Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Physics-Guided Multi-Source Stacking Ensemble for Regional Soil-Moisture Prediction Using Remote-Sensing and Environmental Data

M Pyingkodi
Kongu Engineering College
K. Thenmozhi
kthenmozhi@rvu.edu.in
RV University
Parvathavarthini S
Kongu Engineering College
Suresh Palarima
University of Technology and Applied Sciences-Salalah

Research Article

Keywords: Soil moisture, Multi-source data fusion, Physics-guided modelling, Stacking ensemble, Machine learning, Water Cloud Model

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11567560/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Physics-Guided Multi-Source Stacking Ensemble for Regional Soil-Moisture Prediction Using Remote-Sensing and Environmental Data

Pyingkodi M[a], Thenmozhi K[b] , Parvathavarthini S[c], Suresh Palarima[d]

[a]Associate Professor, Department of MCA, Kongu Engineering College, Erode, Tamil Nadu. India. Email:pyingkodikongu@gmail.com

[b]Assistant Professor, School of Computer Science and Engineering, RV University, Bengaluru, Karnataka. India. E-mail: kthenmozhi@rvu.edu.in / thenmegu@gmail.com

[c]Associate Professor, Department of Computer Technology, Kongu Engineering College, Erode, Tamil Nadu. India. Email: varthinis@gmail.com

[d]College of Computing and Information Sciences, University of Technology and Applied Sciences, Salalah, Sultanate of Oman, suresh.palarimath@utas.edu.om

Corresponding Author:

Dr.K.Thenmozhi, Assistant Professor, School of Computer Science and Engineering, RV University, Bengaluru, Karnataka. India.

E-mail: kthenmozhi@rvu.edu.in / thenmegu@gmail.com

Abstract

Regional soil-moisture prediction is challenging because of its pronounced spatial and temporal variability and the nonlinear relationships between remote-sensing observations and environmental conditions. This study proposes a physics-guided hybrid ensemble approach that integrates WCM-derived features with complementary machine-learning models for regional soil-moisture prediction in Tamil Nadu, India. A dataset comprising 15,575 observations from 391 spatial locations distributed across 28 districts, between October 2015 and August 2022, was developed through a combination of Sentinel-1 SAR, Sentinel-2

optical, CHIRPS rainfall, ERA5-Land temperature, terrain, soil, land-use/land-cover, temporal, and WCM-derived predictors. Monthly NASA–USDA HSL Enhanced SMAP surface soil moisture was used as the reference target. At Level 0, predictions were generated independently using Random Forest (RF), XGBoost, and FT-Transformer, while Ridge Regression served as the Level-1 meta-learner. Out-of-fold (OOF) predictions were used to construct the Level-1 meta-features, thereby reducing the risk of information leakage during stacking. The proposed Hybrid framework achieved the best performance, with an RMSE of 0.0591, MAE of 0.0354, R^2 of 0.9964, NSE of 0.9964, and a correlation coefficient R of 0.9982. XGBoost was the best-performing individual learner, with an RMSE of 0.0779 and an R^2 of 0.9937. Compared with XGBoost, the Hybrid framework reduced RMSE and MSE by 24.1% and 42.3%, respectively. Diagnostic analyses also showed that the Hybrid model performed better in comparison with the observations and reduced the spread of errors. The proposed prediction-level stacking framework effectively combines complementary learning mechanisms for regional soil-moisture prediction. WCM-derived features provide physics-guided information within the multi-source predictor space. However, independent spatial and temporal validation is required to assess the transferability of the framework.

Keywords : *Soil moisture; Multi-source data fusion; Physics-guided modelling; Stacking ensemble; Machine learning; Water Cloud Model*

1. Introduction

Soil moisture is an important land-surface variable that regulates water and energy exchange. It influences evapotranspiration, runoff, vegetation dynamics, and hydrological processes, highlighting the need for accurate soil-moisture information for agricultural monitoring, drought assessment, water-resource management, and hydrological modelling. (Babaeian et al., 2019). However, its high spatial and temporal variability makes accurate monitoring

challenging, particularly across regions with heterogeneous topography. Satellite remote sensing is a useful tool to monitor soil moisture on a large scale. One of the major advantages of SAR is its sensitivity to the near-surface dielectric properties of the land surface, together with its ability to provide observations under a wide range of atmospheric conditions. Sentinel-1 has consequently been widely used for soil moisture estimation. Besides soil moisture, however, the radar backscatter is affected by vegetation attenuation and scattering, surface roughness, crop structure, and acquisition geometry. These factors cause some uncertainty in the radar-soil moisture relationship (Bauer-Marschallinger et al., 2018; Benninga et al., 2022).

Information derived from optical observation complements vegetation and land-surface information. Sentinel-2 multispectral observations can be used to characterize the vegetation status and surface conditions by using spectral bands and vegetation indices. Accordingly, Sentinel-1 and Sentinel-2 can provide complementary information for soil-moisture estimation by combining microwave sensitivity to surface conditions with optical information on vegetation and land cover. However, integrating radar, optical, climatic, terrain, and soil variables produces a heterogeneous predictor space with potentially complex and nonlinear relationships (Bousbih et al., 2018). However, when these variables are combined with radar, optical, climatic, terrain and soil, the predictor space becomes a mixture of different variables, which necessarily have complex and non-linear relationships. Such nonlinear relationships can be effectively modelled using machine learning approaches. Random Forest (RF) can capture complex interactions between environmental predictors and can be used for soil moisture estimation (Carranza et al., 2021, van den Hoogen et al.,2026).The machine learning approach and recent computations are used. Another tree-based learning method, XGBoost, sequentially improves predictions by reducing residual errors while incorporating regularization to control model complexity (Chen & Guestrin, 2016). In this sense, these

properties make RF and XGBoost suitable for modelling heterogeneous environmental predictor spaces. But these purely data-driven models do not explicitly model the physics that influence the radar observations.

The Water Cloud Model (WCM) is a physically interpretable model for radar backscatter from vegetated surfaces that conceptually separates radar backscatter into vegetation and soil-scattering components. Soil moisture estimation using Sentinel-1 data with WCM based approaches has been developed to consider the effects of vegetation (Bai et al., 2017; Bai et al., 2022). The radar response can be significantly affected by vegetation attenuation and scattering and use of this information from the WCM can lead to physically meaningful predictors for data-driven models. However, the simplified WCM formulation may not capture all nonlinear interactions present in heterogeneous regional environments. As more information becomes available from multiple sources, The increasing availability of heterogeneous environmental observations has motivated the use of learning architectures capable of representing nonlinear interactions among structured predictors. Tree-based methods such as RF and XGBoost capture nonlinear relationships through ensembles of decision trees, whereas transformer-based approaches provide an attention-based mechanism for modelling feature interactions. FT-Transformer proposes a novel representation-learning mechanism for structured predictors by applying feature tokenization and self-attention to tabular data, which adapts the Transformer architecture for tabular data (Gorishniy et al., 2021). It is important to mention here that FT-Transformer is not specifically designed for soil moisture prediction, but is developed for tabular learning. It clearly assists local soil-moisture prediction using satellite pictures and weather or environmental data, which allies intently with our proposed remote-sensing and environmental-data framework (Bramley et al.,2024). Its relevance to this study is related to the structured and heterogeneous nature of

the predictor space, where interactions among the remote sensing, climatic, terrain, soil, land-cover and physics-derived variables are anticipated.

Although remote-sensing- and machine-learning-based soil-moisture estimation has advanced substantially, the integration of physics-guided features with complementary learning architectures remains insufficiently investigated for regional soil-moisture prediction. Even though studies have focused on WCM-based retrieval and machine learning methods, the integration of tree-based learners and a tabular Transformer within a prediction-level stacked ensemble has received limited attention. Moreover, the use of the WCM-derived predictors with RF, XGBoost and FT-Transformer to perform the meta-learning provides an opportunity to evaluate whether physically informed features and complementary model representations can jointly improve prediction performance. This paper introduces a hybrid stacked ensemble model for soil moisture prediction over Tamil Nadu, India, which combines multi-source satellite and environmental data with a physics-guided model. The framework takes surface soil moisture as measured by SMAP as the target variable and combines Sentinel-1 SAR, Sentinel-2 optical observations, rainfall, temperature, terrain attributes, soil texture, land-use/land-cover information, temporal variables and WCM-derived predictors. Three base learners are RF, XGBoost, and FT-Transformer are trained independently, and their predictions are subsequently combined using Ridge Regression as the meta-learner, with predictions made by each base learner, and then combined by a meta-learner model, namely Ridge regression. Remote sensing and VIS–NIR sensing offer a quick way to assess soil properties, but their accuracy can vary with soil and environmental conditions. Combining remote-sensing and environmental data through physics-guided ensemble models can improve regional soil-moisture prediction (Ann-Marie Fortuna et al., 2025)

The study has three objectives. First, it constructs a multi-source predictor space, which combines satellite, environmental, temporal, and WCM-derived information to predict soil moisture. Secondly, it assesses the predictive performance of RF, XGBoost and FT-Transformer on this mixed feature space. Third, it answers the question of whether prediction-level stacking of these complementary learners using Ridge Regression improves performance relative to the individual models. The key contribution of this study is the combination of WCM physics-guided feature generation with complementary learning techniques based on trees and Transformer into a stacked ensemble model. The approach does not presume that one model family is better than another. Instead, it tests if these three models, RF, XGBoost and FT-Transformer, are able to capture complementary parts of the multi-source predictor space and whether the combination of the prediction of these three models yields better soil moisture estimation.

2. Data and Study Area

2.1 Study Area

Tamil Nadu is located in southern India and was selected as the study region for regional soil moisture prediction. Based on the spatial coordinates of the observations used in this study, the dataset extends from 8.10° to 13.34° N and 76.48° to 80.24° E**.** The dataset comprises 15,575 observations from 391 unique spatial locations**,** distributed across 28 districts**.** The region exhibits substantial climatic and environmental variability. Tamil Nadu experiences distinct seasonal conditions, with the Northeast Monsoon contributing substantially to rainfall during October–December**,** while rainfall and temperature vary spatially across the state. These variations influence soil moisture availability and its temporal dynamics. The dataset

144 covers October 2015 to August 2022, with observations available across all 12 calendar
145 months, enabling the assessment of seasonal soil moisture variability.

146 The study region also exhibits considerable topographic heterogeneity. Elevation in the
147 dataset ranges from 4 to 2044 m, while slope ranges from 0.03° to 22.56°**.** The dataset further
148 represents four soil-texture classes and six land-use/land-cover classes, capturing differences
149 in soil and surface characteristics that influence infiltration, water storage, drainage,
150 vegetation response, and soil moisture dynamics. Climatic conditions are represented by
151 rainfall and temperature variables. Rainfall values range from 0 to 914.05 mm, whereas
152 temperature ranges from 288.42 to 306.65 K. The dataset also includes seasonal and monsoon
153 indicators, allowing temporal variations in environmental forcing to be incorporated into the
154 prediction framework. This combination of climatic, topographic, soil, vegetation, and land-
155 cover heterogeneity provides an appropriate setting for evaluating the proposed regional soil-
156 moisture prediction framework. This environmental heterogeneity, together with the
157 availability of multi-source satellite observations, makes Tamil Nadu an appropriate study
158 region for assessing the proposed physics-guided hybrid ensemble framework as shown in
159 Fig. 1.

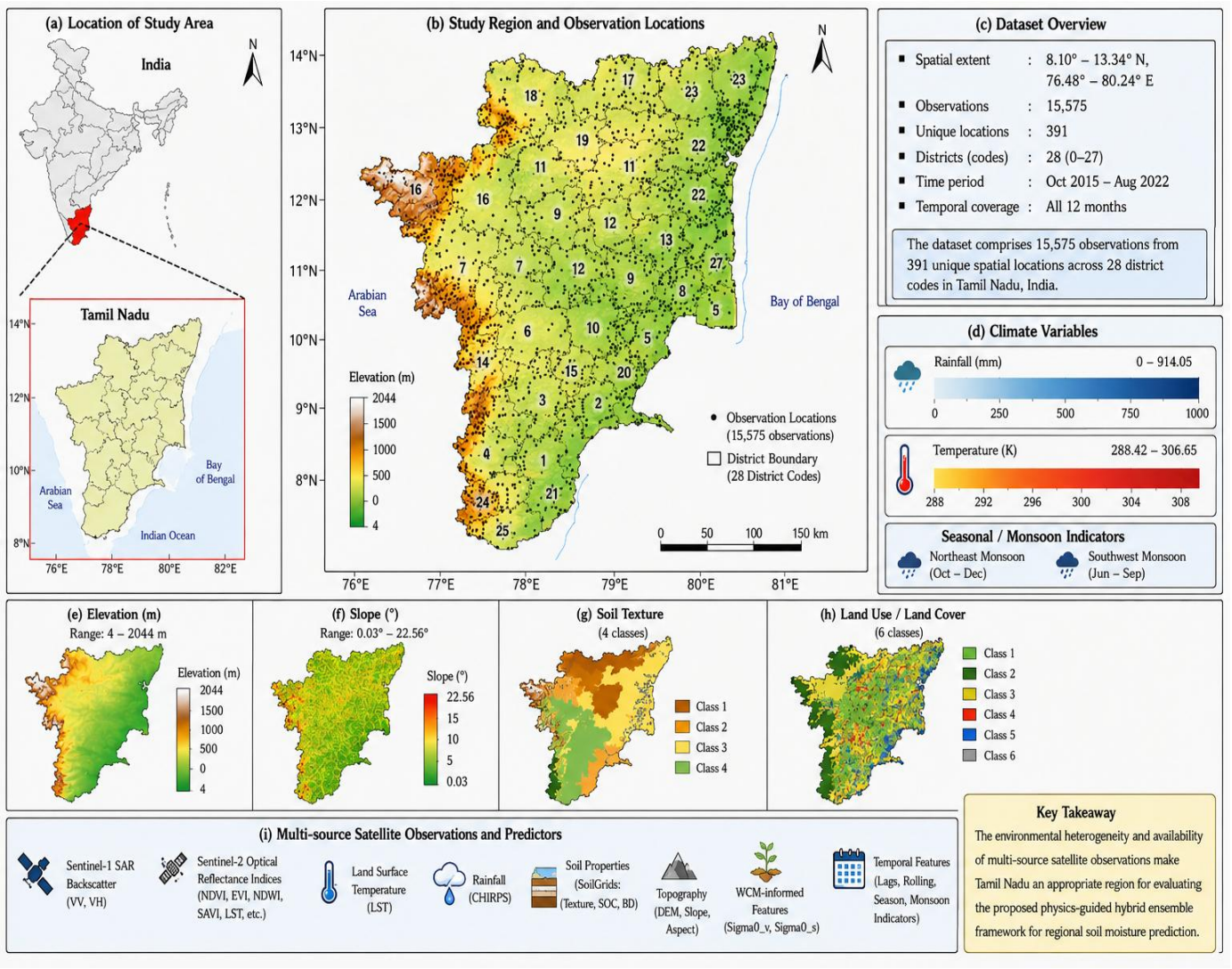


Fig. 1. Study Region And Characteristics Of The Multi-Source Dataset Used For Regional Soil-Moisture Prediction.

2.2 Multi-Source Data

In the present study, integrates active-microwave, optical, precipitation, temperature, topographic, soil, and land-use/land-cover data to characterize the factors controlling soil-moisture variability. These datasets provide complementary information on microwave

scattering, vegetation and spectral characteristics, atmospheric forcing, terrain, soil properties, and land-surface conditions, terrain and soil properties and land surface conditions. Initial geospatial processing and data acquisition are done on Google Earth Engine (GEE) platform. The Sentinel-1 SAR data used were from ESA/Copernicus Sentinel-1 Ground Range Detected (GRD). The data are C-band synthetic aperture radar (SAR) observations in both VV and VH polarization. The nominal resolution of this study is 10 m. VV and VH backscatter were used as key microwave predictors and as inputs for subsequent WCM-based feature generation (Torres et al., 2012). The Sentinel-2 MSI Level-2A surface reflectance data has been extracted from the ESA/Copernicus harmonized collection. The selected spectral bands were B2, B3, B4, B8, B11, and B12. Bands B2, B3, B4, and B8 have a spatial resolution of 10 m, whereas B11 and B12 have a resolution of 20 m. The spectral observations were processed for the characterization of the vegetation and surface conditions, as well as for the calculation of the vegetation and water-related indices that were part of the predictor set (Drusch et al., 2012). The daily precipitation data from CHIRPS were provided by Climate Hazards Center. CHIRPS provides daily precipitation estimates at approximately 0.05° (about 5 km) spatial resolution. Daily precipitation observations were summed to obtain monthly accumulated rainfall, which was subsequently used as a climatic predictor of soil-moisture variability (Funk et al., 2015).

The Daily Aggregated data ERA5-Land were retrieved from the European Centre for Medium-Range Weather Forecasts (ECMWF). The temperature_2m variable was extracted and aggregated to get the monthly mean temperature. ERA5-Land provides hourly land-surface variables at an approximately 9-km spatial resolution, enabling the representation of atmospheric and land-surface conditions at relatively high temporal frequency (Muñoz-Sabater et al., 2021). The Shuttle Radar Topography Mission (SRTM) GL1 digital elevation model was used to obtain topographic information. The data set contains data of about 30 m

(1 arc-second) elevation. Terrain variables: elevation, slope and aspect were added as predictors. Soil information was provided by the OpenLandMap USDA Soil Texture Classification which includes soil texture classes down to 250 m resolution. The SoilTexture variable was added as a categorical soil-property predictor (Farr et al., 2007; Hengl, 2018). The land-use/land-cover data were derived from ESA WorldCover 2021 v200 – an 11-class global land-cover product at 10 m spatial resolution. The LULC variable was added to account for variations in land surface properties that are important for vegetation, surface roughness, infiltration and soil moisture dynamics (Zanaga et al., 2022). Spatial and temporal integration of the resulting datasets were carried out to create the predictor space to be used in the modelling framework. The SMAP-derived surface soil moisture product is considered as a target reference product separately and is explained in Section 2.3. Spatial and temporal characteristics of the data sets, their resolutions and the variables analyzed are summarized in Table 1.

2.3 Reference Soil Moisture Data

The reference data were obtained from the NASA–USDA HSL Enhanced SMAP 10-km Soil Moisture product. The ssm (surface soil moisture) band was used as the reference target and expressed in millimetres (mm). The ssm (surface soil moisture) band was used as the reference target. Monthly mean values were calculated from the available daily observations at each spatial location. The last data set was from October 2015 to August 2022. To avoid the possibility of target leakage, the SoilMoisture variable was not included in the predictor matrix and only used as the dependent variable (Mladenova et al., 2020).

213

214 **Table 1**

215 **Multi-Source Datasets Used For Soil Moisture Prediction**

Data category	Dataset	Provider	Spatial resolution	Temporal processing	Variables used
Active microwave	Sentinel-1 SAR GRD	ESA/Copernicus	10 m	Satellite observations	VV, VH
Optical	Sentinel-2 MSI Level-2A SR Harmonized	ESA/Copernicus	10–20 m	Satellite observations	B2, B3, B4, B8, B11, B12
Precipitation	CHIRPS Daily	UCSB Climate Hazards Center	~5 km	Monthly accumulation	Rainfall
Temperature	ERA5-Land Daily Aggregated	ECMWF	~9 km	Monthly mean	Temperature
Topography	SRTM GL1	NASA/USGS	~30 m	Static	Elevation, Slope, Aspect
Soil	OpenLandMap USDA Soil Texture	EnvirometrIX/OpenLandMap	250 m	Static	SoilTexture

Land cover	ESA WorldCover 2021 v200	ESA	10 m	Static	LULC

2.4 Data Preprocessing and Harmonization

The data from Sentinel-1 were filtered to the study area in Tamil Nadu, IW mode, and VV/VH polarizations. The preprocessing chain implemented in Google Earth Engine (GEE) included orbit correction, border-noise and thermal-noise removal, radiometric calibration, and terrain correction. To get representative VV and VH backscatter values, monthly median composites were created to provide dB values for each polarization, and these were used as predictors and for WCM-based feature generation (Torres et al., 2012; Gorelick et al., 2017). The Sentinel-2 Level-2A harmonized surface-reflectance data was pre-processed using a 20% scene-level cloud-coverage threshold, followed by pixel-level cloud and cirrus masking using the QA60 band. Selected bands (B2, B3, B4, B8, B11, B12) were then made into reflectance with a scale factor of 10,000, and were then composited on a monthly basis. NDVI, EVI, SAVI, and NDWI were calculated from the processed Sentinel-2 bands following their standard formulations (Drusch et al., 2012; Gorelick et al. (2017). Accumulated precipitation was calculated by summing the daily CHIRPS data for each calendar month, at ~0.05° resolution (Funk et al., 2015). The mean temperature was computed by taking the 2-m air temperature from ERA5-Land at a resampling grid of 11.1 km (Muñoz-Sabater et al., 2021). The monthly variables were combined with the other predictors to model soil moisture prediction is given in the Table 2

Table 2

Data Preprocessing and Harmonization

Dataset	Initial processing	Temporal aggregation	Resulting variables
Sentinel-1 SAR	Study-area filtering; IW mode; VV/VH	Monthly	VV, VH
Sentinel-2 MSI SR Harmonized	Cloud filtering (<20%); QA60 cloud/cirrus masking; reflectance	Monthly median	B2, B3, B4, B8, B11, B12
Sentinel-2 derived indices	Index calculation from processed spectral bands	Monthly	NDVI, EVI, SAVI, NDWI
CHIRPS Daily	Study-area and temporal filtering	Monthly sum	Rainfall
ERA5-Land Daily	Study-area and temporal filtering	Monthly	Temperature
SRTM GL1	Terrain-variable derivation	Static	Elevation, Slope,
OpenLandMap	Spatial extraction	Static	SoilTexture
ESA WorldCover 2021	Spatial extraction	Static	LULC
NASA–USDA HSL SMAP	ssm selection	Monthly mean	SoilMoisture

2.5 Static Variables and Temporal Harmonization

The SRTM GL1 DEM was used to extract the topographic variables including elevation, slope, and aspect, soil texture data came from OpenLandMap USDA Soil Texture, and soil use data was sourced from ESA WorldCover 2021 v200. The following variables were used without any temporal variation and were not composited (Farr et al, 2007; Hengl, 2018; Zanaga et al, 2022). Median composites of Sentinel-1 and Sentinel-2 data, monthly sums of CHIRPS data on precipitation and monthly means of ERA5-Land data on temperature were

used for the harmonization of dynamic datasets. Dynamic datasets were harmonized using median composites from Sentinel-1 and Sentinel-2, sums for precipitation for CHIRPS and means for temperature for ERA5-Land, all monthly. To ensure temporal consistency with the predictors, available SMAP surface-soil-moisture observations were aggregated independently to monthly means. The data sets were then layered together based on the spatial and temporal coordinates. The native spatial resolutions were preserved and the spatial integration was done in accordance with the extraction and aggregation steps in the modelling workflow, rather than assuming a common spatial resolution.

3. Soil-Moisture Modelling Framework

The proposed framework combines multi-source environmental observations, physics-based features, and diverse machine-learning models to estimate soil moisture (Fig. 2). Soil-moisture variability was characterized by the Sentinel-1 SAR, Sentinel-2 optical, climatic, terrain, soil and LULC predictors, as well as the temporal predictors. The framework consists of four stages: (i) Multi-source data preparation and integration; (ii) Generation of physics-guided predictors by the Water Cloud Model (WCM); (iii) Independent prediction using Random Forest (RF), XGBoost and FT-Transformer; and (iv) Prediction-level stacking by Ridge Regression. The WCM stage calculates Vegetation Water Content (VWC), vegetation attenuation (τ), vegetation backscatter (σ^0veg), soil backscatter (σ^0soil), and WCM-based soil moisture (WCM_SM). These variables are used as predictors and not as the final retrieval product. Finally, the resulting predictor matrix is independently processed by RF, XGBoost, and FT-Transformer and their predictions are then aggregated by the Ridge Regression meta-learner to leverage complementary information of the heterogeneous base models (Inoubli et al., 2022; Singh et al., 2023; Li et al., 2024; Chaudhary et al., 2022; Zhou et al., 2024).

3.1 Multi-Source Data and Predictor Preparation

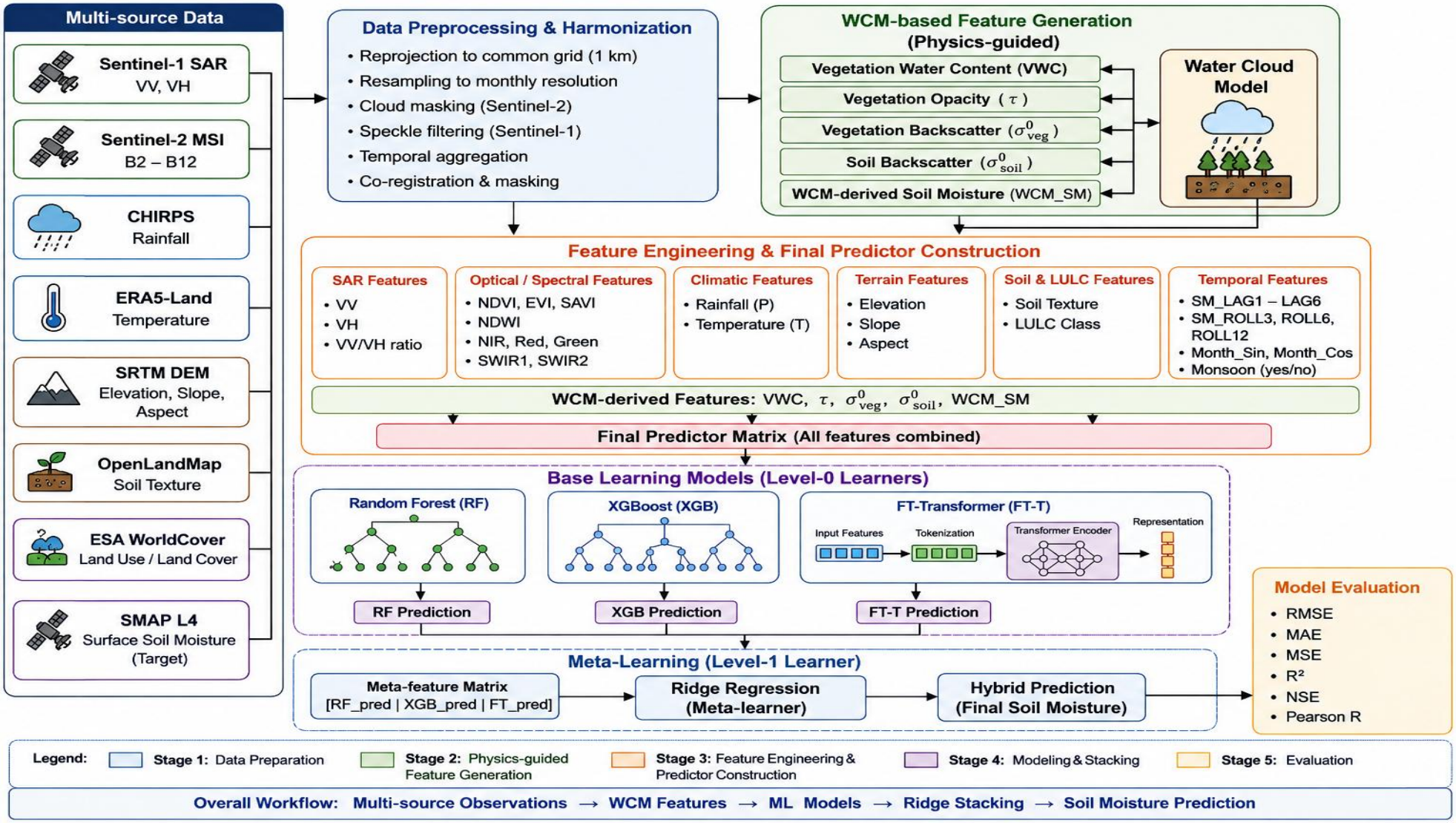
The predictor matrix included Sentinel-1 VV/VH backscatter and VV/VH ratio; Sentinel-2 bands B2, B3, B4, B8, B11, and B12 and derived NDVI, EVI, SAVI and NDWI; monthly CHIRPS rainfall and ERA5-Land temperature; and static variables such as elevation, slope, aspect, soil texture, and LULC. These variables provide complementary information on microwave scattering, vegetation characteristics, climatic forcing, and environmental conditions (Chaudhary et al., 2022; Batchu et al., 2023; Zhou et al., 2024). The reference target was the NASA–USDA HSL Enhanced SMAP surface soil-moisture product, which was aggregated on a monthly time scale from April 2015 to August 2022. To avoid data leakage, the target was not included in the predictor matrix. Median composites of the Sentinel-1/2 data and monthly accumulated rainfall for CHIRPS were used to harmonise dynamic predictors, while monthly mean temperature was harmonized using ERA5-Land data. Predictors and targets were matched based on their spatial and temporal coordinates to obtain the resulting predictors and target. A summary of the complete predictor set is given in Table 3.

Table 3

Multi-Source Datasets Used for Soil-Moisture Prediction

Data source	Dataset / collection	Variables used	Processing	Role
Sentinel-1	S1 GRD	VV, VH, VV/VH	Monthly median	SAR surface
Sentinel-2	S2 SR Harmonized	B2, B3, B4, B8, B11, B12, NDVI, NDWI,	Cloud/cirrus masking; monthly median	Optical and vegetation
SMAP	NASA–USDA HSL SMAP 10.1	ssm → SoilMoisture	Monthly mean	Reference target
CHIRPS	CHIRPS Daily	Rainfall	Monthly sum	Precipitation forcing
ERA5-Land	Daily Aggregated	temperature_2m → Temperature	Monthly mean	Atmospheric forcing

Data source	Dataset / collection	Variables used	Processing	Role
SRTM	SRTM GL1	Elevation, Slope, Aspect	Terrain derivation	Topographic information
OpenLandMap	USDA Soil Texture	SoilTexture	Spatial extraction	Soil property
ESA WorldCover	WorldCover 2021 v200	LULC	Spatial extraction	Land-cover information


Multi-source Data
Sentinel-1 SAR
VV, VH
Sentinel-2 MSI
B2 – B12
CHIRPS
Rainfall
ERA5-Land
Temperature
SRTM DEM
Elevation, Slope, Aspect
OpenLandMap
Soil Texture
ESA WorldCover
Land Use / Land Cover
SMAP L4
Surface Soil Moisture (Target)
Data Preprocessing & Harmonization
• Reprojection to common grid (1 km)
• Resampling to monthly resolution
• Cloud masking (Sentinel-2)
• Speckle filtering (Sentinel-1)
• Temporal aggregation
• Co-registration & masking
WCM-based Feature Generation (Physics-guided)
Vegetation Water Content (VWC)
Vegetation Opacity (τ)
Vegetation Backscatter (σ^0_{veg})
Soil Backscatter (σ^0_{soil})
WCM-derived Soil Moisture (WCM_SM)
Water Cloud Model
Feature Engineering & Final Predictor Construction
SAR Features
• VV
• VH
• VV/VH ratio
Optical / Spectral Features
• NDVI, EVI, SAVI
• NDWI
• NIR, Red, Green
• SWIR1, SWIR2
Climatic Features
• Rainfall (P)
• Temperature (T)
Terrain Features
• Elevation
• Slope
• Aspect
Soil & LULC Features
• Soil Texture
• LULC Class
Temporal Features
• SM_LAG1 – LAG6
• SM_ROLL3, ROLL6, ROLL12
• Month_Sin, Month_Cos
• Monsoon (yes/no)
WCM-derived Features: VWC, τ, σ^0_{veg}, σ^0_{soil}, WCM_SM
Final Predictor Matrix (All features combined)
Base Learning Models (Level-0 Learners)
Random Forest (RF)
XGBoost (XGB)
FT-Transformer (FT-T)
Input Features
Tokenization
Transformer Encoder
Representation
RF Prediction
XGB Prediction
FT-T Prediction
Meta-Learning (Level-1 Learner)
Meta-feature Matrix
[RF_pred | XGB_pred | FT_pred]
Ridge Regression (Meta-learner)
Hybrid Prediction (Final Soil Moisture)
Model Evaluation
• RMSE
• MAE
• MSE
• R²
• NSE
• Pearson R
Legend:
Stage 1: Data Preparation
Stage 2: Physics-guided Feature Generation
Stage 3: Feature Engineering & Predictor Construction
Stage 4: Modeling & Stacking
Stage 5: Evaluation
Overall Workflow: Multi-source Observations → WCM Features → ML Models → Ridge Stacking → Soil Moisture Prediction

285

286

287 **Fig. 2. Schematic overview of the Proposed Work**

3.2 Physics-Guided Feature Generation Using the Water Cloud Model

Radar backscatter from vegetated surfaces is influenced by both canopy scattering and attenuation of the signal reaching and returning from the soil surface. Consequently, Sentinel-1 observations over vegetated areas contain contributions from both vegetation and soil. The Water Cloud Model (WCM) is a semi-empirical representation of these interactions that was first developed to relate microwave backscatter to vegetation and soil conditions (Attema and Ulaby, 1978). The calibration and regional application of WCM-based approaches using Sentinel-1 observations have been investigated in recent studies, and newer studies have explored using WCM combined with machine-learning models for soil-moisture retrieval (Bai et al., 2022; Singh et al., 2023; Li et al., 2024). In contrast to approaches that use the WCM as a stand-alone soil-moisture retrieval model, the present study uses the WCM exclusively for physics-guided feature generation. The implemented formulation represents observed radar backscatter as the combined contribution of vegetation scattering and an attenuated soil-related component (Eq. 1).

$$\sigma^0_{obs} = \sigma^0_{veg} + \tau^2 \sigma^0_{soil} \quad (1)$$

where σ^0_{obs} is the observed radar backscatter, σ^0_{veg} represents the vegetation-scattering component, σ^0_{soil} represents the soil-related scattering component, and τ^2 denotes the two-way vegetation attenuation factor used in the implemented feature-generation procedure. In the conventional WCM notation, this quantity is commonly represented by τ^2. Vegetation Water Content (VWC) was introduced as the vegetation descriptor for the physics-guided feature construction. In the implemented

framework, VWC was calculated from Sentinel-2 vegetation information using a weighted combination of NDVI and EVI (Eq. 2)

$$VWC = 0.6NDVI + 0.4EVI \quad (2)$$

The resulting VWC values were constrained to non-negative values before subsequent calculations. This variable provides a compact representation of vegetation conditions and enables optical information to be incorporated into the WCM-based representation. The use of vegetation information to constrain WCM parameters and vegetation effects is consistent with recent WCM-based remote-sensing studies.

The two-way vegetation attenuation factor was calculated using the implementation-specific relationship (Eq. 3)

$$\tau = exp(-2BVWC) \quad (3)$$

where (B) is the empirical vegetation attenuation coefficient. The empirical coefficient was set to $B = 2.0$ in the implemented configuration.The resulting attenuation factor was constrained to the interval ([0,1]). This formulation represents the reduction in the soil-related radar contribution caused by signal propagation through the vegetation layer.

The vegetation-scattering feature was calculated from VWC and the attenuation factor as (Eq. 4)

$$\sigma^0_{veg} = AVWC(1-\tau) \quad (4)$$

where A denotes the empirical vegetation-scattering coefficient, which was set to A=0.1 in the implemented configuration. This variable was retained as a WCM-derived vegetation-scattering feature. It was not interpreted as an independently measured radar-scattering quantity.The soil-related feature was subsequently generated from the Sentinel-1 VV backscatter and the attenuation factor as (Eq. 5)

$$\sigma^0_{\text{soil}} = \frac{|VV|}{\tau} \quad (5)$$

where (VV) represents the Sentinel-1 VV backscatter variable used in the implementation. Because (Eq. 5) is an implementation-specific feature transformation, σ^0_{soil} is referred to here as a Because Eq. (5) represents an implementation-specific feature transformation, the resulting quantity is referred to here as a WCM-derived soil-related feature rather than as a conventionally inverted soil-backscatter coefficient. This distinction is important because the resulting quantity is used as an input feature for machine learning rather than as an independently retrieved soil-backscatter parameter.

A further physics-guided feature, denoted as WCM_SM, was generated by combining the soil-related scattering feature with NDVI (Eq. 6)

$$WCM_SM = \sigma^0_{soil}(1+NDVI) \quad (6)$$

WCM_SM is not an independently retrieved soil-moisture observation; rather, it is an engineered soil-moisture-related predictor derived from the WCM-based feature formulation. The surface soil moisture reference target for supervised learning continued to be the SMAP-derived surface soil moisture as described in Section 2.3. The reference soil-moisture variable is not replaced by WCM_SM, but rather WCM_SM is used as an additional source of information for the learning models. The WCM procedure thus created 5 physics-guided predictors VWC, τ , σ^0_{veg} , σ^0_{soil}, and

WCM_SM is shown in the Table 3. These five variables were added to the SAR, optical, climatic, terrain, soil, land-cover, and temporal predictors to construct the final predictor matrix. The enhanced predictor matrix was then provided to RF, XGBoost and FT-Transformer, from which predictions were made, and prediction level fusion was performed using the meta-learner.The purpose of the WCM in this study is therefore to introduce physics-inspired models of the vegetation and soil effects into the predictor space. Previous studies have demonstrated the feasibility of integrating physical models with machine-learning approaches to retrieve soil moisture, with WCM being part of a differentiable learning framework to retrieve soil moisture (Li et al., 2024). The WCM-derived variables, on the other hand, are considered as part of the holistic proposed framework shown in the Table 4.

Table 4

WCM-derived predictors used in the proposed framework

Variable	Description	Role in the framework
VWC	Vegetation Water Content	Represents vegetation condition
τ	Two-way vegetation attenuation factor	Represents canopy-related attenuation

Variable	Description	Role in the framework
σ^0_{veg}	Vegetation-scattering feature	Represents vegetation contribution
σ^0_{soil}	Soil-related scattering feature	Represents soil-related radar response
WCM_SM	WCM-derived soil-moisture-related feature	Provides physics-guided information to the learning models

3.4 Feature Engineering and Final Predictor Construction

After data preparation using multiple sources and generation of features using WCM, the data variables were combined to form the predictor matrix for model development. The features include SAR, optical, climate, topography, soil, land cover, temporal and WCM-derived features. The multi-source representations enable multiple soil-moisture variability controlling factors to be expressed in a single modelling context. Machine-learning-based soil-moisture estimation has also been performed on Sentinel-1, Sentinel-2 and several environmental variables, such as temporal data, in recent studies. There were three feature groups: SAR (VV, VH, VV/VH), optical (B2, B3, B4, B8, B11, B12, NDVI, EVI, SAVI, NDWI). Climatic forcing included rainfall and temperature, while the spatial environmental conditions included elevation, slope, aspect, soil texture, and LULC. To incorporate temporal information, lagged and rolling soil-moisture variables were used. The lag features were SM_LAG1, SM_LAG2, SM_LAG3, and SM_LAG6 and the rolling features were SM_ROLL3, SM_ROLL6 and SM_ROLL12. The observations were initially sorted by District, Year and Month. Lagged values were thus derived from the previous observations, antecedent conditions of soil moisture (rolling statistics) were used to represent the conditions within the time period being specified. Use of temporal information follows the trend of more recent soil-moisture studies that

371 indicate temporal constraints and antecedent conditions can be useful in surface soil-moisture
372 estimation. To avoid temporal information from the target observation from being introduced into the
373 predictors, lagged predictors and rolling predictors were created from observations prior to the
374 prediction time. This separation is crucial if time-related variables, derived from targets, are employed
375 in supervised learning. Cyclic transformations of the calendar month (Eq. 7) and (Eq. 8) were used to
376 represent annual seasonality.

$$Month_Sin = sin\left(\frac{2\pi Month}{12}\right) \quad (7)$$

$$Month_Cos = cos\left(\frac{2\pi Month}{12}\right) \quad (8)$$

377

378 A seasonal regime was also included as a Monsoon indicator. In contrast to the cyclic variables, the
379 latter do not require the artificial linear association between December and January, but rather present
380 a continuous description of the annual seasonality. Feature groups of observations, environment, time
381 and WCM-derived feature groups were combined to create the final predictor matrix. The final
382 representation was fed for all three base learners Random Forest, XGBoost, and FT-Transformer.

383 **3.5 Base Learning Models**

384 Three complementary models—Random Forest (RF), XGBoost, and FT-Transformer—were
385 independently trained using the same predictor matrix and SMAP-derived soil-moisture target. RF
386 uses a bagged ensemble of decision trees (Breiman, 2001), whereas XGBoost employs sequential
387 gradient-boosted trees with regularization (Chen & Guestrin, 2016). FT-Transformer provides an

388 attention-based learning mechanism specifically designed for tabular data (Gorishniy et al., 2021).The
389 RF prediction for observation (i) can be represented as (Eq. 9)

$$\hat{y}_i^{RF} = \frac{1}{T}\sum_{t=1}^{T} f_t(x_i) \quad (9)$$

390 where T is the number of regression trees, $f_k \ (x_i)$ denotes the prediction of the t tree, and (x_i)
391 represents the predictor vector for observation (i) (Breiman, 2001). The resulting RF predictions
392 constitute the first base-model prediction stream for the subsequent meta-learning stage.

393 Extreme Gradient Boosting (XGBoost**)** was employed as the second base learner. XGBoost is a
394 gradient-boosting framework that constructs an ensemble of decision trees sequentially, with each
395 successive learner contributing to the reduction of the prediction error (Chen and Guestrin, 2016). The
396 framework also incorporates regularization to control model complexity. In the proposed framework,
397 XGBoost was trained independently using the same predictor matrix and reference target used for RF.
398 Its prediction can be represented as (Eq. 10)

$$\hat{y}_i^{XGB} = \sum_{k=1}^{K} f_k(x_i) \quad (10)$$

399 where (K) represents the number of boosting trees and $f_k \ (x_i)$ denotes the prediction function of the
400 k-th tree (Chen and Guestrin, 2016). The resulting XGBoost predictions form the second base-model
401 prediction stream. Unlike RF, which aggregates independently constructed trees, XGBoost
402 progressively builds the ensemble by learning from the residual errors of preceding trees.

403

$$L = \sum_{i=1}^{N} l(y_{i,}, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (11)$$

In order to pay attention to the feature interactions, FT-Transformer was added as the third base learner in the structured predictor matrix. FT-Transformer offers an alternative to the tree-based RF and XGBoost models (Gorishniy et al., 2021) by using tokens to represent individual features and learning the interactions between these features in the Transformer layers. It was trained using the last set of predictors (remote-sensing, environmental, temporal and WCM-derived). The resulting prediction is the third prediction stream in the ensemble framework, and can be written as (Eq. 12). where f_{FT} denotes the FT-Transformer model, x_i represents the predictor vector, and θ denotes the learned model parameters. The resulting predictions constitute the third base-model prediction stream for the subsequent meta-learning stage as given in the (Eq. 13).

$$\hat{y}_i^{FT} = f_{FT}(x_i; \theta) \quad (12)$$

$$\hat{y}_i^{RF}, \hat{y}_i^{XGB}, \hat{y}_i^{FT} \quad (13)$$

The three base learners independently process the same final predictor matrix. Their outputs can therefore be represented as (Eq. 13) Note that the final predictor matrix is the same for the three models, which train an independent relationship between the prediction of the final predictor matrix and the SMAP derived soil-moisture target without requiring the predictions of other base learners as input. They are kept as different streams and only amalgamated during the next metalearning step. In line with this, the framework proposed in this study uses a two-level stacked learning scheme, including

the Level-0 learners of RF, XGBoost, and FT-Transformer, and the Level-1 learner of Ridge Regression. Stacking combines base-model predictions in a second learning phase which learns their relationship with the target variable (Wolpert, 1992). The predictions at level-0 are indicated by the following equation (Eq. 14)

$$z = [\hat{y}_i^{RF}, \hat{y}_i^{XGB}, \hat{y}_i^{FT}] \quad (14)$$

Here, each row of z represents an observation, the columns are the prediction of RF, XGBoost and FT-Transformer respectively. Ridge Regression was used as the Level-1 learner to combine these prediction streams. Ridge Regression works well with correlated predictors (Hoerl and Kennard, 1970), because it adds an (L_2) regularization term to the least-squares objective. The base models share the same predictor space, so their predictions may be redundant; Ridge regularization offers a robust way of estimating the contributions.The final hybrid prediction is expressed as (Eq. 15). where β_0 is the intercept and β_1, β_2 and β_3 are the coefficients estimated by the Ridge meta-learner. The corresponding regularized objective is (Eq. 16)

$$\hat{y}^{Hybrid} = \beta_0 + \beta_1 \hat{y}^{RF} + \beta_2 \hat{y}^{XGB} + \beta_3 \hat{y}^{FT}, \quad (15)$$

$$\min_{\beta} \left[\left(\sum_{i=1}^{N} y_i - \beta_0 - \sum_{j=1}^{3} \beta_j Z_{ij} \right)^2 + \lambda \sum_{j=1}^{3} \beta_j{}^2 \right] \quad (16)$$

where y_i denotes the reference soil moisture for observation (i), Z_{ij} denotes the prediction of base learner (j), β_j represents the corresponding meta-model coefficient, and λ is the regularization parameter. The (L_2) penalty constrains the coefficient magnitude and can reduce instability when the

prediction streams are correlated (Hoerl and Kennard, 1970). The proposed framework performs fusion at the prediction level**.** RF, XGBoost, and FT-Transformer therefore remain independent at Level 0 and do not receive the predictions of the other base learners as inputs. Only their predicted soil-moisture values are transferred to the Ridge Regression layer. The complete prediction pathway is represented as (Eq. 17)

$$X \rightarrow \begin{cases} RF \rightarrow \hat{y}^{RF} \\ XGBoost \rightarrow \hat{y}^{XGB} \\ FT-Transformer \rightarrow \hat{y}^{FT} \end{cases} \rightarrow Z \rightarrow Ridge \rightarrow \hat{y}^{Hybrid} \quad (17)$$

where X represents the final predictor matrix and $\hat{y}^{Hybrid}$ The final soil-moisture prediction is the last soil-moisture prediction. Stacking is different from simple averaging as it enables the meta-learner to obtain an estimate of the contribution of each of the base-model predictions from the training data. In order to prevent information leakage, the out-of-fold (OOF) prediction values were used as the input values for Level-1 models and each of the meta-features was produced by a model trained using a different observation that was never used to train the corresponding meta-feature. This is to avoid over-optimistic predictions of meta-training and to get a more accurate approximation of the generalization performance. The final framework is thus the combination of the three complementary learning mechanisms: bagged trees (RF), gradient-boosted trees (XGBoost), and attention-based tabular learning (FT-Transformer) using Ridge-based prediction fusion.

Table 5

Predictive Performance of the Individual Learners and Proposed Ridge Stacking.

Model	RF Final	XGBoost	FT Transformer	Ridge Stacking
RMSE	0.091087	0.077863	0.146708	0.059135
MAE	0.052081	0.052256	0.086054	0.035440
MSE	0.008297	0.006063	0.021523	0.003497
R^2	0.991442	0.993747	0.977799	0.996393
NSE	0.991442	0.993747	0.977799	0.996393
R	0.995848	0.996984	0.996803	0.998241

4. Result and Discussion

4.1 Comparative Prediction Performance

The predictive performance of RF, XGBoost, FT-Transformer, and the proposed Hybrid framework was evaluated using RMSE, MAE, MSE, R^2, NSE, and R is shown in the Table 5. XGBoost achieved the best performance among the individual models, with an RMSE of 0.0779 and R^2 of 0.9937. The proposed Hybrid framework performed best overall, achieving RMSE = 0.0591, MAE = 0.0354, R^2 = 0.9964, NSE = 0.9964, and R = 0.9982. Compared with XGBoost, the Hybrid framework reduced

RMSE by approximately 24.1% and MSE by 42.3%. These improvements demonstrate the benefit of combining complementary predictions from RF, XGBoost, and FT-Transformer through Ridge-based stacking. Fig. 3. Compares the predicted and reference soil-moisture values with respect to the 1:1 relationship. The agreement between all models and observations is good. XGBoost shows a lower RMSE when compared to the standalone, higher alignment compared to RF and FT-Transformer, as shown near the slope of 1:1, which aligns with its lower standalone RMSE.FT-Transformer shows more variations from the 1:1 line when compared to XGBoost and RF, which align with its higher standalone RMSE. Hybrid model exhibits the best agreement and least spread distribution among all the models analyzed, with lowest RMSE (0.0591) and highest R^2 (0.9964). The compact distribution shows the gains in predictive consistency that can be obtained by prediction-level stacking. The residual distributions are shown in Fig. 4. The residuals of RF and XGBoost are more centralised and FT-Transformer shows more spread. The residual distribution for the Hybrid model is the narrowest, meaning that the error dispersion is the lowest and the predictions are the most consistent. This result is also confirmed by its RMSE value (0.0591) and its MAE value (0.0354). The absolute prediction errors are displayed in cumulative distributions in Fig. 5. The Hybrid model has higher cumulative proportions at low error levels than the individual models, suggesting higher percentage of low error predictions. The accumulations are comparatively slow for FT-Transformer, whereas the accumulations are intermediate for RF and XGBoost. Summarizing overall, the three diagnostic analyses all show more accurate predictions and more consistent errors for the proposed Hybrid model.

Fig. 6. Compares the models by means of correlation, standard deviation, and centred root-mean-square difference. The highest overall agreement is obtained from the Hybrid model, with the highest

correlation (R=0.9982) with the reference observations. The other models, XGBoost and FT-Transformer, also have good correlations, 0.9970 and 0.9968, respectively, while RF has 0.9958. Hence, the Hybrid model is the closest statistical approximation to the observed soil-moisture variability. The spatial distribution of the reference soil-moisture observations over the study area is shown in Fig. 7. These observations are spatially variable, suggesting that soil moisture conditions vary at different locations. This spatial variability offers valuable reference to understand the modelization of soil-moisture variations with the proposed framework including multiple data sources. This spatial distribution is also indicative of the different environmental conditions included in the data set, reinforcing the need for nonlinear and complementary learning strategies. The Fig. 7 however, is only the distribution of reference observations and cannot per se be used to define spatial generalization at unseen locations. The comparative results in Table 5 demonstrate that the proposed Hybrid framework consistently outperformed the individual RF, XGBoost, and FT-Transformer models across all evaluation metrics. It achieved the lowest RMSE (0.0591), MAE (0.0354), and MSE (0.0035), along with the highest R^2 (0.9964), NSE (0.9964), and R (0.9982). Compared with XGBoost, the strongest individual model, the Hybrid framework reduced RMSE by approximately 24.1%. Overall, Table 5 confirms the effectiveness of prediction-level stacking for improving soil-moisture prediction accuracy and consistency.

4.2 Discussion

On their own, FT-Transformer has more prediction error than XGBoost, but that doesn't mean it is not suited for the job. FT-Transformer, with attention-based feature interactions for tabular learning (Chen & Guestrin, 2016; Gorishniy et al., 2021), is suitable for nonlinear relationships in structured tabular data while XGBoost is suitable for tabular learning. Although its RMSE was higher, FT-

Transformer showed a good correlation R = 0.9968, meaning that there was good agreement between the FT-Transformer and the reference observations. The goal of the proposed framework is not to find the best learner, but to leverage the complementary prediction patterns of RF, XGBoost and FT-Transformer. Over the past few years, the benefits of stacking heterogeneous learners to enhance the prediction accuracy has also been shown in soil-moisture research (Tahmouresi et al., 2024).

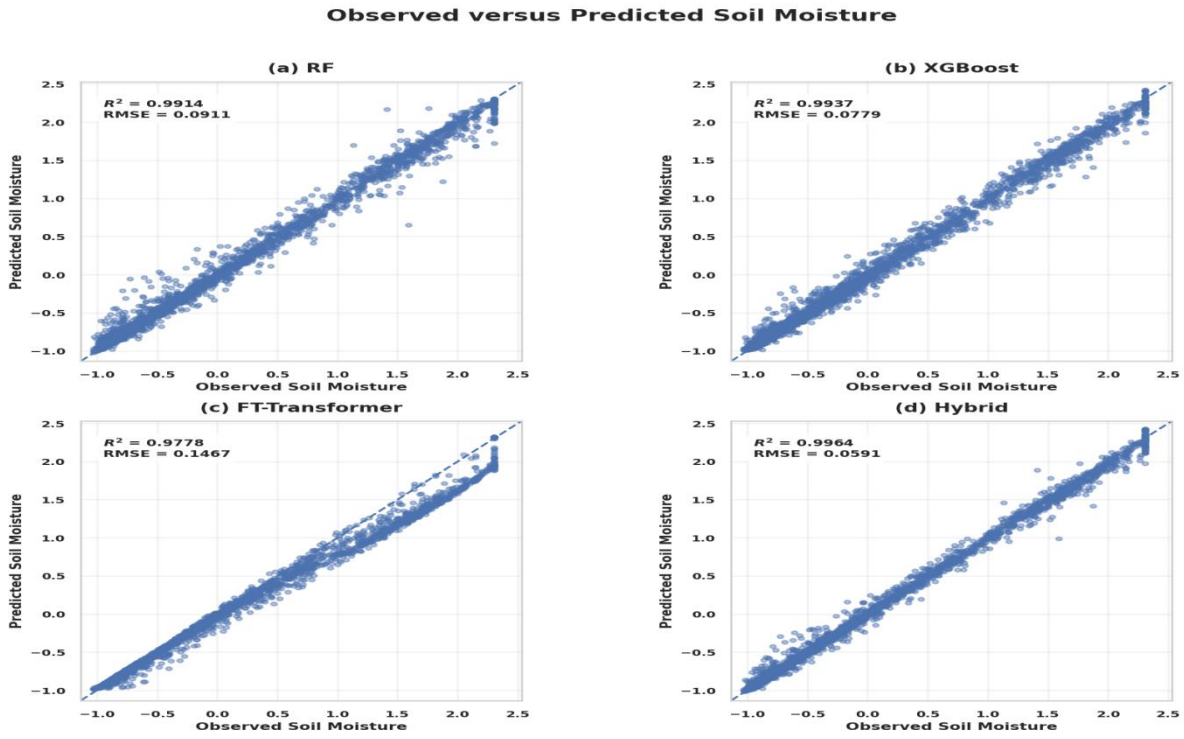
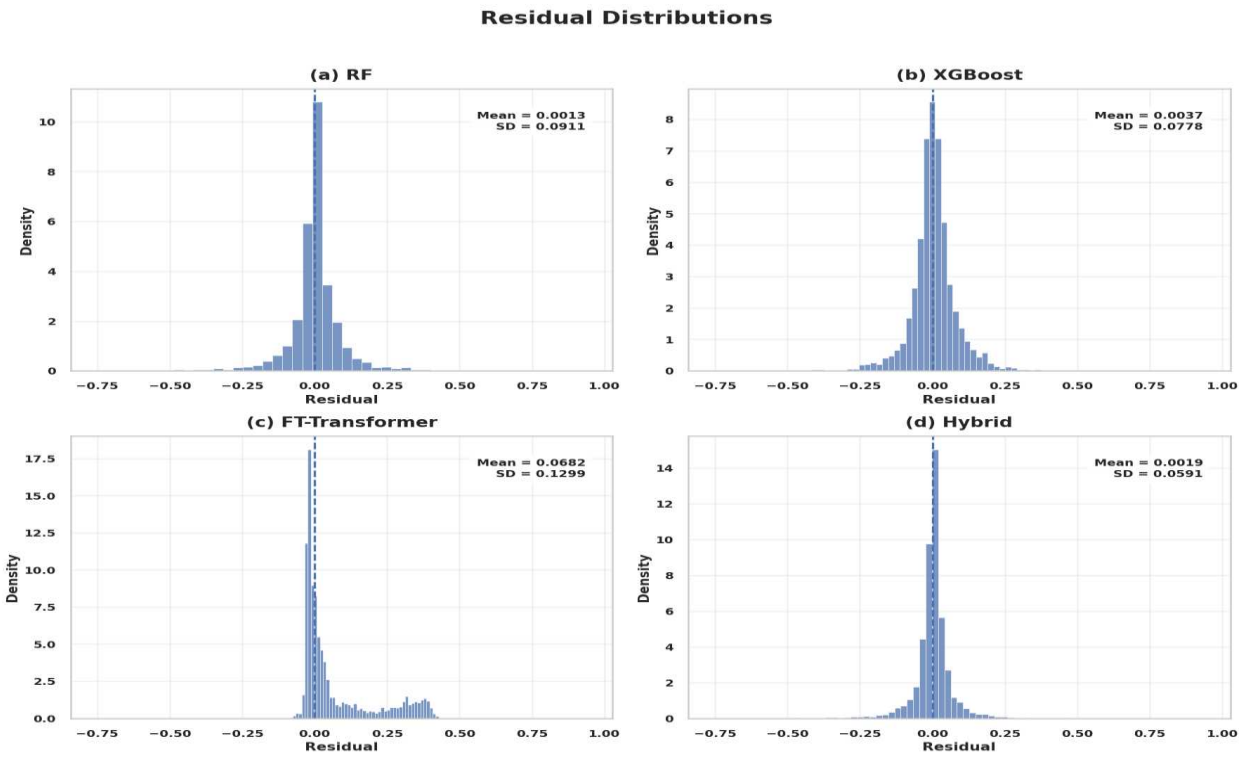


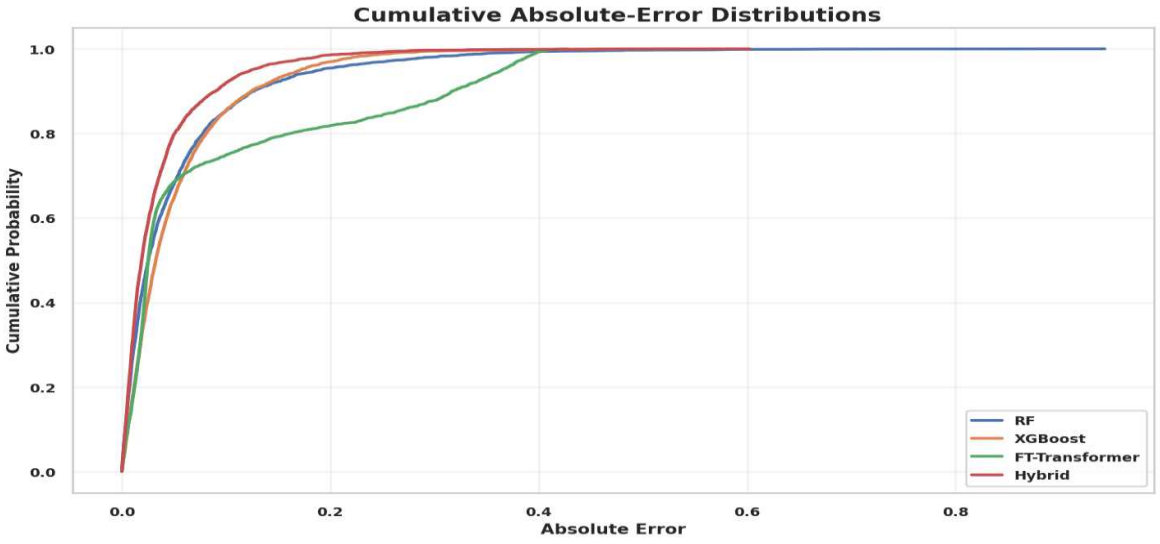
Fig. 3. Observed versus predicted soil-moisture values



517

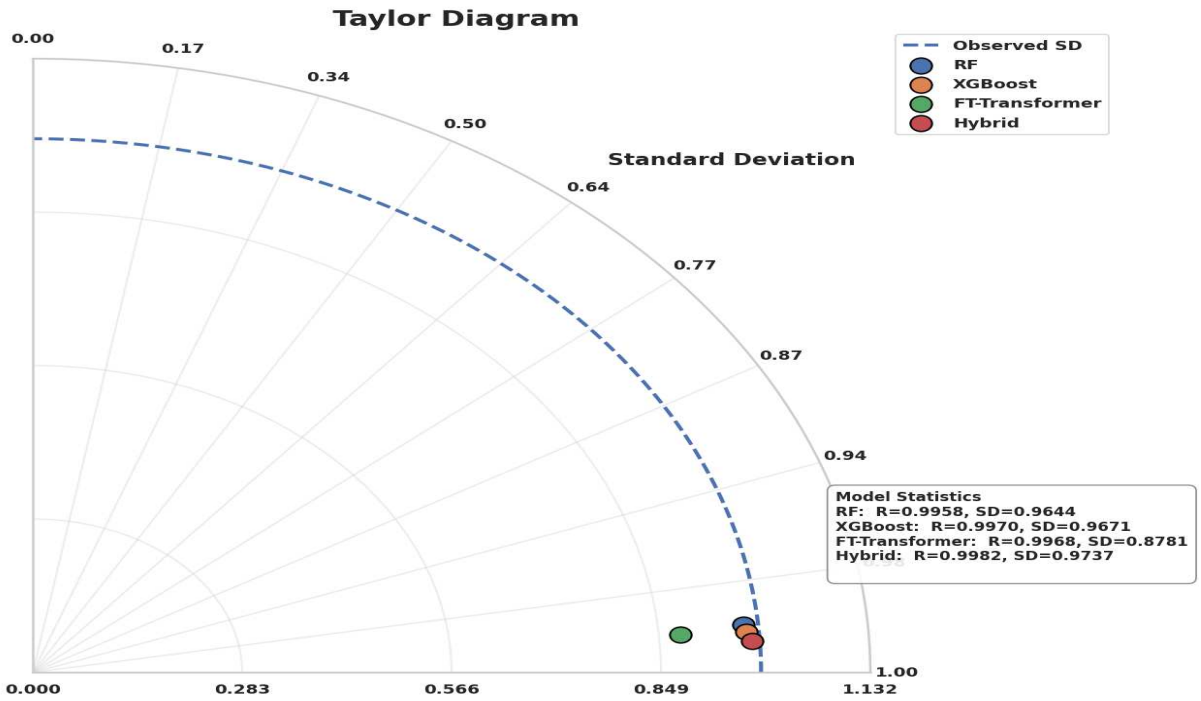
518 **Fig. 4. Residual Distributions of The Evaluated Soil-Moisture Prediction Models**

519



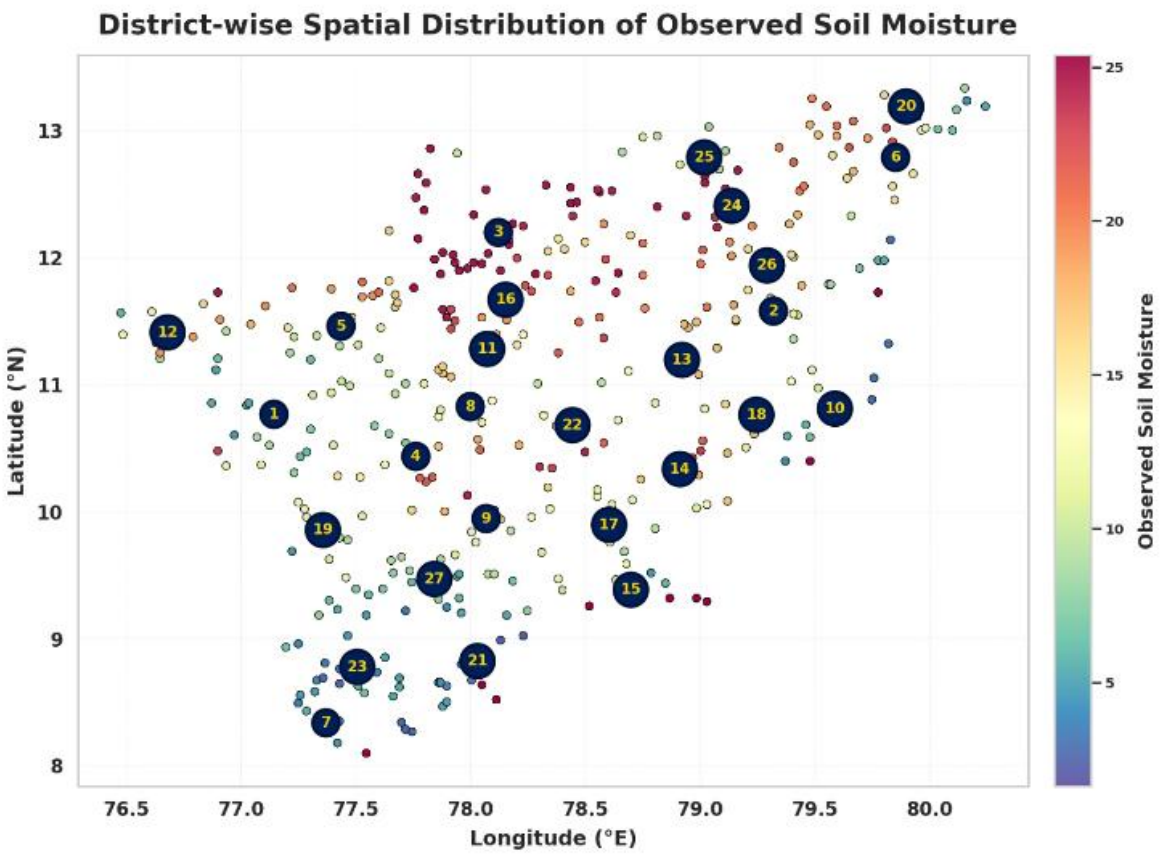
520

521 **Fig. 5. Cumulative Distributions Of Absolute Prediction Errors**



522

523 **Fig. 6. Taylor Diagram Comparing the Statistical Performance of Models**



District No. → District Name

1 = Ariyalur	2 = Coimbatore	3 = Nagapattinam	4 = Namakkal
5 = The Nilgiris	6 = Perambalur	7 = Pudukkottai	8 = Ramanathapuram
9 = Salem	10 = Sivaganga	11 = Tiruvarur	12 = Theni
13 = Viluppuram	14 = Tiruvallur	15 = Thoothukudi	16 = Tiruchirappalli
17 = Tirunelveli	18 = Vellore	19 = Krishnagiri	20 = Tiruvannamalai
21 = Virudhunagar	22 = Dharmapuri	23 = Dindigul	24 = Erode
25 = Kancheepuram	26 = Kanniyakumari	27 = Karur	28 = Madurai

Fig. 7. District - Wise Spatial Distribution Observed Soil Moisture

The features informed by WCM are a physics-based representation of the vegetation and soil contributions to the Sentinel-1 backscatter. Previous studies validated the impact of vegetation-soil interactions on SAR-based soil moisture retrieval, and including these interactions can enhance the accuracy of the soil moisture retrievals (Singh et al., 2023). The optical, environmental, and

topographic predictors are complemented by features derived from the WCM in the proposed framework. Their contribution, therefore, is to deliver feature representations of the physics-informed context and not to deliver independent soil-moisture retrieval. The integration is used for the synergistic use of physical knowledge and data-driven learning in soil moisture prediction. The proposed Hybrid framework obtained an RMSE of 0.0591 and an R^2 of 0.9964, which shows that it has high predictive performance. Recently, the stacking methods for soil moisture estimation have also been reported to be effective. A stacking ensemble approach for satellite SMD was developed by Tahmouresi et al. (2024) which achieved an R^2 of 0.97 and an RMSE of 0.0123 m3/m3. (Li and Yan et al., 2024) concluded that the stacking model further enhanced the accuracy and robustness of soil-moisture estimation compared to individual machine-learning models. The results confirm the effectiveness of model integration and the present study applies a similar approach to the integration of RF, XGBoost, and FT-Transformer with Ridge-based prediction-level stacking. The proposed Hybrid framework shows the feasibility of combining multi-source remote-sensing observations with physics-informed and data-driven features to predict soil moisture in regions. In recent studies, the effectiveness of integrating remote sensing data with environmental variables and ensemble or deep-learning has also been shown in soil-moisture estimation (Li et al., 2024; Xu et al., 2025). The framework could include applications for soil-moisture monitoring and agricultural water management in the region. Yet, independent spatial and temporal validation is needed prior to operation, due to the relevance of model transferability to new locations and time periods (Chrysanthopoulos & Kallioras, 2025).

5. Conclusion

This study proposed the Hybrid framework for soil-moisture prediction combining WCM-informed features with the RF, XGBoost and FT-Transformer by Ridge-based prediction-level stacking. The proposed framework achieved the best performance, with an RMSE of 0.0591, MAE of 0.0354, R^2 of 0.9964, NSE of 0.9964, and R of 0.9982. The Hybrid framework achieved a 24.1% and 42.3% decrease in RMSE and MSE, respectively, compared to the best of the individual learners, which is XGBoost. The results show that heterogeneous learning models can be complementary when applied as stacking for prediction information. Radar–vegetation interactions are further represented in the WCM-informed features as a physics-guided representation, matching previous Sentinel-1 soil-moisture studies. In general, the proposed framework is a promising method for soil-moisture forecasting in the region based on multi-source observations. Independent spatial and temporal validation is required to validate it independently in space and time, as well as test it in other regions, in future studies. The proposed framework will be expanded to wider spatiotemporal application through testing for implementation in various spatial regions, climatic zones, vegetation types and soil properties in the future. Co-combining more SAR and optical observations of high resolution will further improve the surface and environmental dynamic representation. Contributions of soil-moisture features from WCM and antecedent will be quantified via systematic ablation studies. Research will also be conducted to further develop uncertainty aware and advanced physics guided learning approaches to enhance the reliability of prediction. These extensions can improve the transferability of the framework and make it usable in all three areas of regional soil-moisture monitoring, agricultural water management and environmental decision-making

Declarations:

Data Availability: Data and code are available via the following link:

GitHub : https://github.com/pyingkodi/Soil-Mositure-for-Tamil-Nadu

Zenodo : https://zenodo.org/uploads/23983041

Conflict of Interest: No conflict of Interest

Ethical Approval: Ethical approval not required for this study.

Funding: This research received no specific grant from any funding agency.

Clinical Trial Registration: Not applicable. This study does not involve a clinical trial.

Consent to Publish declaration: Not applicable.

Consent to Participate declaration: Not applicable.

References

Ann-Marie Fortuna, Patrick J. Starks & Daniel N. Moriasi .(2025). Estimation of soil organic carbon as a function of soil pretreatment and spectral features of radiometers within the visible and near-infrared spectra,Journal of Soil and Water Conservation, Pages 476-490

Attema, E. P. W., & Ulaby, F. T. (1978). Vegetation is modeled as a water cloud. *Radio Science, 13*(2), 357–364.

590 Babaeian, E., Sadeghi, M., Jones, S. B., Montzka, C., Vereecken, H., & Tuller, M. (2019). Ground,
591 proximal, and satellite remote sensing of soil moisture. *Reviews of Geophysics, 57*(2), 530–616.

592 Bai, X., He, B., Li, X., Zeng, J., Wang, X., Wang, Z., Zeng, Y., & Su, Z. (2017). First assessment of
593 Sentinel-1A data for surface soil moisture estimations using a coupled Water Cloud Model and
594 advanced integral equation model over the Tibetan Plateau. *Remote Sensing, 9*(7), 714.

595 Bai, X., Zheng, D., Liu, X., Fan, L., Zeng, J., & Li, X. (2022). Simulation of Sentinel-1A observations
596 and constraint of water cloud model at the regional scale using a discrete scattering model. *Remote*
597 *Sensing of Environment, 283*, 113308. https://doi.org/10.1016/j.rse.2022.113308

598 Batchu, V., Nearing, G., & Gulshan, V. (2023). A deep learning data fusion model using Sentinel-1/2,
599 SoilGrids, SMAP, and GLDAS for soil moisture retrieval. *Journal of Hydrometeorology, 24*(10),
600 1789–1823. https://doi.org/10.1175/JHM-D-22-0118.1

601 Bauer-Marschallinger, B., Paulik, C., Hochstöger, S., Mistelbauer, T., Modanesi, S., Ciabatta, L.,
602 Massari, C., Brocca, L., & Wagner, W. (2018). Soil moisture from fusion of scatterometer and SAR:
603 Closing the scale gap with temporal filtering. *Remote Sensing, 10*(7), 1030.
604 https://doi.org/10.3390/rs10071030

605 Benninga, H. F. J., van der Velde, R., & Su, Z. (2022). Soil moisture content retrieval over meadows
606 from Sentinel-1 and Sentinel-2 data using physically based scattering models. *Remote Sensing of*
607 *Environment, 280*, 113191.

Bousbih, S., Zribi, M., El Hajj, M., Baghdadi, N., Lili-Chabaane, Z., Gao, Q., & Fanise, P. (2018). Soil moisture and irrigation mapping in a semi-arid region, based on the synergetic use of Sentinel-1 and Sentinel-2 data. *Remote Sensing, 10*(12), 1953.

Bramley, R., Perry, E., Richetti, J., Colaço, A. F., Mowat, D., Ratcliff, C., & Lawes, R. (2024). Within-field extrapolation away from a soil moisture probe using freely available satellite imagery and weather data. *Precision Agriculture*, 25(4), 1877–1893.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Carranza, C., Nolet, C., Pezij, M., & van der Ploeg, M. (2021). Root zone soil moisture estimation with Random Forest. *Journal of Hydrology, 593*, 125840.

Chaudhary, S. K., Srivastava, P. K., Gupta, D. K., Kumar, P., Prasad, R., Pandey, D. K., Das, A. K., & Gupta, M. (2022). Machine learning algorithms for soil moisture estimation using Sentinel-1: Model development and implementation. *Advances in Space Research, 69*(4), 1799–1812. https://doi.org/10.1016/j.asr.2021.08.022

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). Association for Computing Machinery.

Chrysanthopoulos, E., & Kallioras, A. (2025). Temporal and geographic extrapolation of soil moisture using machine learning algorithms. *Catena, 257*, 109156. https://doi.org/10.1016/j.catena.2025.109156

628 Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C.,
629 Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., & Bargellini, P. (2012).
630 Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of*
631 *Environment, 120*, 25–36.

632 Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M.,
633 Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M.,
634 Burbank, D., & Alsdorf, D. (2007). The Shuttle Radar Topography Mission. *Reviews of Geophysics,*
635 *45*(2), RG2004.

636 Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J.,
637 Harrison, L., Hoell, A., & Michaelsen, J. (2015). The climate hazards infrared precipitation with
638 stations—A new environmental record for monitoring extremes. *Scientific Data, 2*, 150066.

639 Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth
640 Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment, 202*, 18–27.

641 Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for
642 tabular data. In *Advances in Neural Information Processing Systems, 34*, 18932–18943.

643 Hengl, T. (2018). *Soil texture classes (USDA system) for 6 soil depths (0, 10, 30, 60, 100 and 200 cm)*
644 *at 250 m* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.2525817

645 Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal
646 problems. *Technometrics, 12*(1), 55–67.

Inoubli, R., Bennaceur Farah, L., Jarray, N., Ben Abbes, A., & Farah, I. R. (2022). A comparison between the use of machine learning techniques and the Water Cloud Model for the retrieval of soil moisture from Sentinel-1A and Sentinel-2A products. *Remote Sensing Letters, 13*(10), 980–990. https://doi.org/10.1080/2150704X.2022.2115862

Li, M., & Yan, Y. (2024). Comparative analysis of machine-learning models for soil moisture estimation using high-resolution remote-sensing data. *Land, 13*(8), 1331.

Li, Z., Yuan, Q., Yang, Q., Li, J., & Zhao, T. (2024). Differentiable modeling for soil moisture retrieval by unifying deep neural networks and water cloud model. *Remote Sensing of Environment, 311*, 114281. https://doi.org/10.1016/j.rse.2024.114281

Mladenova, I. E., Bolten, J. D., Crow, W., Sazib, N., & Reynolds, C. (2020). Agricultural drought monitoring via the assimilation of SMAP soil moisture retrievals into a global soil water balance model. *Frontiers in Big Data, 3*, 10. https://doi.org/10.3389/fdata.2020.00010

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., & Thépaut, J.-N. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data, 13*, 4349–4383.

Singh, S. K., Prasad, R., Srivastava, P. K., Yadav, S. A., Yadav, V. P., & Sharma, J. (2023). Incorporation of first-order backscattered power in Water Cloud Model for improving the Leaf Area Index and Soil Moisture retrieval using dual-polarized Sentinel-1 SAR data. *Remote Sensing of Environment, 296*, 113756. https://doi.org/10.1016/j.rse.2023.113756

667 Tahmouresi, M. S., Niksokhan, M. H., & Ehsani, A. H. (2024). Enhancing spatial resolution of
668 satellite soil moisture data through stacking ensemble learning techniques. *Scientific Reports, 14*,
669 25454.

670 Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B.,
671 Floury, N., Brown, M., Traver, I. N., Deghaye, B., Duesmann, B., Rosich, B., Miranda, N., Bruno, C.,
672 L'Abbate, M., Croci, R., Huchler, M., & Rostan, F. (2012). GMES Sentinel-1 mission. *Remote Sensing*
673 *of Environment, 120*, 9–24.

674 van den Hoogen, J., Hudson, D., & Atzmueller, M. (2026). A comparison of graph construction
675 techniques for applying graph signal processing to soil moisture networks. Discover Computing, 29,
676 Article 17.

677 Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259.
678 https://doi.org/10.1016/S0893-6080(05)80023-1

679 Xu, Y., Cai, S., Huang, J., Liu, J., Shang, J., Yang, Z., & Zhang, Z. (2025). A multimodal deep
680 learning approach for soil moisture downscaling using remote sensing and weather data. *Journal of*
681 *Geophysical Research: Machine Learning and Computation, 2*(3), e2025JH000639.
682 https://doi.org/10.1029/2025JH000639

683 Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G.,
684 Wevers, J., Cartus, O., Santoro, M., Fritz, S., Lesiv, M., Herold, M., Tsendbazar, N.-E., Xu, P.,
685 Ramoino, F., & Arino, O. (2022). *ESA WorldCover 10 m 2021 v200* [Data set]. Zenodo.
686 https://doi.org/10.5281/zenodo.7254221

687 Zhou, Y., Wang, B., Zhu, W., Feng, L., He, Q., Zhang, X., Wu, T., & Yan, N. (2024). Spatial-
688 temporal constraints for surface soil moisture mapping using Sentinel-1 and Sentinel-2 data over
689 agricultural regions. *Computers and Electronics in Agriculture, 219*, 108835.
690 https://doi.org/10.1016/j.compag.2024.108835