the transfer function. The problem of regression might be stated as follows:

$$\text{Minimize: } \frac{1}{2}\|\mathrm{w}\|^2 + \mathrm{C}\sum_{\mathrm{i=1}}^{\mathrm{N}}(\xi_\mathrm{i} + \xi_\mathrm{i}^*) \tag{9}$$

$$\text{Subject to: } \begin{cases} \{\mathrm{y_i} - \mathrm{f(x)} \le \varepsilon + \xi_\mathrm{i} \\ \mathrm{f(x)} - \mathrm{y_i} \le \varepsilon + \xi_\mathrm{i} \\ \xi_\mathrm{i}, \xi_\mathrm{i}^* \ge 0, \mathrm{i} = 1,2,3, \ldots., \mathrm{N} \end{cases} \tag{10}$$

where ε is the boundary value, ξ_i and ξ_i^* are the slack variables and C is the penalty parameter. Using the Lagrange multipliers, the optimization problem largely converts to quadratic programming, and a nonlinear regression function solution can be expressed as:

$$\mathrm{f(x)} = \sum_{\mathrm{i=1}}^{\mathrm{N}}(\alpha_\mathrm{i} - \alpha_\mathrm{i}^*)\mathrm{K(x,y)} + \mathrm{b} \tag{11}$$

where K (x, x_i) is the kernel function and α_i, α_i^* are the dual variables.

It should be noted that the SVR method's generalization ability considerably depends on the choice of the kernel function. This study employed various kernel functions, including linear, Polynomial (Poly), Sigmoid, and Gaussian (RBF).

2.4.2.3. Random forest regression (RFR)

RF is an ensemble learning method based on statistical theory that combines the concepts of DTs and bagging to solve classification and regression problems. Leo Breiman developed this algorithm at the end of the 90s. RF is an advanced method based on the combination of bootstrap aggregation. The algorithm generates multiple bootstrap samples with replacements from the original training data set (about 67 %) to create multiple regression trees (ntree), known as 'in bag' data. In comparison, the excluded data set (about 33 %) is known as 'out-of-bag' data (OOB) (Breiman 2001). In addition, the OOB samples are used to measure the variable importance, which indicates the predictive power of each variable and is used to optimize the selection of input parameters.

Consequently, this model simplifies the method, reduces the computational costs of analysis, and helps understand the relationship between variables and the dependence of one variable on another. The variable importance measure is based on the percentage increase in mean squared error (% IncMSE). Finally, the predicted value of an observation is calculated by averaging over all the trees. Breiman (2001) has provided a more detailed description of RF methods and parameters.

2.4.2.4. Extreme gradient boosting (XGBoost)

In recent years, the XGBoost model, a scalable tree-boosting system, has emerged. Chen and Guestrin (2016) introduced this algorithm to enhance the efficacy and speed of gradient-boosted decision trees. The XGBoost algorithm employs additive learning to produce a powerful learner by combining several feeble learners, i.e., each tree (Chen and Guestrin, 2016). In the sequential modeling procedure of XGBoost, each DT relies on the prior tree's result to generate an improved predictor (Zhu and Zhu, 2019). Also, the XGBoost model increases the weight of incorrectly classified training samples (with significant errors) and decreases the weight of those satisfactorily classified. The prior wrongly categorized subsamples are processed several times with increased vigilance to minimize the error rate (Chen and Guestrin, 2016). The final prediction computed by XGBoost is based on the sum of all decision trees' weighted contributions. The calculation formula is as follows:

$$f_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = f_i^{t-1} + f_i(x_i) \quad (12)$$

where $f_i(x_i)$ is the learner of step t, $f_i(x_i)$ and f_i^{t-1} are steps t and t-1, respectively, and x_i is the input variable.

The XGBoost model has the advantages of powerful generalization ability, overfitting prevention, high expandability, and fast computing speed.

2.5. Accuracy assessment of model performance

To quantitatively evaluate the performance of the developed models, the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) were used as evaluation metrics.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(O_i - E_i)^2}{\sum_{i=1}^{N}(O_i - \overline{O})^2} \quad (13)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(O_i - E_i)^2}{N}} \quad (14)$$

$$MAE = \frac{\sum_{i=1}^{N}|O_i - E_i|}{N} \quad (15)$$

where N shows the number of the output data series; O_i and $\overline{O}$ are the measured and average of actual SMC, respectively, and E_i is predicted SMC.

3. Results and Discussion

3.1. Spatial variability of soil moisture

Fig. 3 provides the average measured soil moisture at observation depths of 0-10, 10-30, 30-50, 50-70, and 70-90 cm. At different depths, the measured soil moisture ranged from 0.11 to 0.63 cm^3cm^{-3}, with a standard deviation ranging from 0.11 to 0.12 cm^3cm^{-3}. Also, the maximum soil moisture was recorded during September at all depths, when the crop evapotranspiration (ET_c) peaked. Fig. 4 shows the relationship between soil moisture's spatial coefficient of variation (CV) and mean soil moisture during the sugarcane growing season. Fatichi et al. (2015) stated that precipitation, topography, and vegetation significantly impact soil moisture variability. Nevertheless, according to the area's topographical and climatic conditions, the soil moisture changes are mainly driven by ET_c (irrigation), vegetation, and soil properties. Also, the highest CV was observed between July and September, coinciding with the peak crop growth and water requirement. According to Fig. 4, RZSM has a lower CV than SSM, which could be due to significant land-atmosphere interactions in a semi-arid region affecting the surface layer (Srivastava et al., 2020).

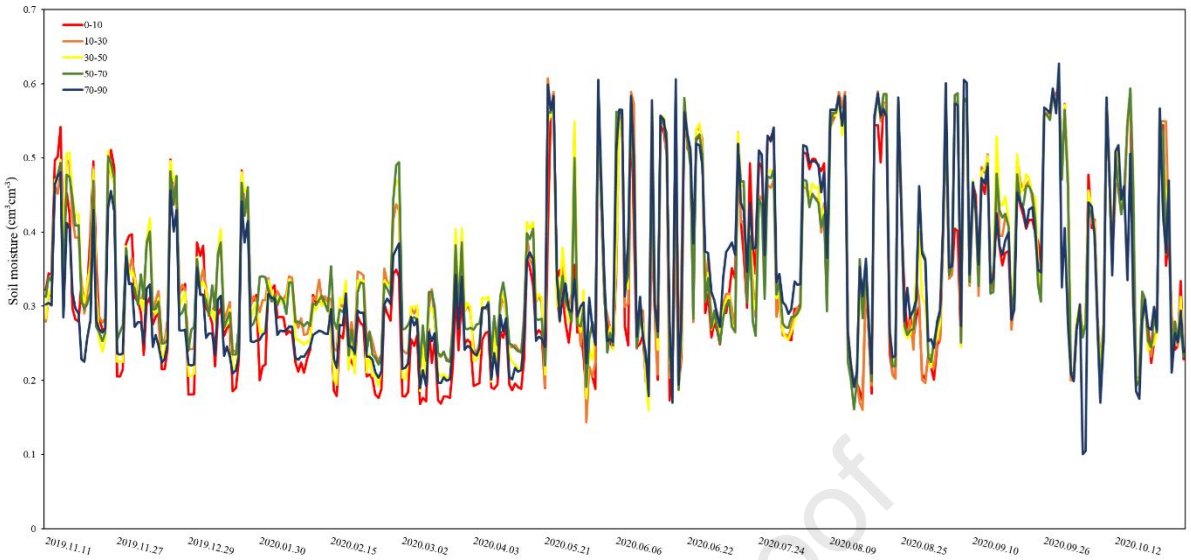


Fig. 3. Time series of soil moisture content at different depths during satellite overpasses (0–10 cm, 10–30 cm, 30–50 cm, 50–70 cm, and 70–90 cm)

338

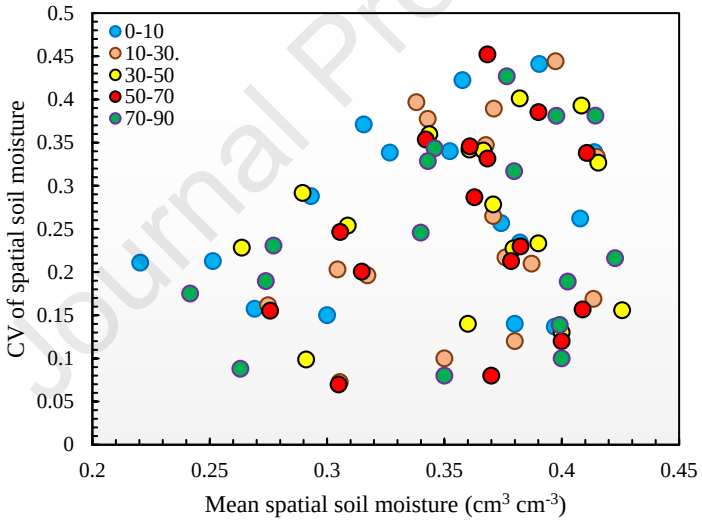


Fig. 4. The spatial CV of soil moisture versus mean soil moisture at different depths 0-10, 10-30, 30-50, 50-70, and 70-90 cm.

339

340 **3.2. Feature selection**

341 The IGR approach was used to assess the sensitivity of the variables in estimating soil moisture in both scenarios.
342 This can aid us in determining which potential predictive features to include in our model. The analysis results in
343 Scenario-1 are shown in Fig. 5, which uses a 5-fold CV technique to determine the IGR of each parameter.
344 According to Fig. 5, the most influential factors for soil moisture estimation detected by IGR were NTR at 0-10 cm
345 depth, R at 10-50 cm depth, and Wetness at 50-90 cm depth. At 0-10 cm depth, factors such as NTR (0.65), Wetness
346 (0.63), Greenness (0.61), NMDI (0.50), TVDI (0.44), Brightness (0.43), LST, B, and NSMI (0.41) demonstrated

higher IGR values, indicating their more significant influence compared to other variables. The relevance of SWIR2 and Albedo (0.39), G and R (0.38), SWIR1, and EVI (0.34) were moderate. Also, MNDWI (0.24), Ks (0.23), R (0.27), θ_{FC} (0.20), and θ_{PWP} (0.19) all demonstrate a below modest level of significance. Some input parameters include VSDI, T, RH, E, W, S_h, Rn, sand, silt, clay, θ_r, α, n, BD, and φ have an IGR value of 0.00, so they are ineffective in predicting soil moisture at 0-10 cm depth. Similarly, Fig. 5 indicates the effective parameters for the rest of the soil depth (10-90 cm). The IGR technique generally indicates that when compared to meteorological variables and soil physical and hydraulic parameters, NTR, Wetness, TVDI, and SWIR2 had the strongest predictive power for soil moisture at all depths investigated (Fig. 5). Some research also demonstrated that spectral information, directly and indirectly, affect soil moisture (Fathololoumi et al., 2020; Ryu et al., 2021). Particularly, spectral indices that can reduce background error to extract new information and facilitate interpreting and processing satellite images can generally reveal soil moisture changes better than individual spectral bands (Nguyen et al., 2022).

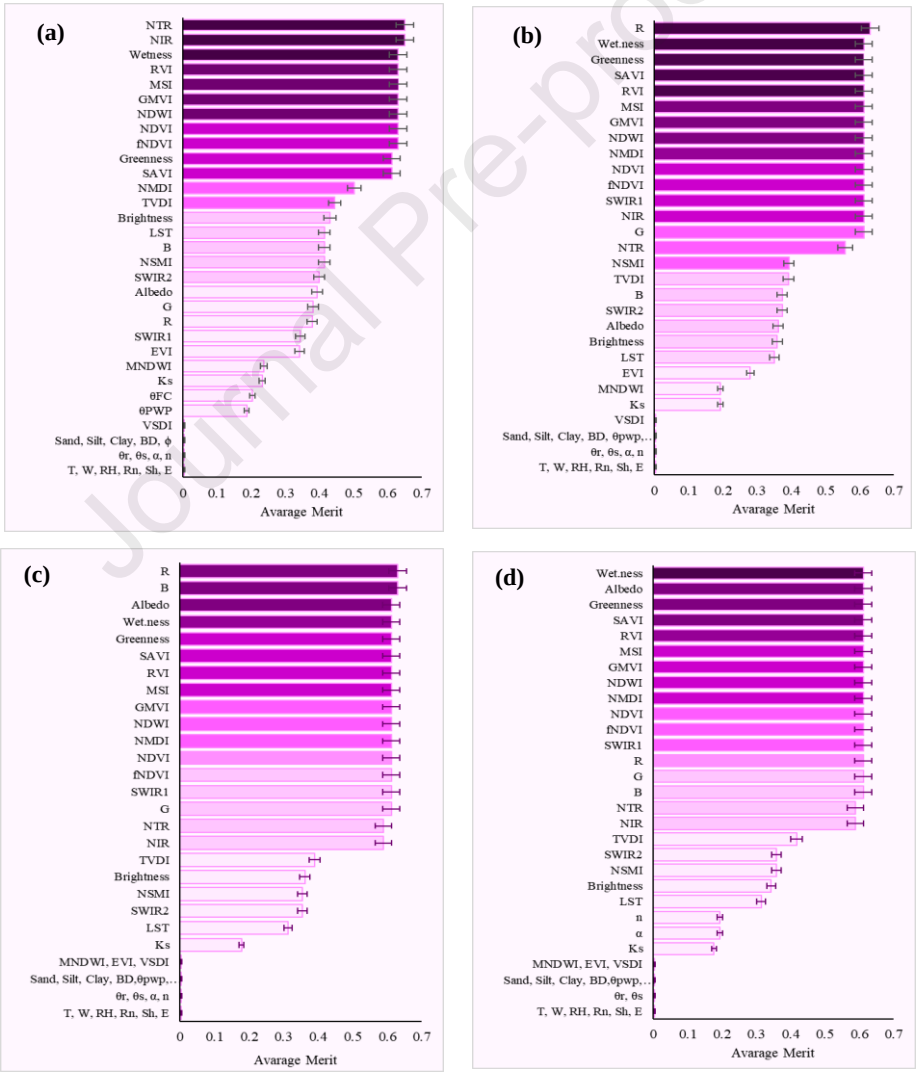
The relationship between VIs and soil moisture in the root zone is affected by vegetation type and climatic conditions (Liu et al., 2012). However, in this study, the response of VIs to soil moisture variations in the root zone was almost similar across different depths. These findings could be attributed to complex processes such as root uptake, capillary action, soil drainage, and ET_c (Alavi et al., 2024; Chang et al., 2012; Holzman et al., 2014). Furthermore, Holzman et al. (2021) reported that the correlation between VIs and soil moisture in the root zone can be linked to the structure and depth of the root system and physical constraints like soil resistance. Specifically, sugarcane roots can extend up to 4.25 meters deep, with approximately 50% of the root biomass located within the first meter of soil (Laclau and Laclau, 2009). As a result, the significant influence of VIs at various soil depths up to 90 cm in our study is likely due to the extensive water absorption by sugarcane roots within this depth. For example, Santos et al. (2014) demonstrated that the soil layer between 0 and 60 cm is where most water uptake by roots occurs in coffee plants, and the highest correlation between the EVI-2 index and soil moisture was observed within this depth.

From a physiological perspective, a reduction in soil moisture prompts stomata to respond to chemical signals produced by drying roots, leading to stomatal closure, reduced internal CO_2 concentration, and decreased transpiration. This process consequently reduces leaf chlorophyll content while the leaf water status remains relatively stable (Davies and Zhang, 1991). Additionally, water stress induces photochemical changes immediately following stomatal closure, but a decrease in leaf water content only occurs when soil moisture reaches a critical threshold (Chaves et al., 2002; Liu et al., 2012). However, a scheduled irrigation management strategy was implemented in our study area, thereby minimizing water stress on the crops. This approach maintained consistent soil moisture levels across the study area (Fig. 3), likely contributing to the observed uniformity in VIs responses across different soil depths.

Compared to other factors, meteorological variables were inefficient in estimating soil moisture across all depths (Fig. 5). Although Araya et al. (2021) reported that precipitation is one of the most crucial variables in determining soil moisture, in our study area, which is in a hot and arid climate, there was no significant precipitation during the

study period. This likely explains the minimal impact of meteorological variables on soil moisture estimation in this region. Generally, meteorological variables play a significant role primarily at larger spatial scales, such as watersheds. In contrast, at smaller spatial scales, like farms and limited areas, factors such as topography and soil properties have a more pronounced effect on soil moisture variability (Karthikeyan and Mishra, 2021).

Among all the physical soil properties examined, θ_{FC} and θ_{PWP} appear to have the most significant impact on soil moisture at the 0-10 cm depth (Fig. 5). Babaeian et al. (2021) confirmed that the increase in correlation is achieved due to the sequential addition of soil physical and hydraulic properties, including soil texture, θ_{FC}, and θ_{PWP}. Soil particle size is crucial in agricultural fields because it affects the pore diameter and their ability to store and retain water. This has been detailed in the study by Wang et al. (2022). However, in the present study, soil particle size did not significantly influence soil moisture prediction at different depths, likely due to the limited variation in soil texture within the study area.



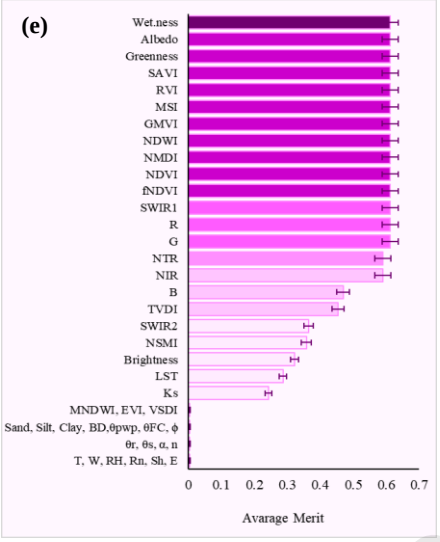


Fig. 5. Evaluating influential factors in soil moisture prediction at different depths (a) 0-10 cm, (b) 10-30 cm, (c) 30-50 cm, (d) 50-70 cm, and (e) 70-90 cm using IGR in Scenario-1.

394

395 Nonetheless, other studies have documented significant correlations between soil moisture and physical soil
396 properties (Acharya et al., 2021; Karthikeyan and Mishra, 2021), indicating that such results may be location-
397 specific.

398 Despite the limited variation in soil texture in the study area, Ks was still identified as a significant factor across all
399 soil depths. This finding suggests that Ks, due to its fundamental role in controlling water movement within the soil,
400 remains a key variable in predicting soil moisture, even under conditions where soil texture does not vary
401 significantly (Acharya et al., 2021). Given the heavy soil texture of the fields, containing 33 to 46% clay, Ks is
402 expected to be a critical factor in estimating RZSM. In Scenario-1, four distinct Models were defined based on the
403 IGR value for each ML method to estimate soil moisture at different depths. It is worth mentioning that different
404 variables may have varying levels of influence depending on the SMC. As a result, the Models were designed for
405 each depth of soil. Table 4 lists the Model inputs in a hierarchical sequence, including meteorological factors, soil
406 hydraulic and physical properties, spectral bands, and the indices obtained from the Landsat-8 satellite.

407 The analysis results in Scenario-2 are shown in Fig. 6 that VH and VV polarization and $DpRVI_C$ from Sentinel-1 and
408 NDVI and TVDI from Landsat-8 images indicate a more substantial relationship with the RZSM than topographic
409 parameters. The TVDI (0.68) and NDVI (0.16) were more significant with high IGR values than other variables of
410 influence, according to the findings of IGR at 0-10 cm depth. Also, the relevance of the local incidence angle and
411 VH (0.08), $DpRVI_C$ (0.062), and VV (0.058) were moderate (Fig. 6). The optical features' high correlation,
412 particularly the vegetation and thermal indices, appears to result from their inherent and indirect effect on soil
413 moisture estimation. Moreover, because vegetation cover dominates the researched area, vegetation and thermal
414 indices could more accurately represent and model the existing reality.

415

Table 4. Combination of different models using the IGR technique in Scenario-1.

Depths (cm)	Model 1	Model 2	Model 3	Model 4
		Input features		
0-10	NTR, NIR, Wetness, RVI, MSI, GVMI, NDWI, NDVI, fNDVI, Greenness, SAVI, NMDI, TVDI, Brightness, LST, B, NSMI, SWIR2, Albedo, G, R, SWIR1, EVI, MNDWI, K_s, θ_{FC}, θ_{PWP}, VSDI, Sand, Silt, Clay, BD, ϕ, θr, θs, α, n, T, W, RH, Rn, S_h, E	NTR, NIR, Wetness, RVI, MSI, GVMI, NDWI, NDVI, fNDVI, Greenness, SAVI, NMDI, TVDI, Brightness, LST, B, NSMI, SWIR2, Albedo, G, R, SWIR1, EVI, MNDWI, K_s, θ_{FC}, θ_{PWP}	NTR, Wetness, Greenness, NMDI, TVDI, Brightness, LST, B, NSMI, SWIR2, Albedo, G, R, SWIR1, EVI, MNDWI, K_s, θ_{FC}, θ_{PWP}	NTR, Wetness, Greenness, NMDI, TVDI, Brightness, LST, B, NSMI, SWIR2, Albedo, G, R, SWIR1, EVI, MNDWI
10-30	R, Wetness, Greenness, SAVI, RVI, MSI, GVMI, NDWI, NMDI, NDVI, fNDVI, SWIR1, NIR, G, NTR, NSMI, TVDI, B, SWIR2, Albedo, Brightness, LST, EVI, MNDWI, VSDI, K_s, θ_{FC}, θ_{PWP}, Sand, Silt, Clay, BD, ϕ, θr, θs, α, n, T, W, RH, Rn, S_h, E	R, Wetness, Greenness, SAVI, RVI, MSI, GVMI, NDWI, NMDI, NDVI, fNDVI, SWIR1, NIR, G, NTR, NSMI, TVDI, B, SWIR2, Albedo, Brightness, LST, EVI, MNDWI, K_s	R, Wetness, NTR, NSMI, TVDI, B, SWIR2, Albedo, Brightness, LST, EVI, MNDWI, K_s	R, Wetness, NTR, NSMI, TVDI, B, SWIR2, Albedo, Brightness, LST, EVI
30-50	R, B, Albedo, Wetness, Greenness, SAVI, RVI, MSI, GVMI, NDWI, NMDI, NDVI, fNDVI, SWIR1, G, NTR, NIR, TVDI, Brightness, NSMI, SWIR2, LST, K_s, EVI, MNDWI, VSDI, θ_{FC}, θ_{PWP}, Sand, Silt, Clay, BD, ϕ, θr, θs, α, n, T, W, RH, Rn, S_h, E	R, B, Albedo, Wetness, Greenness, SAVI, RVI, MSI, GVMI, NDWI, NMDI, NDVI, fNDVI, SWIR1, G, NTR, NIR, TVDI, Brightness, NSMI, SWIR2, LST, K_s	R, NTR, TVDI, Brightness, NSMI, SWIR2, LST, K_s	R, NTR, TVDI, Brightness, NSMI, SWIR2, LST
50-70	Wetness, Albedo, Greenness, SAVI, RVI, MSI, GVMI, NDWI, NMDI, NDVI, fNDVI, SWIR1, R, G, B, NTR, NIR, TVDI, SWIR2, NSMI, Brightness, LST, n, α, K_s, MNDWI, EVI, VSDI, Sand, Silt, Clay, BD, θ_{FC}, θ_{PWP}, ϕ, θs, θr, T, W, RH, Rn, S_h, E	Wetness, Albedo, Greenness, SAVI, RVI, MSI, GVMI, NDWI, NMDI, NDVI, fNDVI, SWIR1, R, G, B, NTR, NIR, TVDI, SWIR2, NSMI, Brightness, LST, n, α, K_s	Wetness, NTR, TVDI, SWIR2, NSMI, Brightness, LST, K_s	Wetness, NTR, TVDI, SWIR2, NSMI, Brightness, LST
70-90	Wetness, Albedo, Greenness, SAVI, RVI, MSI, GVMI, NDWI, NMDI, NDVI, fNDVI, SWIR1, R, G, NTR, NIR, B, TVDI, SWIR2, NSMI, Brightness, LST, K_s, MNDWI, EVI, VSDI, Sand, Silt, Clay, BD, θ_{FC}, θ_{PWP}, ϕ, θs, θr, n, α, T, W, RH, Rn, S_h, E	Wetness, Albedo, Greenness, SAVI, RVI, MSI, GVMI, NDWI, NMDI, NDVI, fNDVI, SWIR1, R, G, NTR, NIR, B, TVDI, SWIR2, NSMI, Brightness, LST, K_s	Wetness, G, NTR, B, TVDI, SWIR2, NSMI, Brightness, LST, K_s	Wetness, G, NTR, B, TVDI, SWIR2, NSMI, Brightness, LST

416

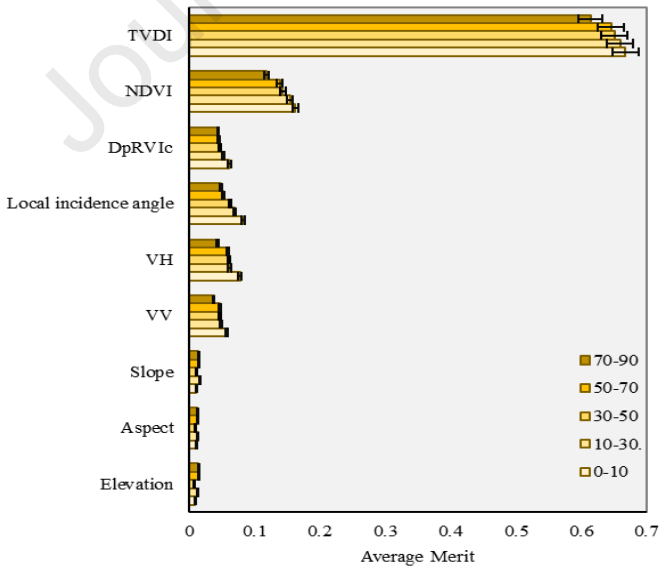


Fig. 6. Evaluating influential factors in predicting soil moisture at different depths using IGR in Scenario-2.

417

418 This was in agreement with the study of Wakigari and Leconte (2022), which found less importance of VV and VH
419 than NDVI in soil moisture retrieval in the Susquehanna watershed. In contrast, Bai et al. (2019) highlighted VV,
420 LST, and VI as the most robust predictors for estimating soil moisture in regions with sparse vegetation. Their study
421 demonstrates that the effectiveness of these variables is heavily influenced by local environmental conditions and
422 vegetation cover. This underscores the importance of adopting context-specific strategies when selecting input
423 parameters to ensure accurate soil moisture estimation. Since the topographical parameters were less effective in
424 estimating soil moisture, they were excluded from the estimation. On the other hand, estimating soil moisture from
425 SAR data remains challenging because of the effects of vegetation and surface roughness. In previous studies,
426 researchers used NDVI to reduce the impact of vegetation cover in soil moisture retrieval using derived data from
427 Sentinel-1 (Bhogapurapu et al., 2022; Liu et al., 2022; Wakigari and Leconte, 2022). The first model was defined
428 based on prior research, including VV, VH, local incidence angle, $DpRVI_C$, and NDVI. The second model was
429 developed using the first model's parameters and TVDI. The goal behind adding TVDI to the retrieval Models was
430 to introduce a new effective parameter that may reduce vegetation cover's impact due to the dominance of dense
431 vegetation in the study area, considering the use of surface temperature in addition to vegetation cover.

432

433 **3.3. Hyperparameters tuning**

434 Table 5 presents the results of the 5-fold-CV-based hyperparameter selection described in section 2.4.1. The optimal
435 values of the user-defined parameter M for the M5P algorithm at all depths and in both scenarios were determined
436 by the 5-fold-CV method as 4. For the SVR algorithm in the first Scenario, the kernel type was linear for estimating
437 surface and near-SSM, and the RBF was determined for deeper layers. In Scenario-2, the Poly and linear kernels at
438 0-70 cm and 70-90 cm depths, respectively, had the lowest RMSE values for soil moisture estimation. On the other
439 hand, the number of trees varied from 200 to 900 at different depths in both XGBoost and RFR algorithm Scenarios.
440 More details about setting hyperparameters in the ML algorithms are given in Table 5.

441 **Table 5.** Hyperparameters tuning in different ML methods

ML methods	Hyper-parameters	Scenario-1					Scenario-2				
		0-10	**10-30**	**30-50**	**50-70**	**70-90**	**0-10**	**10-30**	**30-50**	**50-70**	**70-90**
SVR	Kernel type	Linear	Linear	RBF	RBF	RBF	Poly	Poly	Poly	Poly	Linear
	C	3.145	1.310	7.515	9.486	8.099	1.212	1.873	2.142	1.020	1.672
	ɣ	-	-	0.695	0.494	0.627	0.680	0.067	0.584	0.256	-
	d	-	-	-	-	-	1.726	2.399	1.034	2.648	-
	ε	-	-	0.448	0.368	0.148	0.065	0.080	0.057	0.607	-
XGBoost	Colsample_bytree	0.798	0.799	0.798	0.899	0.688	0.599	0.798	0.799	0.599	0.799
	gamma	0.380	0.388	0.380	0.398	0.278	0.161	0.388	0.388	0.161	0.388
	Learning_rate	0.058	0.058	0.058	0.062	0.054	0.158	0.058	0.058	0.158	0.058
	Max_depth	7	7	7	9	7	3	7	7	3	7

	N_Estimators	246	240	246	250	240	688	246	246	688	246
	Max_delta_step	6	6	6	7	7	2	6	6	2	6
	Subsample	0.70	0.70	0.70	0.80	0.60	0.65	0.70	0.70	0.65	0.70
RFR	N_Estimators	900	900	900	100	100	300	200	800	700	200
	Max_Depth	50	90	80	40	10	60	70	110	80	10
	Min_Samples_Leaf	2	2	2	2	2	2	2	2	4	1
	Min_Samples_split	2	10	5	2	2	5	2	10	10	10

Note: ML: Machine learning

3.4. Models' performance

To assess the accuracy of the models, the estimated soil moisture for the surface, near-surface, and root zone predicted using the specified models in Scenario-1 was compared with the corresponding on-site measured values (Fig. 7). As illustrated in Fig. 7, the M5P algorithm yielded similar results for soil moisture estimation with Model-1 (R^2=0.90, RMSE=0.04, MAE=0.028 cm^3cm^{-3}) and Model-4 (R^2=0.88, RMSE=0.043, MAE=0.03 cm^3cm^{-3}). In other words, the exclusion of meteorological factors, soil physical and hydraulic properties, spectral indices (RVI, MSI, GVMI, NDWI, NDVI, fNDVI, SAVI, and SWI), and the spectral band (NIR) did not significantly impact the results. Furthermore, the M5P algorithm based on Model-1 provided reliable estimates for soil moisture at depths of 10-30 cm (R^2=0.84, RMSE=0.049, MAE=0.033 cm^3cm^{-3}), 30-50 cm (R^2=0.82, RMSE=0.053, MAE=0.035 cm^3cm^{-3}), 50-70 cm (R^2=0.80, RMSE=0.053, MAE=0.036 cm^3cm^{-3}), and 70-90 cm (R^2=0.80, RMSE=0.053, MAE=0.04 cm^3cm^{-3}). These findings indicate that incorporating all input parameters into the M5P algorithm significantly enhances the model's accuracy. Although the M5P model, which uses decision trees to partition data and linear regression models for prediction in each leaf, has a relatively simple structure, its decision-making power is more pronounced compared to many widely used models (such as ANN), effectively interacting with various and complex features (Yukseler et al., 2023).

The highest accuracy of the SVR algorithm in estimating SSM (R^2=0.79, RMSE=0.059, MAE=0.043 cm^3cm^{-3}) and near-SSM (R^2=0.75, RMSE=0.069, MAE=0.046 cm^3cm^{-3}) was observed in Model-1. The poor performance of Models-2 to 4 can be attributed to excluding some input parameters. According to Fig. 7, the best performance of SVR in estimating soil moisture at 30-50 cm depth (R^2= 0.74, RMSE=0.064, MAE= 0.049 cm^3cm^{-3}), 50-70 cm depth (R^2=0.73, RMSE=0.062, MAE=0.047 cm^3cm^{-3}) and 70-90 cm depth (R^2=0.74, RMSE=0.064, MAE=0.049 cm^3cm^{-3}) was observed using Model-4. Excluding meteorological variables, soil physical and hydraulic parameters, and some spectral indices from the input data (Model-4) in soil moisture estimation by the SVR algorithm led to increasing R^2 and decreasing RMSE and MAE trends for all depths in the root zone (30-90 cm).

According to Fig. 7, the highest accuracy of the XGBoost algorithm in estimating soil moisture at 0-10 cm depth (R^2=0.98, RMSE=0.020, MAE=0.015 cm^3cm^{-3}), 10-30 cm depth (R^2=0.97, RMSE=0.025, MAE=0.016 cm^3cm^{-3}) and 30-50 cm depth (R^2=0.93, RMSE=0.035, MAE=0.028 cm^3cm^{-3}) were observed in Model-3. Also, the omission of meteorological parameters, physical and hydraulic soil properties, and some spectral indices (e.g., 1, 2, and 3) in

Model-4 improved the XGBoost algorithm's soil moisture estimation accuracy at depths of 50-70 (R^2=0.91, RMSE=0.035, MAE=0.029 cm^3cm^{-3}) and 70-90 (R^2=0.89, RMSE=0.042, MAE=0.029 cm^3cm^{-3}).

As shown in Fig. 7, the best performance of RFR in estimating soil moisture at 0-10 cm depth (R^2=0.98, RMSE=0.018, MAE=0.011 cm^3cm^{-3}) and 10-30 cm depth (R^2=0.97, RMSE=0.024, MAE=0.015 cm^3cm^{-3}) was observed using Model-3 and Model-2, respectively. Also, the RFR method more accurately estimates soil moisture at 30-50 cm depth (R^2=0.93, RMSE=0.036, MAE=0.024 cm^3cm^{-3}), 50-70 cm depth (R^2= 0.93, RMSE=0.035, MAE=0.022 cm^3cm^{-3}) and 70-90cm depth (R^2=0.91, RMSE=0.039, MAE=0.025 cm^3cm^{-3}) based on Models 3, 4, and 4, respectively (Fig. 7). The effect of Ks was evident as soil depth increased from 0 to 50 cm when using the SVR, XGBoost and RFR algorithms to estimate soil moisture. However, by increasing the depth from (50-70 cm) to (70-90 cm), the highest value of R^2 and the lowest value of MAE and RMSE were obtained only by combining spectral indices and spectral bands (Model-4). Spectral information was the only parameter that affected the SVR, XGBoost, and RFR algorithms in estimating soil moisture at high depths.

Fig. 8 demonstrates the correlations between M5P, SVR, XGBoost, and RFR predicted and observed soil moisture at each studied depth employing the specified Models in the testing phase of Scenario-1. The mean values of R^2 (RMSE) between predicted and on-site measured SSM produced from the M5P, SVR, XGBoost, and RFR methods, respectively, were 0.83 (0.054 cm^3cm^{-3}), 0.74 (0.066 cm^3cm^{-3}), 0.87 (0.042 cm^3cm^{-3}) and 0.89 (0.041 cm^3cm^{-3}), based on the test data. The most accurate soil moisture prediction at 0-10 cm depth is observed by RFR using Model-3 (R^2=0.89, RMSE= 0.04, MAE=0.027 cm^3cm^{-3}), and the lowest could be seen at 70-90 cm depth by SVR using Model-4 (R^2=0.51, RMSE=0.085, MAE=0.062 cm^3cm^{-3}). RFR's soil moisture prediction accuracy was significantly higher than M5P and SVR (Fig. 8). Furthermore, the proximity of the data to the line of the perfect agreement indicates the accuracy with which the model estimates the measured data.

Nevertheless, considerable scattering of data points from the agreement line demonstrates poor performance of M5P and SVR algorithms in modeling measured soil moisture data, thus lacking generalization. The main difference between the M5P, SVR, and RFR algorithms is the large number of input parameters and the consequent inability to detect soil moisture heterogeneity due to the variability of spectral data and physical and hydraulic soil attributes.

On the other hand, based on Fig. 7 and 8, XGBoost produced nearly similar results to RFR in RZSM estimation. Notably, RFR has fewer tunable hyperparameters than XGBoost, which reduces its complexity and makes it easier to use. Also, RFR outperformed XGBoost in the estimation during the testing phase based on the metrics (R^2, RMSE, and MAE) (Fig. 8). However, the XGBoost algorithm's results were also satisfactory. Araya et al. (2021) confirmed that RS data and soil characteristics could affect soil moisture estimation using the RFR algorithm. Wang et al. (2022) evaluated the performance of five ML models (polynomial regression, ridge regression, lasso regression, elastic net regression, and RFR) in SSM estimation using RS data and various parameters, such as geographical location, elevation, vegetation coverage, soil texture, and seasonal patterns. The results demonstrated that the RFR model with an MAE less than 0.02 cm^3cm^{-3} performed better than other ML models.

The permutation feature importance technique was employed to evaluate the significance of different input variables in predicting soil moisture across various depths using the RFR algorithm. Fig. 9 illustrates the importance of input variables in the top models. Despite variations in the importance rankings at different depths, TVDI consistently emerged as the most significant factor across all examined depths. This consistency underscores TVDI's critical role in accurately reflecting soil moisture depletion, primarily due to its effect on plant stomatal regulation (Krishnan and Indu, 2023). Additionally, during more advanced stages of water stress, where moisture levels near the root zone significantly decrease, a notable reduction in VIs can be observed (Holzman et al., 2014). These results align with Ryu et al.'s (2021) findings, which demonstrate TVDI's capability to effectively monitor both short-term and long-term variations in soil moisture.

The Brightness and NSMI variables were identified as the second most important factors at depths of 0-30 cm and 30-90 cm, respectively (Fig. 9). The Brightness index plays a more crucial role in SSM estimation due to its sensitivity to surface conditions, such as soil texture, organic matter content, and surface wetness. This index influences the soil's radiation budget and energy balance, affecting surface evaporation and transpiration processes, ultimately leading to changes in soil moisture levels (Firozjaei, 2019). On the other hand, the NSMI, with its ability to capture variations in soil moisture content, is particularly effective at greater depths where direct surface influence is diminished.

Previous research has highlighted significant differences in soil reflectance between SWIR2 and SWIR1 bands. These differences are primarily attributed to water absorption characteristics, which correlate linearly with SSM levels (Sadeghi et al., 2015; Yue et al., 2019). Adab et al. (2020) investigated the relationship between SSM and optical-thermal data in this context. Their findings revealed that SWIR1 and SWIR2 bands strongly correlate with SSM in barren and vegetated areas, respectively. This correlation indicates that SWIR bands are effective indicators for differentiating SMC across various land cover types. Furthermore, Sadeghi et al. (2017) showed that SWIR reflectance is sensitive to the water content within leaves and the internal leaf structure. As a result, SWIR bands can effectively reflect changes in soil moisture within the root zone, making them valuable for monitoring vegetation health and soil moisture dynamics.

Besides spectral characteristics, the RFR algorithm also identified physical and hydraulic soil properties such as θ_{FC} and Ks as highly significant, especially at depths of 0-10 cm and 0-50 cm, respectively. The θ_{FC} variable plays a crucial role in determining SSM levels by indicating the maximum water retention capacity of the soil in its surface layers. This capacity directly impacts moisture conditions around plant roots, influencing irrigation management practices. Meanwhile, Ks is vital in regulating water flow within subsurface soil layers, thereby controlling moisture's spatial and temporal distribution at various depths. This highlights Ks's significance in water resource management and optimizing water use, particularly at the farm scale. Similar findings by Moazenzadeh et al. (2022) suggest that soil moisture variability is a function of matric potential and hydraulic properties, differing with soil depth and time, especially in unsaturated environments.

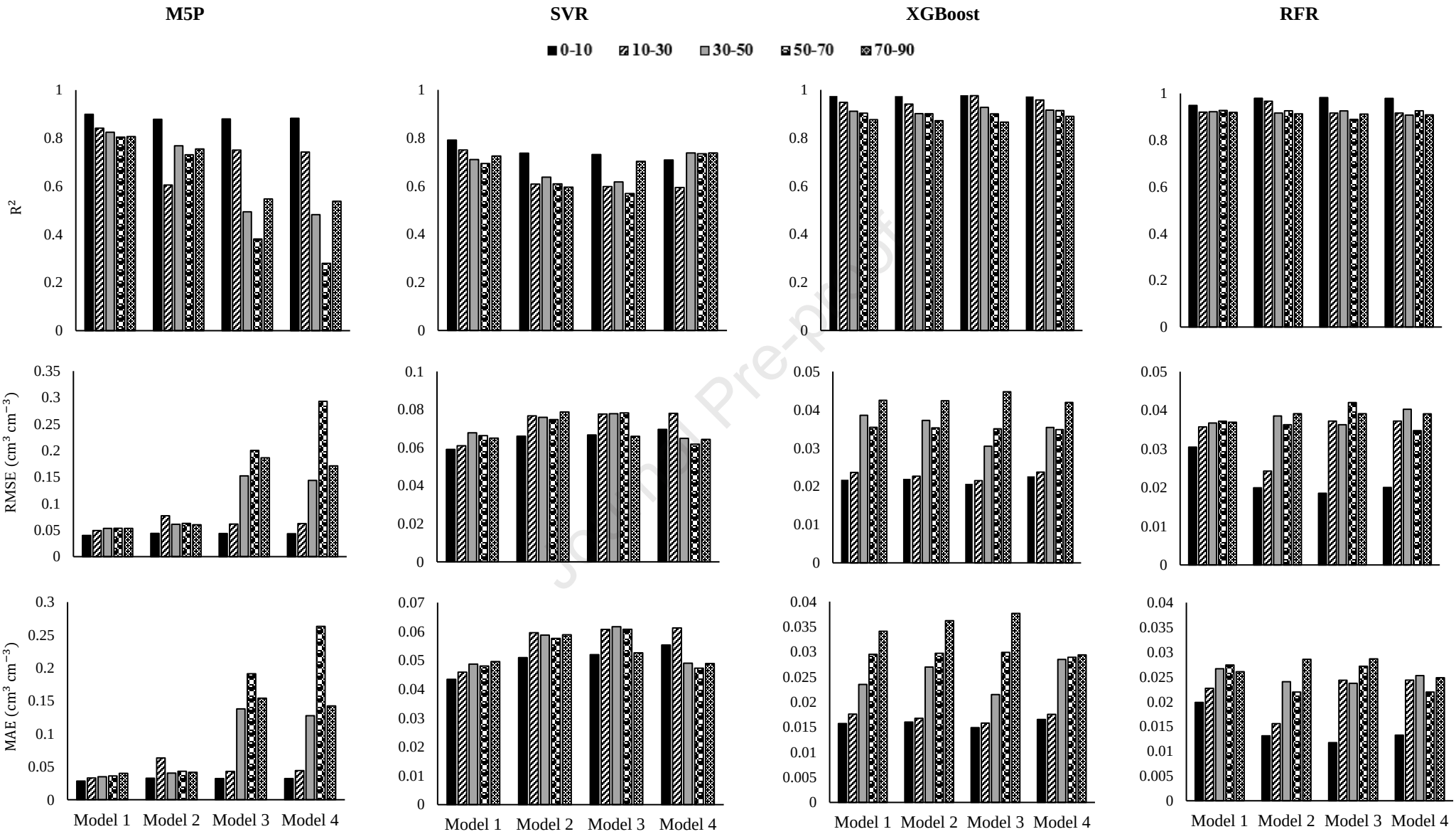
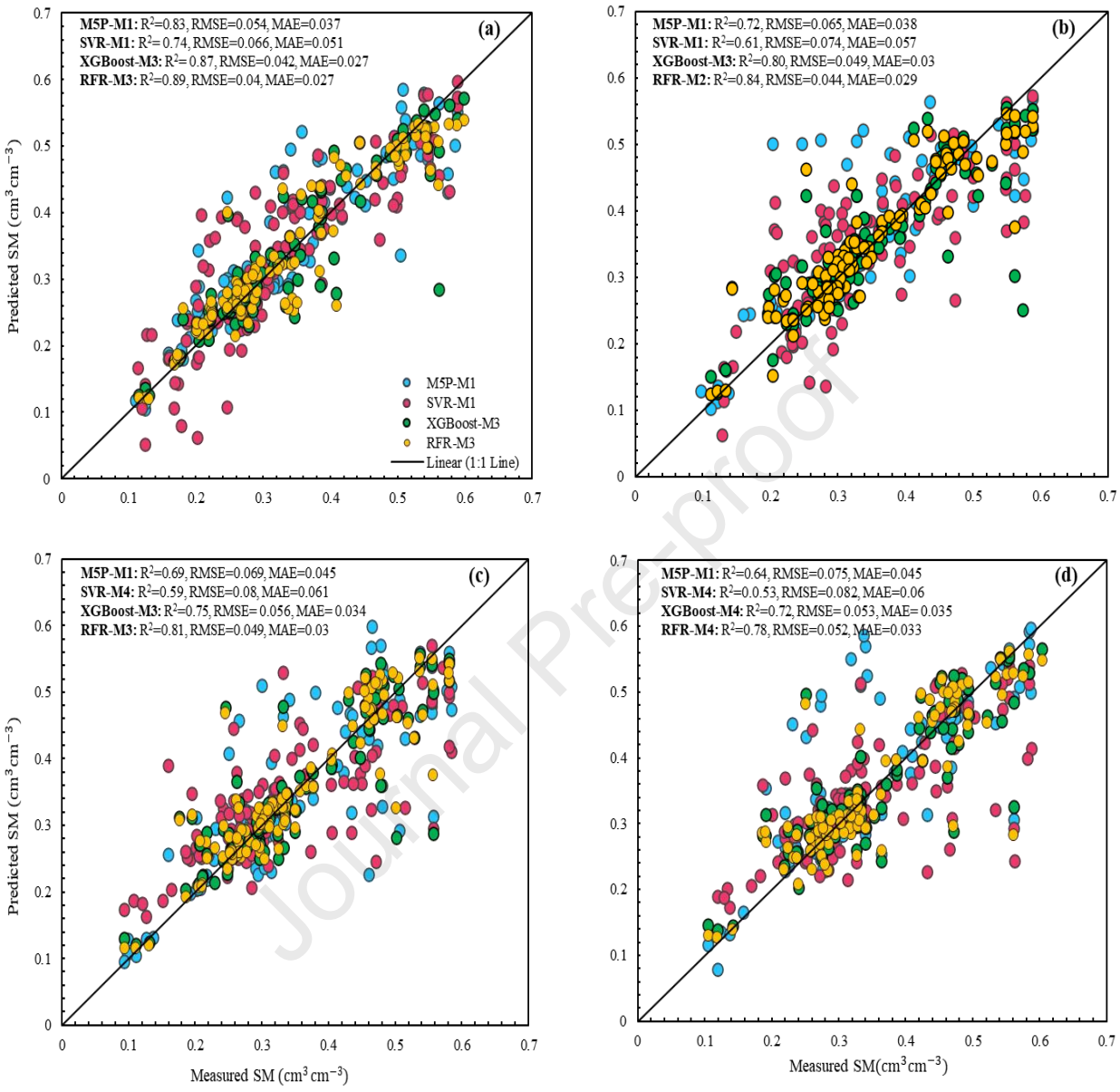


Fig. 7. Accuracy metrics between estimated and measured soil moisture with the M5P, SVR, XGBoost, and RFR algorithms in the training period in Scenario-1.

538



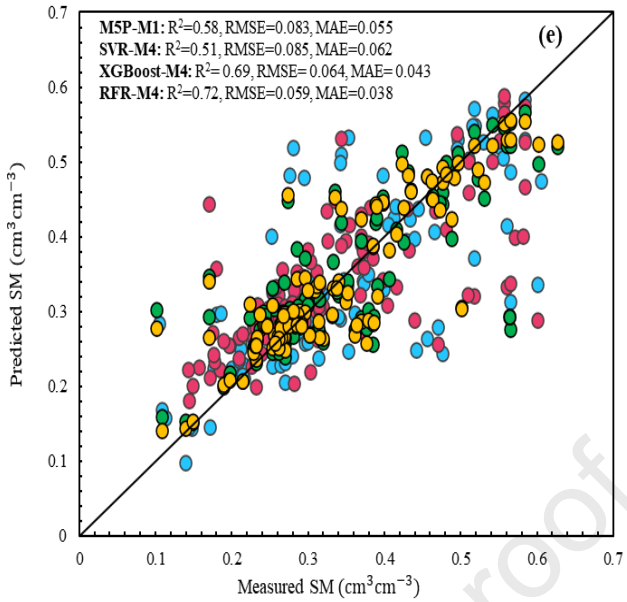
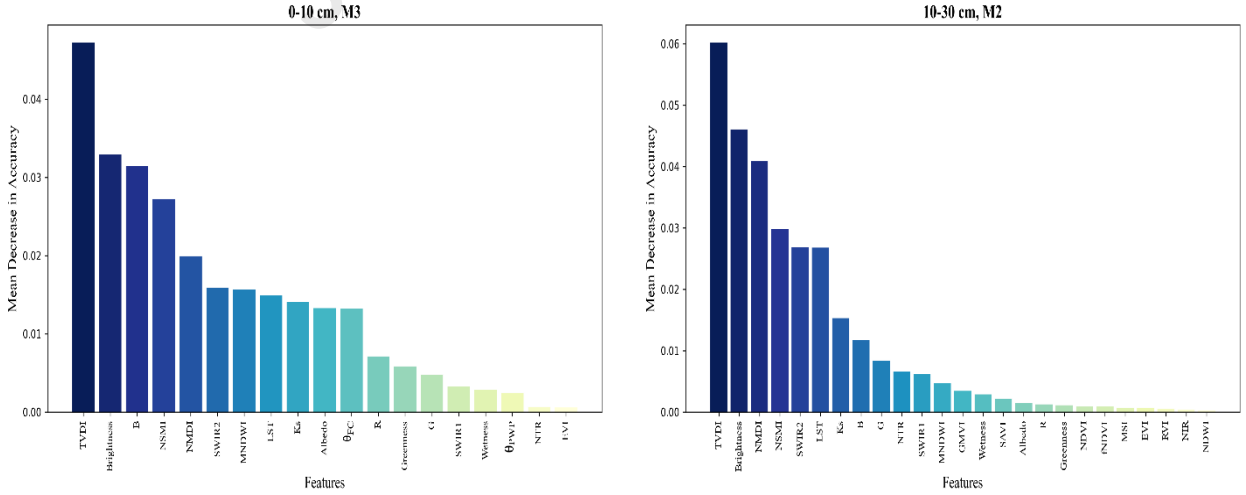


Fig. 8. Comparison of measured and predicted soil moisture by M5P, SVR, XGBoost, and RFR algorithms in Scenario-1 during the testing period at different depths; (a) 0-10 cm, (b) 10-30 cm, (c) 30-50 cm, (d) 50-70 cm and (e) 70-90 cm. The solid line is the 45-bisector line.

Although meteorological variables did not significantly influence the selected models, this observation does not negate their potential importance. The limited impact of these variables might be due to the regional scale of the study and the relatively homogenous climatic conditions observed. However, meteorological variables will likely play a more significant role in soil moisture prediction in more humid regions or areas with substantial and variable precipitation patterns. This warrants further investigation under diverse climatic conditions to fully understand their impact.



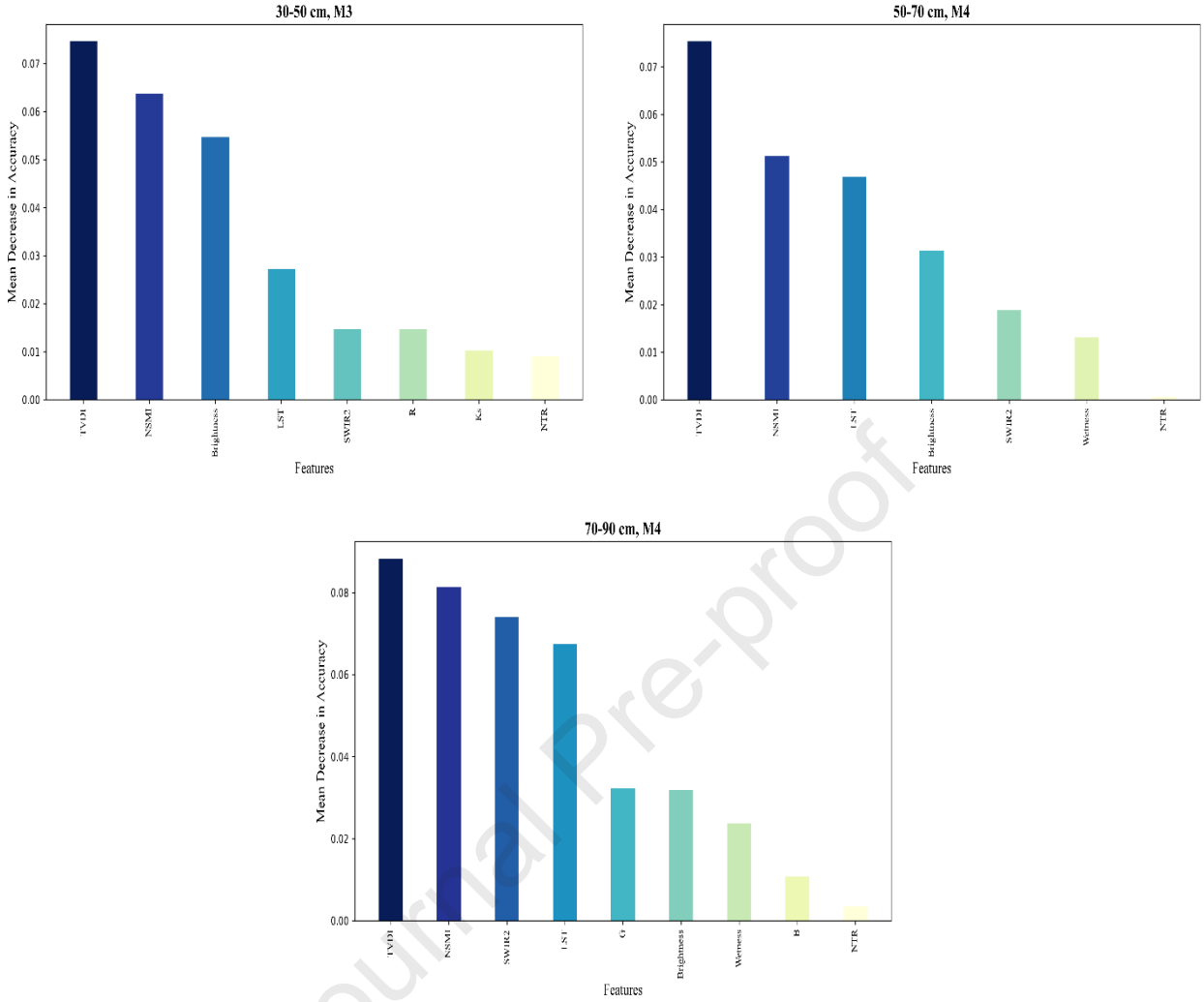


Fig 9. RF model variable importance in Scenario-1, showing only the variables with the highest predictive performance at each depth.

In Scenario-2, soil moisture was retrieved from combined Sentinel-1 SAR and Landsat-8 VIs using the RZSM estimation models. Comparing the ML algorithms results in Models 1 and 2, the accuracy of the soil moisture estimation from Model-2 was significantly higher than that from Model-1. For example, Adding TVDI to Model-2 led to an increase in the R^2 value and a decrease in MAE and RMSE values compared to Model-1 in all soil depths and ML algorithms (Model-1: R^2_{M5P}=0.40, R^2_{SVR}=0.49, $R^2_{XGBoost}$=0.95, R^2_{RFR}=0.74; Model-2: R^2_{M5P} =0.50, R^2_{SVR} =0.65, $R^2_{XGBoost}$= 0.97, R^2_{RFR} =0.97) (Fig. 10).

These results can be attributed to the fact that the TVDI is commonly more sensitive to soil moisture (Ryu et al., 2021). Furthermore, they can reduce the impact of vegetation cover in estimating SSM from SAR data more efficiently. A comparison between the results from the training data set in Model-2 (Fig. 10) shows that SMC could be predicted with high accuracy for RFR and XGBoost (R^2=0.97, RMSE=0.024, MAE=0.018 cm^3cm^{-3}), while SVR (R^2=0.65, RMSE=0.076, MAE=0.06 cm^3cm^{-3}) and M5P (R^2=0.50, RMSE=0.092, MAE=0.07 cm^3cm^{-3}) had the

lowest accuracy of estimation soil moisture at 0-10 cm depth. Moreover, based on the statistical criteria, the RFR and XGBoost algorithms were superior to the other two algorithms (M5P and SVR) at 10-90 cm depths during the training stage. As the depth increased, the RMSE value in RFR decreased by 22% on average compared to XGBoost. Consequently, the RFR algorithm was deemed the optimal ML method for estimating SMC in Scenario-2.

Fig. 11 depicts the estimated versus measured SMC scatter plots from four specified ML algorithms using Scenario-2's optimal features during the testing phase. As shown in Fig. 11, the M5P and SVR estimations at all depths were overestimated and underestimated for the low and high soil moisture ranges. The M5P algorithm did not perform satisfactorily despite using the pruning method. In contrast, the SVR algorithm had more effectively estimated soil moisture using Sentinel-1 SAR data than the M5P algorithm. By comparing Fig. 8 and 11, it is evident that both Scenarios' estimates of the SSM produced similar results. However, as the depth increased, the R^2 values in Scenario-1 increased by 9.52%, 9.87%, 11.53%, and 11.11% at 10-30 cm, 30-50 cm, 50-70cm, and 70-90 depths, respectively, compared to Scenario-2.

Moreover, the RFR algorithm performed a more accurate prediction with IGR-optimized parameters derived from two multiple sensors than the other ML algorithms. Adding TVDI to the proposed RFR algorithm increased the R^2 value in 0-10 cm depth (R^2 = 0.89). In contrast, Liu et al. (2022) did not attain a better R^2 value than 0.72 in estimating SSM using the ANN prediction model with VV, VH, and NDVI input data in Australia. This indicates that the superiority of the proposed approach may be attributed to the combined use of TVDI in the RFR model. Also, the R^2 values decreased from 14% to 28% for depths from 10-30 cm to 70-90 cm compared to this value at 0–10 cm depth.

The ML model's important outcome was to provide a detailed soil moisture map. Therefore, the optimal SMC prediction model (RFR) in both Scenarios was operated on each pixel in the area of research to produce a spatial distribution of RZSM at various times. The differences between the multiple soil depths could be seen on the modeled soil moisture maps (Fig. 12). Fig. 12 depicts RZSM distribution on September 10, the peak growth period for sugarcane crops, when fields were irrigated every 5-7 days. Some farms were under-irrigated or irrigated, while others had soil moisture less than θ_{FC} and required irrigation (Fig. 12). Less soil moisture is shown on 0-30 cm depth maps than on 30-70 cm depth maps. The reason could be the higher topsoil layer evaporation (Fig. 12a and 12b). By our expectations, the percentage of soil moisture close to saturation is lower at 70-90 cm depth. This is due to the location of the drains that control soil moisture at this depth.

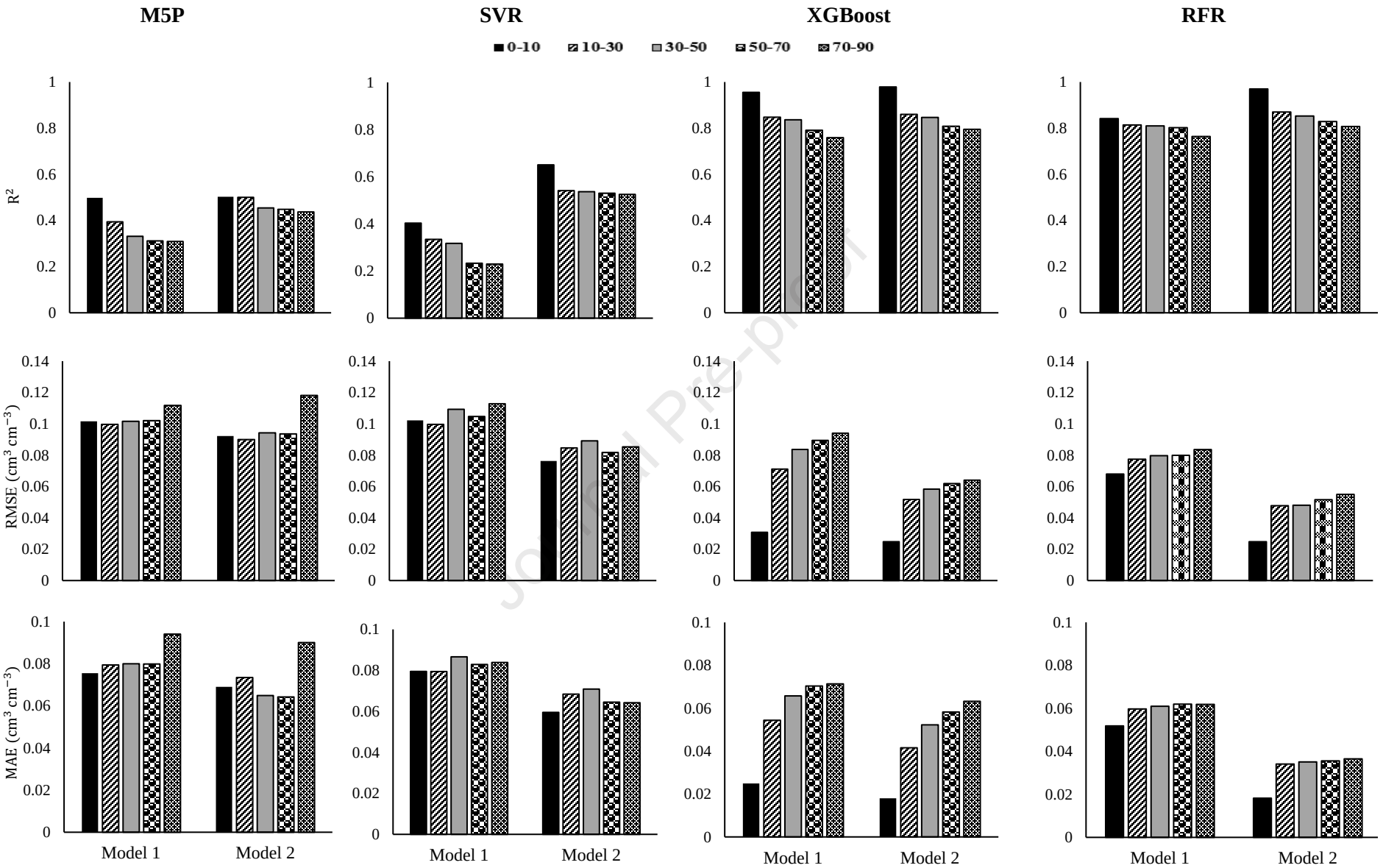
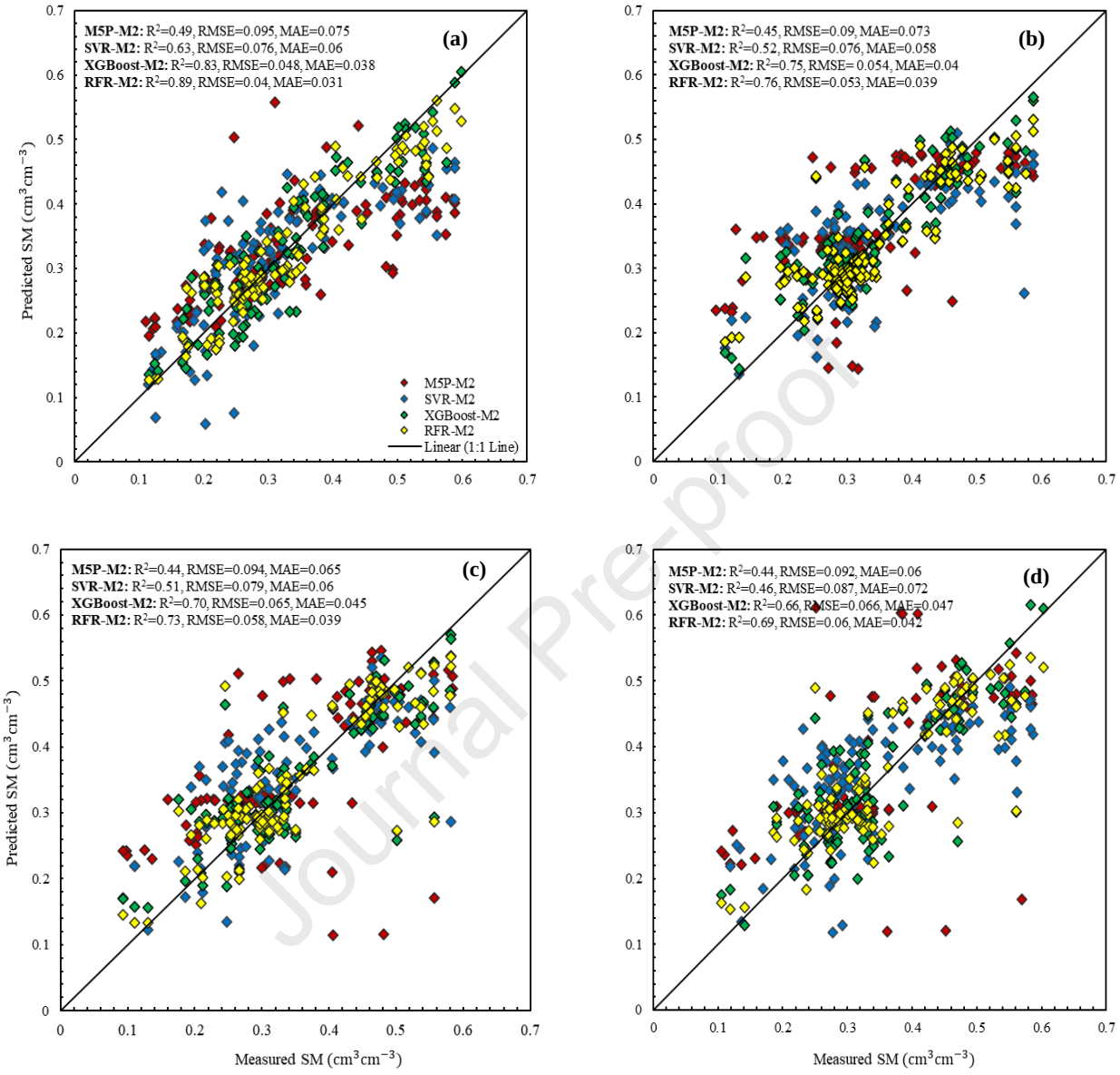


Fig. 10. Accuracy metrics between estimated and measured soil moisture with the M5P, SVR, XGBoost, and RFR algorithms in the training period in Scenario-2.

586



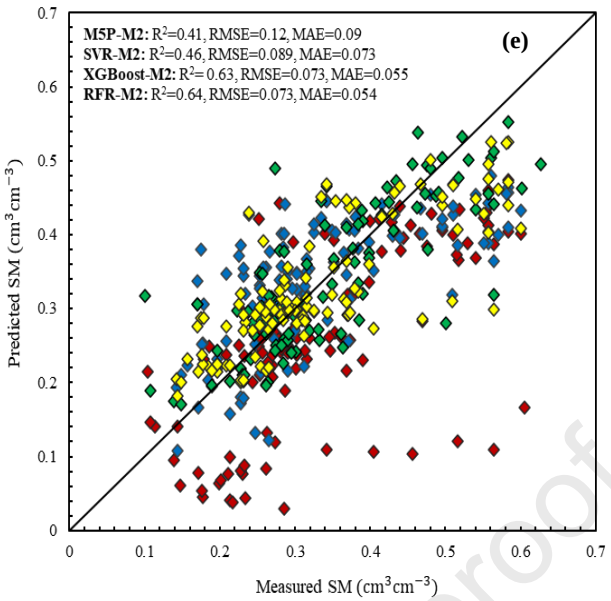


Fig. 11. Comparison of measured and predicted soil moisture by M5P, SVR, XGBoost, and RFR algorithms in Scenario-2 during the testing period at different depths; (a) 0-10 cm, (b) 10-30 cm, (c) 30-50 cm, (d) 50-70 cm and (e) 70-90 cm. The solid line is the 45-bisector line.

3.5. Adding measured surface soil moisture in modeling RZSM

Due to existing hydraulic connections between the top layer (0-10 cm) and the root zone (10-90 cm), RZSM estimations can be substantially more accurate based on actual surface layer soil moisture data. The RFR algorithm was used to test this supposition due to its superior performance in both Scenarios (Section 3.4). The median value is represented by the straight line that runs through each box. The whiskers stretch from the 5th to 95th percentile values, while the boxes indicate the 25th and 75th percentiles (interquartile area). The results indicated that adding the SSM improves the RZSM estimates in both Scenarios, and the best performance of the training phase was at 10-30 cm depth (Scenario-1: R^2=0.98, RMSE=0.016, MAE=0.009 cm^3cm^{-3}; Scenario-2: R^2=0.98, RMSE=0.018, MAE=0.011 cm^3cm^{-3}) (Fig.13a and 13d). A surprising result was that the addition of SSM to Scenario-2 (combination of Sentinel-1 extracted VV, VH, local incidence angle, and $DpRVI_C$ and Landsat-8 extracted NDVI and TVDI) improved the R^2 values in 10-30 cm, 30-50 cm, 50-70 cm, and 70-90 cm depths by 17.39%, 19.76%, 21.59%, and 22.89%, respectively (Fig. 13d-13e).

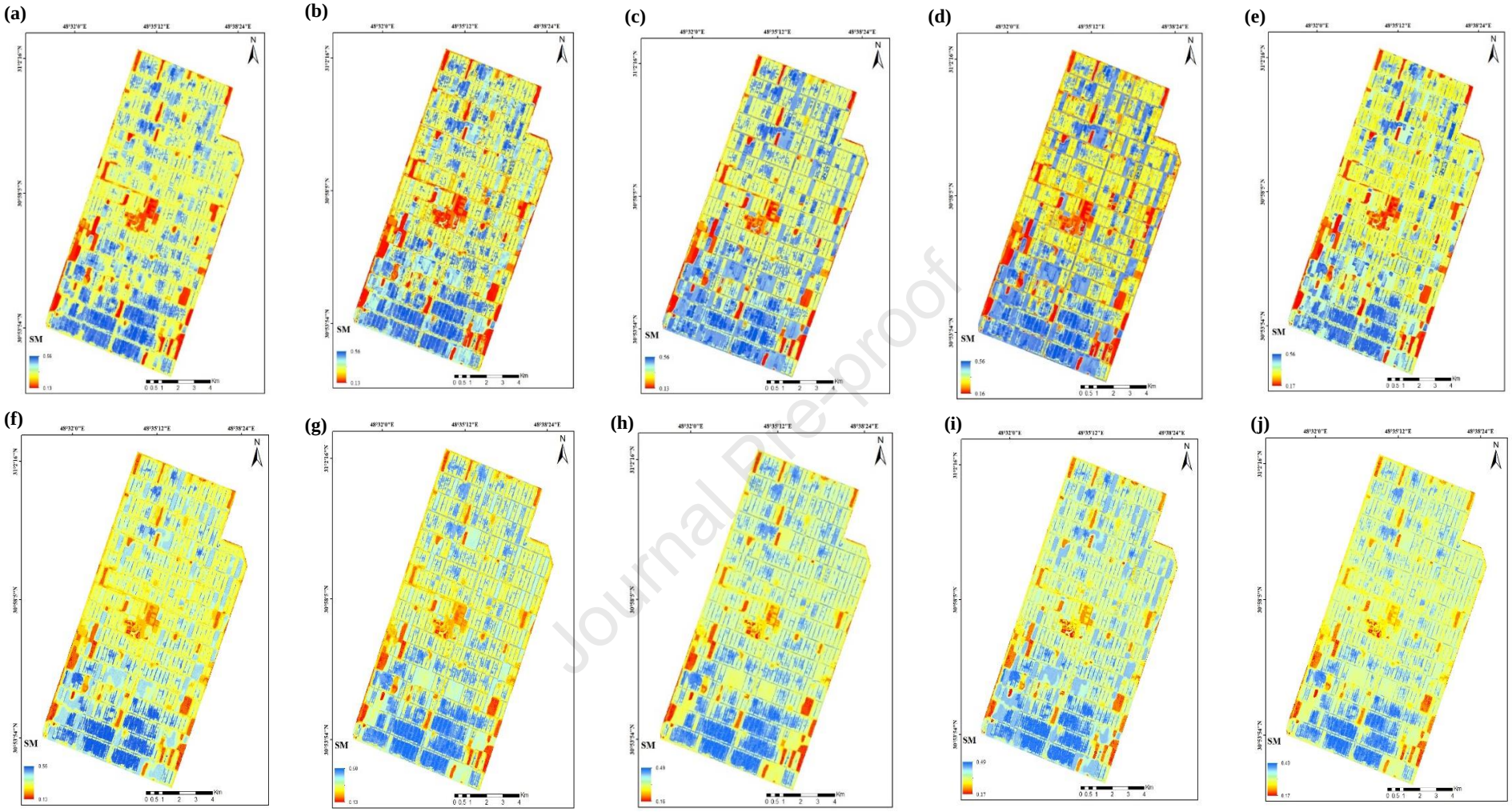


Fig. 12. Spatial pattern of estimated soil moisture with RFR algorithm for both Scenarios on September 10, 2020. a to e (Scenario-1) and f to j (Scenario-2) indicate 0-10 cm, 30-10 cm, 30-50 cm, 50-70 cm, and 70-90 cm depths, respectively.

600

601

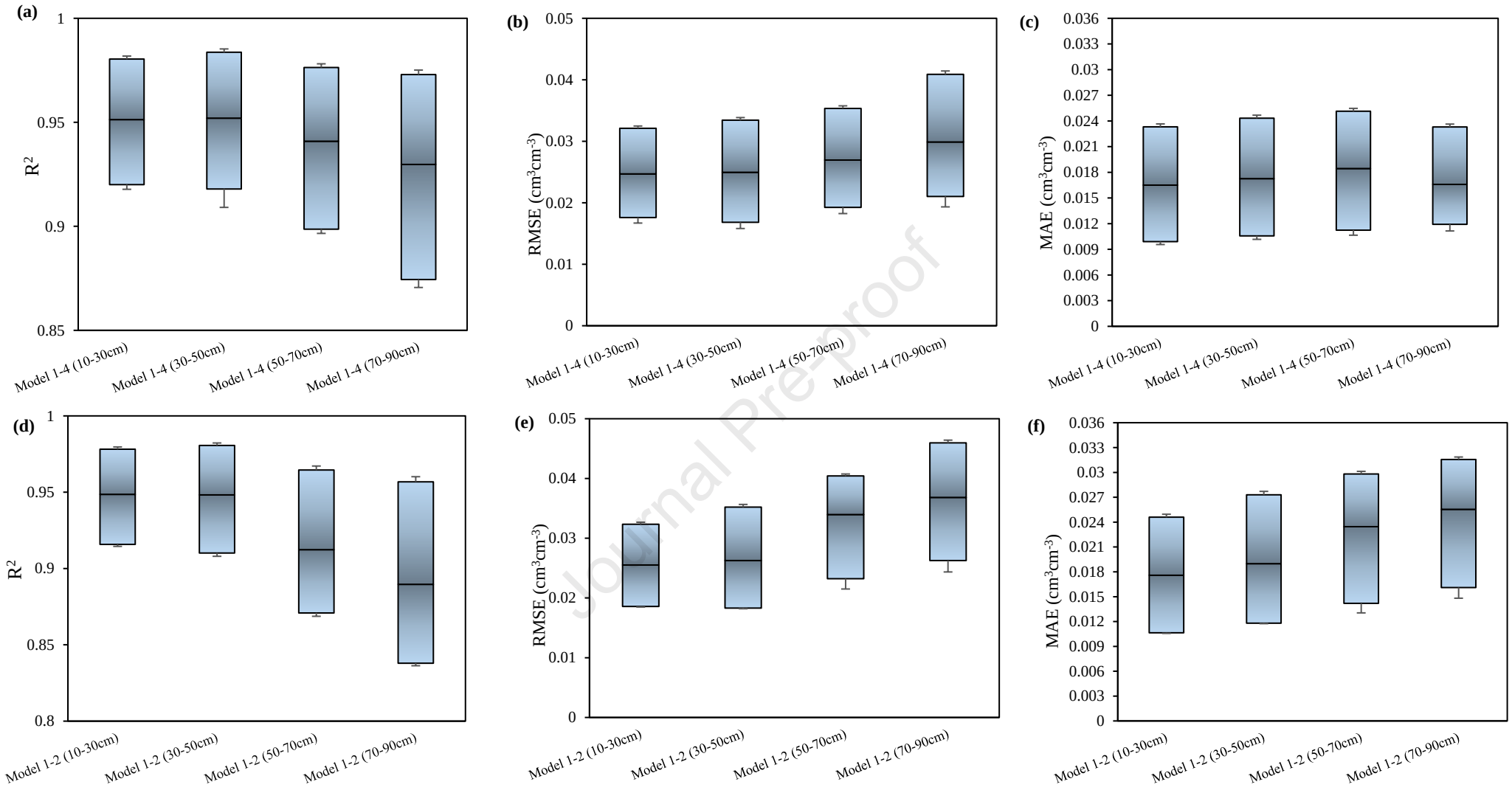


Fig. 13. The highest accuracy of retrieved soil moisture with the RFR method by adding surface soil moisture (0-10 cm) to the datasets of Models 1 to 4 in Scenario-1 (a to c) and Models 1 to 2 in Scenario-2 (d to f) at all depths.

602

This improved performance is because evaporative factors mainly control soil moisture in the thin surface layer. On the other hand, the hydraulic connection of this layer with the lower soil layers may be lost due to extreme evaporation conditions. Comparing the results of the RFR algorithm by adding the SSM in the first and second Scenarios showed that the soil moisture estimation at 0-50 cm depth in both Scenarios performed almost similarly, and no significant difference was observed in the results. However, Scenario-1 outperformed Scenario-2 as the depth was increased. RZSM maps incorporating RFR based on SSM can better recognize tracks created by agricultural vehicles and large irrigation canals with an appropriate spatial resolution (Fig. 14).

Overall, the proposed approach indicated the impressive precision of ML algorithms without relying on system dynamics assumptions. Additionally, the method can provide a comprehensive assessment of the probability distribution of the spatial map for RZSM using optical, thermal, and SAR RS data. This research selected the input variables based on the IGR method. Meanwhile, Chen et al. (2021) used three feature selection methods, including Pearson's correlation coefficient, recursive feature elimination, and RF, to enhance the accuracy of soil moisture estimation by ML models such as SVR, RF, and GBRT. According to their findings, the RF model with the RF feature selection method was the most effective machine-learning model in estimating soil moisture (R^2=0.79). However, this level of accuracy is somewhat lower than the current research's performance accuracy of 0.89, demonstrating that the IGR method had effectively identified the influencing variables in soil moisture estimation.

In addition to spectral data, the RFR algorithms used θ_{FC}, θ_{PWP}, and Ks parameters to estimate RZSM in the optical-thermal (Scenario-1) RS. Therefore, it can be expected that the soil texture, characterized as medium to heavy within the study area, could significantly impact the estimates. Several researchers have emphasized that a very sandy or clayey texture is the most crucial factor affecting the uncertainty of the model (Tunçay et al., 2023). Also, Gu et al. (2021) have demonstrated that soil thermal characteristics can lead to overestimating soil water content values in clayey soils. Therefore, lighter-textured soils may affect the values of the spectral indices due to their lower water-holding capacity, leading to different results.

Tao et al. (2023) used a multi-feature ensemble learning algorithm to retrieve soil moisture during the principal phases of grape growth using spectral indices, topography, and evapotranspiration, which had a lower performance accuracy (R^2=0.75) than RFR performance in this study (R^2=0.89). Also, their investigation employed satellite imagery data with a lower spatial resolution (500 m) compared to the present study. They noted that topography and ET are the two most critical environmental factors that affect soil moisture retrieval in mountainous regions. On the other hand, Araya et al. (2021) underscored the significance of precipitation as the primary input variable for estimating soil moisture in Mediterranean areas. It should be noted that no effective rainfall was recorded during the research period in the warm and arid study area. Hence, it is imperative to emphasize that several environmental parameters, including topography, ET, and precipitation, can affect the accuracy of soil moisture estimation under diverse environmental conditions (Araya et al., 2021; Fathololoumi et al., 2020; Fatichi et al., 2015; Tao et al., 2023). Consequently, it is crucial to incorporate these aspects considering the conditions into the modeling process to enhance the accuracy of the models.

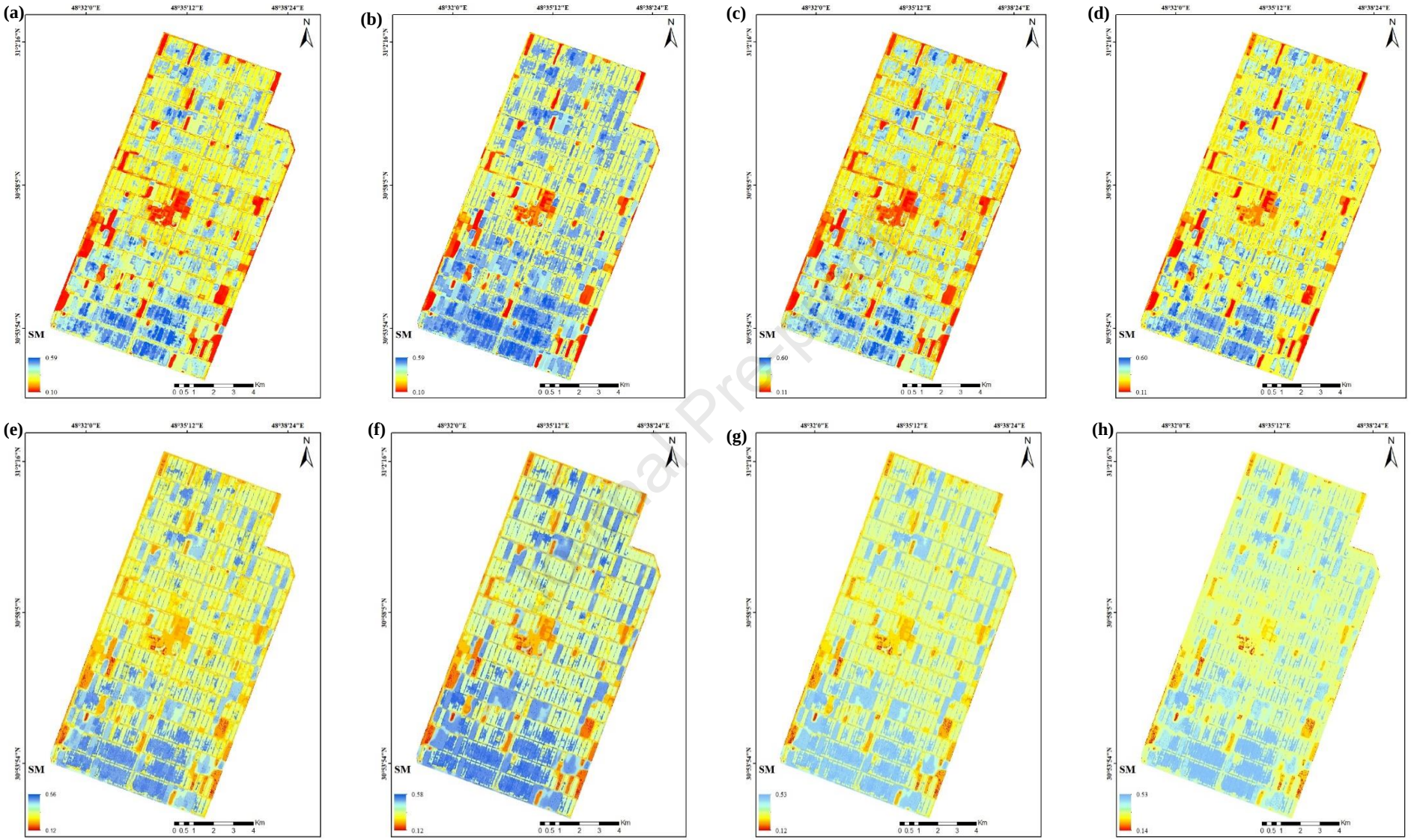


Fig. 14. Spatial pattern of estimated RZSM using surface soil moisture measurement as an auxiliary input variable using RFR algorithm for both Scenarios. a to d (Scenario-1) and e to h (Scenario-2) indicate 10-30 cm, 30-50 cm, 50-70 cm, and 70-90 cm depths, respectively.

Studies have shown that modeling soil moisture using SAR data in combination with the NDVI yields satisfactory results (R^2 = 0.70) owing to the strong interactions and correlations between this index and soil moisture (Liu et al., 2022; Tao et al., 2023). Also, compared to other spectral indices, it has been noted that LST and features related to LST, such as TVDI, play a more significant role in the soil moisture retrieval models (Liu et al., 2022; Tao et al., 2023). This implies that LST-related features are more significant indicators in soil moisture modeling than spectral indices. Therefore, this research hypothesized that combining NDVI and TVDI with SAR data could improve the accuracy of the RZSM retrieval model performance, and the findings confirmed this theory (R^2=0.89).

The results showed that in addition to TVDI, indices such as NSMI, Brightness, and the LST and SWIR2 bands play a crucial role in estimating RZSM due to their fundamental influence on plant stomatal regulation. These optical-thermal data are sensitive to water content in both vegetation and soil, making them highly effective in environments with dense vegetation, where the canopy significantly influences the signal. Combined with the RFR model, these optical-thermal indices accurately estimated RZSM (R^2=0.72-0.84). In contrast, combining TVDI with SAR data resulted in lower accuracy (R^2=0.64-0.76) than the optical-thermal approach, primarily due to the different nature of SAR signals, which penetrate the vegetation canopy and interact directly with the soil surface. Overall, these findings emphasize the potential of SWIR2-based indices in estimating RZSM and highlight the limitations of SAR data in densely vegetated areas. For instance, Krishnan and Indu (2023) used TVDI to estimate RZSM (0-200 cm), but their accuracy was lower compared to the present study (R^2=0.16-0.64). This comparison further underscores the importance of combining vegetation and moisture indices with TVDI in ML models to improve estimation accuracy.

According to the results, the ML models' RZSM retrieval accuracy decreased significantly compared to the SSM estimates. On the other hand, research has previously demonstrated that various indices and models have been developed to overcome the vertical limitations of microwave satellite soil moisture products and accurately estimate RZSM. These approaches involve establishing a connection between the time series of SSM and RZSM (Baldwin et al., 2019). Manfreda et al. (2014) developed a soil moisture analytical relationship (SMAR) model using the SSM series to determine RZSM. However, most of their applications are specifically tailored for satellite-based input datasets, often with a coarser spatial resolution (Baldwin et al., 2019). Therefore, including SSM as an additional input parameter along with spectral data and soil physical and hydraulic properties in ML models can help in modeling intricate and uncertain data, ultimately enhancing the accuracy of RZSM retrieval. Guo et al. (2023) employed the SMAR model to retrieve RZSM in the Xiliaohe River Basin. A genetic algorithm calibrated the model parameters. Then, the spatial parameters were estimated using the RFR method with the soil properties, meteorological variables, and vegetation characteristics as explanatory variables. The findings indicated that their performance accuracy (RMSE=0.06 cm^3cm^{-3}) was lower than the RFR's performance in the current study (RMSE=0.03 cm^3cm^{-3}).

The findings of this study have the potential to significantly impact decision-making in precision agriculture, especially in environments with limited resources such as water and fertilizers. The application of ML models like RFR, which demonstrated high accuracy in estimating RZSM using SSM (measured or satellite-derived), allows for precise soil moisture monitoring at various depths on a farm-scale basis. This level of precision enables farmers to

make data-driven irrigation decisions, ensuring that water is used efficiently and crops receive the optimal amount of moisture necessary for growth. Such an approach not only optimizes resource use and reduces input costs but also enhances the sustainability and accessibility of precision agriculture in resource-constrained environments.

3.6. Advantages and Limitations

Most soil moisture retrieval studies using Sentinel-1 SAR and Landsat-8 have focused on areas with low vegetation cover, and few have specifically addressed SMC in the root zone under dense vegetation cover. This study presents a practical approach for estimating RZSM in agricultural lands with high spatial resolution, particularly in areas with dense vegetation. The findings indicate that the implemented approaches are not only reliant on RS data but also consider the physical and hydraulic properties of the soil, enabling adaptation of these methods under similar conditions. Also, these results can serve as an effective tool for correcting gaps in long-term time-series data. In the SAR RS approach, TVDI was combined with SAR data to further enhance the accuracy of RZSM estimation. This novel approach provided valuable insights into RZSM estimation in densely vegetated areas and represents a novel step toward improving the accuracy and downscaling of microwave-based soil moisture products for operational use in agricultural areas.

However, the proposed approaches have inherent limitations. One significant limitation is the need for concurrent access to SAR and optical-thermal data. While Sentinel-1 SAR data are resilient to environmental factors, the optical and thermal bands used from Landsat-8 are highly sensitive to atmospheric conditions, such as cloud cover and aerosols. This sensitivity can limit the applicability of this method in regions with frequent adverse weather conditions, affecting the accuracy and consistency of the soil moisture estimates. Future studies may address this by incorporating indices like OPTRAM, which uses red, near-infrared, and shortwave infrared bands, bypassing the need for thermal data (Sadeghi et al., 2017). This would enable Sentinel-1 to be combined with sensors like Sentinel-2, which lack thermal bands with a higher temporal resolution than Landsat-8. If the application of OPTRAM is validated, these features could extend the method's applicability across a broader range of climatic conditions and enhance its effectiveness for operational agricultural monitoring.

Another limitation is related to the study area's specific environmental conditions, including its climate, soil type, and vegetation characteristics. This might restrict the findings' generalizability to other geographic regions with different climates, soil compositions, and vegetation types. Although meteorological variables were found to be less influential in RZSM estimation in this study, their role may be more pronounced in regions with different climate regimes or in environments subject to significant temporal variability in weather patterns. Hence, it is essential to carry out additional validation and calibration efforts in various regions and vegetation types.

4. Conclusions

This study presented a practical approach for estimating root zone soil moisture (RZSM) in cropland with dense vegetation cover at high spatial resolution. Machine learning (ML) techniques effectively identified RZSM in sugarcane fields by integrating in situ measured data (meteorological data and soil physical and hydraulic properties) with satellite information (Landsat-8 and Sentinel-1 images), leading to significant cost and time savings. The key findings of this study are summarized as follows:

- The RFR algorithm outperformed other algorithms (M5P, SVR, and XGBoost) in both the optical-thermal (Scenario-1) and the combined SAR and optical-thermal (Scenario-2) remote sensing (RS) approaches.
- The RFR algorithm performed better in estimating RZSM using the optical-thermal approach than the combined SAR and optical-thermal approach.
- Spectral parameters such as TVDI, Brightness, and NSMI significantly impacted RZSM estimation more than meteorological data and soil physical and hydraulic properties.
- In the combined SAR and optical-thermal approach, TVDI and NDVI were more critical for estimating soil moisture than other parameters, such as incidence angle and VH polarization.
- Including SSM as auxiliary input data in both Scenario-1 and Scenario-2 significantly increased the accuracy of RZSM estimation.

Despite some limitations related to weather conditions and the effects of vegetation cover, this study offers a novel and operational approach for estimating RZSM at the field scale. This methodology, by leveraging freely accessible satellite imagery, has the potential to significantly contribute to agricultural water resource management and the advancement of precision agriculture practices.

Funding

"Shahid Chamran University of Ahvaz" funded this study.

CRediT authorship contribution statement

Atefeh Nouraki: Conceptualization, Methodology, Data curation, Validation, Writing-original draft. **Mona Golabi:** Writing-review & editing. **Mohammad Albaji:** Writing-review & editing. **Abd Ali Naseri:** Writing-review & editing. **Saeid Homayouni:** Methodology, Writing-review & editing.

Declaration of interests

The authors declare that there are no competing interests.

Data Availability Statement

Data will be made available on request.

Acknowledgment

The authors thank the Research Council of the Shahid Chamran University of Ahvaz for financial support (GN: SCU.WI1401.281). We would also like to acknowledge the Research Center of Iran's Sugarcane and the HFSAIC Manager for providing the current research equipment.

References

Acharya, U., Daigh, A. L., Oduor, P. G., 2021. Machine Learning for Predicting Field Soil Moisture Using Soil, Crop, and Nearby Weather Station Data in the Red River Valley of the North. Soil Syst.5, 57.

Adab, H., Morbidelli, R., Saltalippi, C., Moradian, M., & Ghalhari, G. A. F., 2020. Machine learning to estimate surface soil moisture from remote sensing data. Water, 12(11), 3223.

Agam, N., Kustas, W. P., Anderson, M. C., Li, F., Neale, C. M., 2007. A vegetation index-based technique for spatial sharpening of thermal imagery. Remote Sens. Environ.107, 545-558.

Al-Aizari, A. R., Alzahrani, H., AlThuwaynee, O. F., Al-Masnay, Y. A., Ullah, K., Park, H. J., ... & Liu, X., 2024. Uncertainty reduction in Flood susceptibility mapping using Random Forest and eXtreme Gradient Boosting algorithms in two Tropical Desert cities, Shibam and Marib, Yemen. Remote Sens., 16(2), 336.

Alavi, M., Albaji, M., Golabi, M., Naseri, A. A., & Homayouni, S., 2024. Estimation of sugarcane evapotranspiration from remote sensing and limited meteorological variables using machine learning models. J. Hydrol, 629, 130605.

Araya, S. N., Fryjoff-Hung, A., Anderson, A., Viers, J. H., Ghezzehei, T. A., 2021. Advances in soil moisture retrieval from multispectral remote sensing using unoccupied aircraft systems and machine learning techniques. Hydrol. Earth Syst. Sci. 25, 2739-2758.

Babaeian, E., Paheding, S., Siddique, N., Devabhaktuni, V. K., & Tuller, M. 2021. Estimation of root zone soil moisture from ground and remotely sensed soil information with multisensor data fusion and automated machine learning. Remote sensing of environment, 260, 112434.

Baghdadi, N., El Hajj, M., Zribi, M., 2017. Bousbih, S. Calibration of the Water Cloud Model at C-Band for Winter Crop Fields and Grasslands. Remote Sens. 9, 969.

Bai, J., Cui, Q., Zhang, W., & Meng, L., 2019. An approach for downscaling SMAP soil moisture by combining Sentinel-1 SAR and MODIS data. Remote Sens., 11(23), 2736.

Baig, M. H. A., Zhang, L., Shuai, T., Tong, Q., 2014. Derivation of a tasselled cap transformation based on Landsat 8 at-satellite reflectance. Remote Sens. Lett. 5, 423-431.

Baldwin, D., Manfreda, S., Lin, H., & Smithwick, E. A., 2019. Estimating root zone soil moisture across the Eastern United States with passive microwave satellite data and a simple hydrologic model. Remote Sens., 11(17), 2013.

Bazzi, H., Baghdadi, N., Najem, S., Jaafar, H., Le Page, M., Zribi, M., ... & Spiliotopoulos, M., 2022. Detecting irrigation events over semi-arid and temperate climatic areas using Sentinel-1 data: case of several summer crops. Agro, 12(11), 2725.

Bhogapurapu, N., Dey, S., Homayouni, S., Bhattacharya, A., Rao, Y. S., 2022. Field-scale soil moisture estimation using sentinel-1 GRD SAR data. Adv. Space. Res.

Breiman, L., 2001 Random forests. Mach Learn. 45, 5–32.

Brocca, L., Hasenauer, S., Lacava, T., Melone, F., Moramarco, T., Wagner, W., Dorigo, W., Matgen, P., Martínez-Fernández, J., Llorens, P., Latron, J., Martin, C., Bittelli, M., 2011. Soil moisture estimation through ASCAT and AMSR-E sensors: an intercomparison and validation study across Europe. Remote Sens. Environ. 115, 3390–3408.

Carranza, C., Nolet, C., Pezij, M., van der Ploeg, M., 2021. Root zone soil moisture estimation with Random Forest. J. Hydrol. 593, 125840.

Chang, T.-Y., Wang, Y.-C., Feng, C.-C., Ziegler, A.D., Giambelluca, T.W., Liou, Y.-A., 2012. Estimation of root zone soil moisture using apparent thermal inertia with MODIS imagery over a tropical catchment in Northern Thailand. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens, 5, 752–761.

Chaves, M.M., Pereira, J.S., Maroco, J., Rodrigues, M.L., Ricardo, C.P.P., Osório, M.L., Carvalho, I., Faria, R., Pinheiro, C., 2002. How plants cope with water stress in the field? Photosynthesis and growth. Ann. Bot. 89, 907–916.

Chen, L., Xing, M., He, B., Wang, J., Shang, J., Huang, X., & Xu, M., 2021. Estimating soil moisture over winter wheat fields during growing season using machine-learning methods. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens, 14, 3706-3718.

Chen, T., & Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Cortes, C., Vapnik, V., 1995. Support-Vector Networks. Mach. Learn. 20, 273–297.

Davies, W.J., Zhang, J., 1991. Root signals and the regulation of growth and development of plants in drying soil. Annu. Rev. Plant Physiol 42, 299–306

El Hajj, M., Baghdadi, N., Zribi, M., Rodríguez-Fernández, N., Wigneron, J. P., Al-Yaari, A., ... & Calvet, J. C. 2018. Evaluation of SMOS, SMAP, ASCAT and Sentinel-1 soil moisture products at sites in Southwestern France. Remote Sens., 10(4), 569.

Fathololoumi, S., Vaezi, A. R., Alavipanah, S. K., Ghorbani, A., Biswas, A., 2020. Comparison of spectral and spatial-based approaches for mapping the local variation of soil moisture in a semi-arid mountainous area. Sci. Total Environ. 724, 138319.

Fatichi, S., Katul, G. G., Ivanov, V. Y., Pappas, C., Paschalis, A., Consolo, A., ... & Burlando, P., 2015. Abiotic and biotic controls of soil moisture spatiotemporal variability and the occurrence of hysteresis. Water Resour. Res., 51(5), 3505-3524.

Firozjaei, M.K., Kiavarz, M., Nematollahi, O., Karimpour Reihan, M., Alavipanah, S.K., 2019. An evaluation of energy balance parameters, and the relations between topographical and biophysical characteristics using the mountainous surface energy balance algorithm for land (sebal). Int. J. Remote Sens. 1-31.

Flint, A. L., Flint, L. E., 2002. 2.2 Particle Density. Methods of soil analysis: Part 4 Physical methods, 5, 229-240.

Foumelis, M., Blasco, J. M. D., Desnos, Y. L., Engdahl, M., Fernández, D., Veci, L., ..., Wong, C., 2018. ESA SNAP-StaMPS integrated processing for sentinel-1 persistent scatterer interferometry. In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. 1364-1367

Gao, B.C., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. 58, 257-266.

Gibson, J. D., 2020. Mutual Information Gain and Linear/Nonlinear Redundancy for Agent Learning, Sequence Analysis, and Modeling. Entropy, 22(6), 608

Gill, M.K., Asefa, T., Kemblowski, M., and McKee, M., 2006. Soil moisture prediction using support vector machines. J. Am. Water Resour. Assoc. 42, 1033-1046.

Gu, H., Lin, Z., Guo, W., Deb, S., 2021. Retrieving surface soil water content using a soil texture adjusted vegetation index and unmanned aerial system images. Remote Sens. 13, 145.

Guo, J., Bai, Q., Guo, W., Bu, Z., Zhang, W., 2022. Soil moisture content estimation in winter wheat planting area for multi-source sensing data using CNNR. Comput Electron Agr, 193, 106670.

Guo, X., Fang, X., Zhu, Q., Jiang, S., Tian, J., Tian, Q., & Jin, J. 2023. Estimation of Root-Zone Soil Moisture in Semi-Arid Areas Based on Remotely Sensed Data. Remote Sens, 15(8), 2003.

826 Haubrock, S.N., Chabrillat, S., Lemmnitz, C., Kaufmann, H., 2008. Surface soil moisture quantification models
827 from reflectance data under field conditions. Int. J. Remote Sens. 29, 3-29.

828 Holzman, M. E., Rivas, R. E., & Bayala, M. I., 2021. Relationship between TIR and NIR-SWIR as indicator of
829 vegetation water availability. Remote Sens., 13(17), 3371.

830 Holzman, M.E.; Rivas, R.; Bayala, M., 2014. Subsurface soil moisture estimation by VI-LST method. IEEE Geosci.
831 Remote Sens. Lett., 11, 1951–1955.

832 Hosseini, M., McNairn, H., 2017. Using multi-polarization C-and L-band synthetic aperture radar to estimate
833 biomass and soil moisture of wheat fields. Int. J. Appl. Earth Obs. Geoinf. 58, 50-64.

834 Huete, A. R., 1988. A soil-adjusted vegetation index (SAVI). Remote Sensing of Environment, 25(3), 295-309.

835 Huete, A.R., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., & Ferreira, L.G., 2002. Overview of the radiometric and
836 biophysical performance of the MODIS vegetation indices. Remote Sens. Environ, 83(1-2), 195-213

837 Hunt, Jr.E.R., Rock, B.N., 1989. Detection of changes in leaf water content using near-and middle-infrared
838 reflectances. Remote Sens. Environ. 30, 43-54.

839 Jimenez-Munoz, J. C., Cristobal, J., Sobrino, J. A., Sòria, G., Ninyerola, M., & Pons, X., 2008. Revision of the
840 single-channel algorithm for land surface temperature retrieval from Landsat thermal-infrared data. IEEE
841 Transactions on geoscience and remote sensing, 47(1), 339-349.

842 Jordan, C.F., 1969. Derivation of leaf-area index from quality of light on the forest floor. Ecology, 50, 663-666.

843 Karthikeyan, L., & Mishra, A. K., 2021. Multi-layer high-resolution soil moisture estimation using machine learning
844 over the United States. Remote Sens. Environ. 266, 112706.

845 Kisi, O.O., Dailr, A.H.H., Cimen, M.M., Shiri, J., 2012. Suspended sediment modeling using genetic programming
846 and soft computing techniques. J. Hydrol.450, 48–58.

847 Klute, A., 1988. Methods of soil analysis 2d ed., pt. 1; physical and mineralogical methods. Soil Science, 146(2),
848 138

849 Krishnan, S., & Indu, J., 2023. Assessing the potential of temperature/vegetation index space to infer soil moisture
850 over Ganga Basin. J. Hydrol, 621, 129611.

851 Kumari, N., Srivastava, A., & Dumka, U. C., 2021. A long-term spatiotemporal analysis of vegetation greenness
852 over the Himalayan Region using Google Earth Engine. Clim., 9(7), 109.

853 Laclau, B. P., and Laclau, J.-P., 2009. Growth of the whole root system for a plant crop of sugarcane under rainfed
854 and irrigated environments in Brazil. Field Crops Res. 114, 351–360. doi: 10.1016/j.fcr.2009.09.004.

855 Liang, S., 2001. Narrowband to broadband conversions of land surface albedo I: Algorithms. Remote Sens. Environ.
856 76, 213-238.

857 Liu, Q., Gu, X., Chen, X., Mumtaz, F., Liu, Y., Wang, C., Yu, T., Zhang, Y., Wang, D., Zhan, Y., 2022. Soil Moisture
858 Content Retrieval from Remote Sensing Data by Artificial Neural Network Based on Sample Optimization. Sens.
859 22, 1611.

860 Liu, S., Roberts, D. A., Chadwick, O. A., & Still, C. J., 2012. Spectral responses to plant available soil moisture in a
861 Californian grassland. Int. J. Appl. Earth Obs. Geoinf., 19, 31-44.

862 Manfreda, S., Brocca, L., Moramarco, T., Melone, F., & Sheffield, J., 2014. A physically based approach for the
863 estimation of root-zone soil moisture from surface measurements. Hydrol Earth Syst Sci, 18(3), 1199-1212.

864 Melesse, A. M., Khosravi, K., Tiefenbacher, J. P., Heddam, S., Kim, S., Mosavi, A., & Pham, B. T., 2020. River
865 water salinity prediction using hybrid machine learning models. Water, 12(10), 2951.

866 Min, X., Shangguan, Y., Li, D., & Shi, Z., 2022. Improving the fusion of global soil moisture datasets from SMAP,
867 SMOS, ASCAT, and MERRA2 by considering the non-zero error covariance. Int. J. Appl. Earth Obs. Geoinf., 113,
868 103016.

Moazenzadeh, R., Mohammadi, B., Safari, M. J. S., & Chau, K. W., 2022. Soil moisture estimation using novel bio-inspired soft computing approaches. Eng. Appl. Comput. Fluid Mech., 16(1), 826-840.

Moran, M.S., Clarke, T.R., Inoue, Y., Vidal, A., 1994. Estimating crop water deficit using the relation between surface-air temperature and spectral vegetation index. Remote Sens. Environ. 49 (3), 246–263

Nguyen, T. T., Ngo, H. H., Guo, W., Chang, S. W., Nguyen, D. D., Nguyen, C. T., ... & Hoang, N. B., 2022. A low-cost approach for soil moisture prediction using multi-sensor data and machine learning algorithm. Sci. Total Environ, 833, 155066.

Quinlan, J.R., 1992. Learning with continuous classes. In Proceedings of the Australian joint conference on artificial Intelligence. 343–348.

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W., 1974. Monitoring vegetation systems in the Great Plains with ERTS. NASA Spec. Publ, 351(1), 309.

Ryu, S., Kwon, Y.J., Kim, G., Hong, S., 2021. Temperature Vegetation Dryness Index-Based Soil Moisture Retrieval Algorithm Developed for Geo-KOMPSAT-2A. Remote Sens. 13, 2990.

Sadeghi, M., Babaeian, E., Tuller, M., Jones, S., 2017. The optical Trapezoid model: a novel approach to remote sensing of soil moisture applied to Sentinel-2 and Landsat- 8 observations. Remote Sens. Environ. 198, 52–68.

Sadeghi, M., Jones, B.S., Philpot, W.D., 2015. A linear physically-based model for remote sensing of soil moisture using shortwave infrared bands. Remote Sens. Environ. 164, 66–76.

Santos, W. J. R., Silva, B. M., Oliveira, G. C., Volpato, M. M. L., Lima, J. M., Curi, N., & Marques, J. J., 2014. Soil moisture in the root zone and its relation to plant vigor assessed by remote sensing at management scale. Geoderma, 221, 91-95.

Schaap, M.G., Leij, F.J., van Genuchten, M.Th., 2001. ROSETTA: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. J. Hydrol. 251, 163–176

Srivastava, A., Saco, P. M., Rodriguez, J. F., Kumari, N., Chun, K. P., & Yetemen, O., 2021. The role of landscape morphology on soil moisture variability in semi-arid ecosystems. Hydrol. Process. 35(1), e13990.

Tao, S., Zhang, X., Feng, R., Qi, W., Wang, Y., & Shrestha, B., 2023. Retrieving soil moisture from grape growing areas using multi-feature and stacking-based ensemble learning modeling. Comput Electron Agr, 204, 107537.

Toth, B., Szatmari, G., Takacs, K., Laborczi, A., Mako, A., Rajkai, K., Pasztor, L., 2019. Mapping soil hydraulic properties using random forest based pedotransfer functions and geostatistics. Hydrol. Earth Syst. Sci. 23, 2615–2635.

Tunçay, T., Alaboz, P., Dengiz, O., & Başkan, O., 2023. Application of regression kriging and machine learning methods to estimate soil moisture constants in a semi-arid terrestrial area. Comput Electron Agr, 212, 108118.

van Genuchten, M.T.h., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J. 44, 892–898

Wakigari, S. A., Leconte, R., 2022. Enhancing Spatial Resolution of SMAP Soil Moisture Products through Spatial Downscaling over a Large Watershed: A Case Study for the Susquehanna River Basin in the Northeastern United States. Remote Sens. 14, 776.

Wang, L., Fang, S., Pei, Z., Wu, D., Zhu, Y., & Zhuo, W., 2022. Developing machine learning models with multisource inputs for improved land surface soil moisture in China. Comput Electron Agr, 192, 106623.

Wang, L., Qu, J.J., 2007. NMDI: A normalized multi-band drought index for monitoring soil and vegetation moisture with satellite remote sensing. Geophys Res. Lett. 34.

Wang, Y., Witten, I. H., 1996. Induction of model trees for predicting continuous classes.

Xu, H., 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. Int. J. Remote Sens. 27, 3025-3033.

Yeh, I.C., Lien, C.H., 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Syst. Appl. 36, 2473-2480.

Yukseler, U., Toprak, A., Gul, E., & Dursun, O. F., 2023. Flood hazard mapping using M5 tree algorithms and logistic regression: a case study in East Black Sea Region. Earth Sci. Inform., 16(3), 2033-2047.

Zhang, L.J., Zeng, Y.J., Zhuang, R.D., Szabo, B., Manfreda, S., Han, Q.Q., Su, Z.B., 2021. In Situ Observation-Constrained Global Surface Soil Moisture Using Random Forest Model. Remote Sens. (Basel) 13.

Zhang, N., Hong, Y., Qin, Q., Liu, L., 2013. VSDI: a visible and shortwave infrared drought index for monitoring soil and vegetation moisture based on optical remote sensing. Int. J. Remote Sens. 34, 4585-4609.

Zhu, S., & Zhu, F., 2019. Cycling comfort evaluation with instrumented probe bicycle. Transp. Res. A Policy Pract, 129, 217–231.

Table 3. Vegetation, soil, and moisture indices from optical and thermal RS observations.

Variable	Acronym	Equation	Reference
Normalized Difference Vegetation Index	NDVI	$\frac{\rho_{NIR} - \rho_{Red}}{\rho_{NIR} + \rho_{Red}}$	Rouse et al. (1974)
Fraction Normalized Difference Vegetation Index	fNDVI	$(1 - NDVI)^{0.625}$	Agam et al. (2007)
Normalized Multi-Band Drought Index	NMDI	$\frac{\rho_{NIR} - (\rho_{SWIR1} - \rho_{SWIR2})}{\rho_{NIR} + (\rho_{SWIR1} - \rho_{SWIR2})}$	Wang & Qu (2007)
Normalized Difference Water Index	NDWI	$\frac{\rho_{NIR} - \rho_{SWIR1}}{\rho_{NIR} + \rho_{SWIR1}}$	Gao (1996)
Modified Normalized Difference Water Index	MNDWI	$\frac{\rho_{Green} - \rho_{SWIR1}}{\rho_{Green} + \rho_{SWIR1}}$	Xu (2006)
Normalized Soil Moisture Index	NSMI	$\frac{\rho_{SWIR1} - \rho_{SWIR2}}{\rho_{SWIR1} + \rho_{SWIR2}}$	Haubrock et al. (2008)
Visible and Shortwave Infrared Drought Index	VSDI	$1 - [(\rho_{SWIR1} - \rho_{Blue}) + (\rho_{Red} - \rho_{Blue})]$	Zhang et al. (2013)
Global Vegetation Moisture Index	GVMI	$\frac{(\rho_{NIR} + 0.1) - (\rho_{SWIR1} + 0.02)}{(\rho_{NIR} + 0.1) + (\rho_{SWIR1} + 0.02)}$	Ceccato et al. (2002)
Moisture Stress Index	MSI	$\frac{\rho_{SWIR1}}{\rho_{NIR}}$	Hunt. & Rock (1989)
Enhanced Vegetation Index	EVI	$\frac{2.5(\rho_{NIR} - \rho_{Red})}{\rho_{NIR} + 6\,\rho_{Red} - 7.5\rho_{Blue} + 1}$	Huete et al. (2002)
Soil Adjusted Vegetation Index	SAVI	$1.25\frac{(\rho_{NIR} - \rho_{Red})}{(\rho_{NIR} + \rho_{Red} + 0.25)}$	Huete (1988)
Ratio Vegetation Index	RVI	$\frac{\rho_{NIR}}{\rho_{Red}}$	Jordan (1969)
NIR Transformed Reflectance	NTR	$\frac{(1 - \rho_{NIR})^2}{2\,\rho_{NIR}}$	Kubelka & Munk (1931) and Babaeian et al. (2021)
Temperature Vegetation Dryness Index	TVDI	$\frac{LST - LST_{Min}}{LST_{Max} - LST_{Min}}$, $LST = \gamma \times \left[\frac{1}{\varepsilon} \times (\psi_1 \times L_{sensor} + \psi_2) + \psi_3\right] + \delta$	Moran et al. (1994), Jiménez-Muñoz et al. (2008)
Albedo	-	$0.356\,\rho_{Blue} + 0.130\,\rho_{Green} + 0.373\rho_{Red} + 0.085\rho_{NIR} + 0.072\rho_{SWIR1} + 0.072\rho_{SWIR2} - 0.0018$	Liang (2001)
Brightness	-	$0.3029\rho_{Blue} + 0.2786\rho_{Green} + 0.4733\rho_{Red} + 0.5599\rho_{NIR} + 0.508\rho_{SWIR1} + 0.1872\rho_{SWIR2}$	Baig et al. (2014)
Greenness	-	$-0.2941\rho_{Blue} - 0.243\rho_{Green} - 0.5424\rho_{Red} + 0.7276\rho_{NIR} + 0.0713\rho_{SWIR1} - 0.1608\rho_{SWIR2}$	
Wetness	-	$0.1511\rho_{Blue} + 0.1973\rho_{Green} + 0.3283\rho_{Red} + 0.3407\rho_{NIR} - 0.7117\rho_{SWIR1} - 0.4559\rho_{SWIR2}$	

Table 4. Combination of different models using the IGR technique in Scenario-1.

Depths (cm)	Model 1	Model 2	Model 3	Model 4
		Input features		
0-10	NTR, NIR, Wetness, RVI, MSI, GVMI, NDWI, NDVI, fNDVI, Greenness, SAVI, NMDI, TVDI, Brightness, LST, B, NSMI, SWIR2, Albedo, G, R, SWIR1, EVI, MNDWI, K_s, θ_{FC}, θ_{PWP}, VSDI, Sand, Silt, Clay, BD, ϕ, θr, θs, α, n, T, W, RH, Rn, S_h, E	NTR, NIR, Wetness, RVI, MSI, GVMI, NDWI, NDVI, fNDVI, Greenness, SAVI, NMDI, TVDI, Brightness, LST, B, NSMI, SWIR2, Albedo, G, R, SWIR1, EVI, MNDWI, K_s, θ_{FC}, θ_{PWP}	NTR, Wetness, Greenness, NMDI, TVDI, Brightness, LST, B, NSMI, SWIR2, Albedo, G, R, SWIR1, EVI, MNDWI, K_s, θ_{FC}, θ_{PWP}	NTR, Wetness, Greenness, NMDI, TVDI, Brightness, LST, B, NSMI, SWIR2, Albedo, G, R, SWIR1, EVI, MNDWI
10-30	R, Wetness, Greenness, SAVI, RVI, MSI, GVMI, NDWI, NMDI, NDVI, fNDVI, SWIR1, NIR, G, NTR, NSMI, TVDI, B, SWIR2, Albedo, Brightness, LST, EVI, MNDWI, VSDI, K_s, θ_{FC}, θ_{PWP}, Sand, Silt, Clay, BD, ϕ, θr, θs, α, n, T, W, RH, Rn, S_h, E	R, Wetness, Greenness, SAVI, RVI, MSI, GVMI, NDWI, NMDI, NDVI, fNDVI, SWIR1, NIR, G, NTR, NSMI, TVDI, B, SWIR2, Albedo, Brightness, LST, EVI, MNDWI, K_s	R, Wetness, NTR, NSMI, TVDI, B, SWIR2, Albedo, Brightness, LST, EVI, MNDWI, K_s	R, Wetness, NTR, NSMI, TVDI, B, SWIR2, Albedo, Brightness, LST, EVI
30-50	R, B, Albedo, Wetness, Greenness, SAVI, RVI, MSI, GVMI, NDWI, NMDI, NDVI, fNDVI, SWIR1, G, NTR, NIR, TVDI, Brightness, NSMI, SWIR2, LST, K_s, EVI, MNDWI, VSDI, θ_{FC}, θ_{PWP}, Sand, Silt, Clay, BD, ϕ, θr, θs, α, n, T, W, RH, Rn, S_h, E	R, B, Albedo, Wetness, Greenness, SAVI, RVI, MSI, GVMI, NDWI, NMDI, NDVI, fNDVI, SWIR1, G, NTR, NIR, TVDI, Brightness, NSMI, SWIR2, LST, K_s	R, NTR, TVDI, Brightness, NSMI, SWIR2, LST, K_s	R, NTR, TVDI, Brightness, NSMI, SWIR2, LST
50-70	Wetness, Albedo, Greenness, SAVI, RVI, MSI, GVMI, NDWI, NMDI, NDVI, fNDVI, SWIR1, R, G, B, NTR, NIR, TVDI, SWIR2, NSMI, Brightness, LST, n, α, K_s, MNDWI, EVI, VSDI, Sand, Silt, Clay, BD, θ_{FC}, θ_{PWP}, ϕ, θs, θr, T, W, RH, Rn, S_h, E	Wetness, Albedo, Greenness, SAVI, RVI, MSI, GVMI, NDWI, NMDI, NDVI, fNDVI, SWIR1, R, G, B, NTR, NIR, TVDI, SWIR2, NSMI, Brightness, LST, n, α, K_s	Wetness, NTR, TVDI, SWIR2, NSMI, Brightness, LST, K_s	Wetness, NTR, TVDI, SWIR2, NSMI, Brightness, LST
70-90	Wetness, Albedo, Greenness, SAVI, RVI, MSI, GVMI, NDWI, NMDI, NDVI, fNDVI, SWIR1, R, G, NTR, NIR, B, TVDI, SWIR2, NSMI, Brightness, LST, K_s, MNDWI, EVI, VSDI, Sand, Silt, Clay, BD, θ_{FC}, θ_{PWP}, ϕ, θs, θr, n, α, T, W, RH, Rn, S_h, E	Wetness, Albedo, Greenness, SAVI, RVI, MSI, GVMI, NDWI, NMDI, NDVI, fNDVI, SWIR1, R, G, NTR, NIR, B, TVDI, SWIR2, NSMI, Brightness, LST, K_s	Wetness, G, NTR, B, TVDI, SWIR2, NSMI, Brightness, LST, K_s	Wetness, G, NTR, B, TVDI, SWIR2, NSMI, Brightness, LST

Ethical Statement for Remote Sensing Applications: Society and Environment

Now, I, Mona Golabi, consciously assure that for the manuscript, Enhanced Root Zone Soil Moisture Monitoring Using Multitemporal Remote Sensing Data and Machine Learning Techniques, the following is fulfilled:

1) This material is the author's original work, which has not been previously published elsewhere.

2) The paper is not currently being considered for publication elsewhere.

3) The paper reflects the author's research and analysis wholly and truthfully.

4) The paper properly credits the meaningful contributions of co-authors and co-researchers.

5) The results are appropriately placed in the prior and existing research context.

6) All sources used are correctly disclosed (correct citation). Literally copying of text must be indicated as such by using quotation marks and giving proper references.

7) All authors have been personally and actively involved in substantial work leading to the paper and will take public responsibility for its content.

The violation of the Ethical Statement rules may result in severe consequences.

I agree with the above statements and declare that this submission follows the Remote Sensing Applications: Society and Environment policies outlined in the Guide for Authors and in the Ethical Statement.

Date: May 04, 2024

Corresponding author’s signature:

Mona Golabi

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: