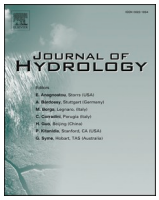
Journal of Hydrology 625 (2023) 130010

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol



Research papers

# Soil moisture at 30 m from multiple satellite datasets fused by random forest

Jing Ning [a], Yunjun Yao [a,*], Qingxin Tang [b], Yufu Li [c,*], Joshua B. Fisher [d], Xiaotong Zhang [a], Kun Jia [a], Jia Xu [e], Ke Shang [f], Junming Yang [a], Ruiyang Yu [a], Lu Liu [a], Xueyi Zhang [a], Zijing Xie [a], Jiahui Fan [a]

[a] State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China
[b] School of Geography and Environment, Liaocheng University, Liaocheng 252000, China
[c] Jincheng Meteorological Administration, Jincheng 048026, China
[d] Schmid College of Science and Technology, Chapman University, University Drive, Orange, CA 92866, USA
[e] Department of Infrastructure Engineering, Faculty of Engineering & IT, University of Melbourne, Melbourne, VIC 3010, Australia
[f] School of Space Information, Space Engineering University, Beijing 101416, China

ARTICLE INFO

This manuscript was handled by Emmanouil Anagnostou, Editor-in-Chief

*Keywords:*
Soil moisture
High spatial resolution
Machine learning
Agricultural drought

ABSTRACT

As a key variable used to characterize the climate process between the land surface and atmosphere, the surface soil moisture (SM) plays an irreplaceable role in the fields of hydrology, meteorology and agriculture. However, the relatively coarse spatial resolution of SM products currently limit the application in water resource management at the field scale. In this study, we proposed a high-spatial-resolution SM retrieval framework based on the random forest algorithm (RF-SM) to integrate in situ SM dataset from in the International Soil Moisture Network (ISMN), Landsat 8 optical and thermal observations, soil properties from SoilGrids V2.0, meteorological variables from the fifth generation of the European ReAnalysis (ERA5) dataset and four coarse-scale SM products including the Soil Moisture Active/Passive (SMAP), the Soil Moisture and Ocean Salinity (SMOS), European Space Agency Climate Change Initiative (ESA CCI) and U.S. National Climate Assessment Land Data Assimilation System (NCA-LDAS). Compared to three other machine learning (ML) algorithms [extremely randomized trees (Extra-Trees), gradient boosting regression tree (GBRT) and extreme gradient boosting (XGBoost)], the random forest (RF) algorithm exhibited the best performance against a subset of 100 validation sites with a Kling–Gupta efficiency (KGE) of 0.69 and root-mean-square-error (RMSE) of 0.063 $m^3/m^3$. In terms of different land cover types and typical sites, RF-SM also showed a better accuracy than any of the individual SM product. Finally, the retrieval framework was applied to map the 30-m resolution SM spatial distributions in five substudy areas in the U.S. The results suggest that it is feasible to retrieve accurate SM at a 30-m spatial resolution from multiple satellite datasets based on the RF algorithm, which has important practical significance for agricultural drought monitoring at the field scale.

## 1. Introduction

Soil moisture (SM), at the heart of biological, physical and chemical processes on the land surface (Dorigo et al., 2017), drives complex climate processes and feedback loops between the ground and the atmosphere by affecting the regulation of water and the distribution of energy fluxes (Peng et al., 2017; Seneviratne et al., 2010; van der Molen et al., 2011). SM has become a crucial variable in hydrology (Brocca et al., 2017; Purdy et al., 2018; Sadeghi et al., 2020), meteorology (Alizadeh and Nikoo 2018; Baldocchi et al., 2022; Berg and Sheffield 2018) and agricultural water management (Holzman et al., 2018; Parida et al., 2008; Rahman et al., 2017). Traditional SM observations rely on ground measurement equipped with professional instruments and mature technology (Garrido et al., 1999; Topp and Reynolds 1998; Valente et al., 2006; Zhou et al., 2001), which are the most accurate data sources for surface model calibrations and satellite-based SM retrievals (Robock et al., 2000). However, due to the limitations associated with the number and cost of monitoring equipment, it is difficult to provide

* Corresponding authors.
*E-mail addresses:* boyyunjun@163.com (Y. Yao), qxtlyf@163.com (Y. Li).

https://doi.org/10.1016/j.jhydrol.2023.130010
Received 9 May 2023; Received in revised form 16 July 2023; Accepted 23 July 2023
Available online 9 August 2023

SM information in remote areas and over complex ground surfaces, and point-based observation databases cannot dynamically reflect spatial heterogeneities at relatively large scales (Peng et al., 2017). Therefore, there is an urgent demand for a scientifically feasible framework for accurately retrieving and mapping SM at relatively high spatial resolution that can be applied to agricultural drought monitoring and water resource management at the field scale (Martínez-Fernández et al., 2016; Seneviratne et al., 2010).

Remote sensing has solved the observation bottleneck at the point scale to obtain spatiotemporally continuous SM data from the field scale to the regional and global scales (Mohanty et al., 2017). Optical and thermal infrared remote sensing can reflect rich spectral and thermal information about the land surface at a high spatial resolution (Verrelst et al., 2015). Various drought indices (Ghulam et al., 2007; Wang et al., 2007) and thermal inertia models (Kang et al., 2017; Lu et al., 2018) have been proposed for SM retrieval. However, these methods have limited ranges of applicability due to cloud-contaminated images and the specificity of the surface conditions and humidity background (Attila et al., 2014; Price 1985). Microwave remote sensing is most suitable for generating operational all-weather SM products. To date, global SM datasets based on passive microwave sensors have been produced, such as the Advanced Microwave Scanning Radiometer-Earth Observing System (AMSR-E) (Njoku et al., 2003), the Soil Moisture and Ocean Salinity (SMOS) (Kerr et al., 2016) and the Soil Moisture Active/Passive (SMAP) products (Chan et al., 2018). However, the spatial resolution of these products is relatively coarse (25–50 km) and cannot meet the requirements of research and application at the field or regional scales (several kilometers or even meters) (Li et al., 2021; Peng et al., 2017). In addition, numerical simulations and data assimilations can also provide spatiotemporally continuous SM data. Land surface data assimilation systems, such as the Global Land Data Assimilation System (GLDAS) (Rodell et al., 2004) and the U.S. National Climate Assessment Land Data Assimilation System (NCA-LDAS) (Kumar et al., 2019), combine SM from satellite observations with the information from land surface process models and their associated meteorological forcing to improve the estimation accuracy of satellite data or a single model (Reichle and Koster 2005; Reichle et al., 2004). Although data assimilation methods can reduce coarse-scale observations to several to tens of kilometers, the quality of the forcing datasets and targeted bias correction methods must be considered under the assimilation framework (Lievens et al., 2015; Peng et al., 2017).

In agricultural monitoring and water resource management at the field and regional scales, the requirements for the quality and applicability of SM products are relatively high. Therefore, numerous studies have focused on the disaggregation or downscaling of coarse-spatial-resolution SM data (Sabaghy et al., 2018). Based on the studies of predecessors, the following three groups can be roughly delineated: (1) satellite product-based methods, (2) surface property-based methods, and (3) data assimilation-based methods (Peng et al., 2017; Sabaghy et al., 2018). Satellite product-based methods take full advantage of high-spatial-resolution remote sensing datasets to estimate downscaled SM. In terms of microwave remote sensing, exploiting the strength of active microwave sensors is the first choice for downscaling passive microwave remote sensing products. For example, Tomer et al. (2016) proposed the Merge Active and Passive microwave Soil Moisture (MAPSM) algorithm by combining RADARSAT-2 with high-spatial-resolution data and SMOS with high-temporal-resolution data to represent the direction and amplitude of drying and wetting changes in SM and accurately estimate SM at a spatial resolution of 500 m. In contrast, optical/thermal remote sensors have fine spatial resolutions and many studies have used surface parameters such as vegetation variables and the land surface temperature (LST) to downscale microwave-based SM (Colliander et al., 2017). For instance, Chauhan et al. (2003) downscaled a product at a 25-km resolution to 1 km by relating low-resolution SM from microwave radiometers to LST, albedo and vegetation index values in a “universal triangle” approach (Carlson et al., 1994). Similarly Merlin et al. (2008) proposed a physics-based downscaling algorithm based on the linear relationship between the soil moisture index and surface SM to capture the spatial variations in SM in the top 0–5 cm at a fine scale (Merlin et al., 2009, 2010; Merlin et al., 2005).

Surface property-based methods often take into account the effects of attributes such as the soil texture and vegetation coverage on the spatial structure of SM in downscaling strategies. For example, Ranney et al. (2015) considered the spatial variation characteristics of vegetation and soil properties based on the Equilibrium Moisture from Topography (EMT) model (Coleman and Niemann 2013), and improved the SM downscaling ability of the model by inputting fine-resolution auxiliary datasets. Subsequently, Guevara and Vargas (2019) provided a prediction approach for SM spatial patterns relying on geomorphology-derived terrain parameters to improve the spatial resolution of SM data derived from passive satellite observations with a coarse grid to a 1-km resolution. Surface hydrological models form the core of data assimilation-based methods and are also an important basis for high-resolution SM estimations (Reichle et al., 2001). For instance, Sahoo et al. (2013) used three-dimensional ensemble Kalman filter technology to assimilate the AMSR-E coarse-resolution product into the Noah land surface model and estimated fine-scale (1-km) SM at the watershed scale. Additionally, Kornelsen et al. (2015) developed an SM-downscaling technique based on bias correction in data assimilation systems and successfully applied the technique to satellite observations at subpixel scales.

Recently, machine learning (ML) methods, capable of characterizing complex relationships between multiple input features and target variables, have been widely used to downscale SM from satellite products (Ali et al., 2015; Peng et al., 2021). For instance, Im et al. (2016) examined the abilities of three ML algorithms including the random forest, boosted regression tree and Cubist to downscale of AMSR-E SM data by using Moderate Resolution Imaging Spectroradiometer (MODIS) products (surface albedo, land surface temperature, vegetation indices and evapotranspiration data at 1-km resolution) and found that the downscaling SM with 1 km resolution showed a higher correlation with ground measurements than the original AMSR-E data. Similarly, Zhao et al. (2018) proposed a SMAP SM- downscaling study based on the random forest (RF) algorithm by integrating optical/thermal infrared observation products from MODIS and topographic data, which presented high spatial heterogeneities and detailed temporal patterns at a 1-km spatial resolution. However, water resource management at the field scale requires more detailed SM data (Vergopolan et al., 2020). To retrieve SM data with a high precision and spatial resolution, an increasing amount of auxiliary information and refined surface parameter datasets have been incorporated. Recently, Zhang et al. (2022) developed a high-resolution (30-m) SM retrieval framework that integrates band reflectance from Landsat and SMAP products, the fifth generation of the European ReAnalysis (ERA5) dataset, and ancillary datasets including topography, precipitation and soil properties. However, different SM products have their own strengths due to differences in the retrieval and estimation algorithms (Al-Yaari et al., 2019; Cui et al., 2017; Zhang et al., 2021). The performance of SM retrieval is limited by data quality and inherent uncertainties when a single product is introduced as an average moisture conditions over large areas. Moreover, the accurate retrieval of SM state requires consideration of various meteorological stresses in the process of land–atmosphere interaction (Entekhabi et al., 1996). Therefore, it is particularly urgent to develop a downscaling framework from high-spatial-resolution satellite data, coarse-scale SM products and multiple auxiliary datasets based on ML algorithms to obtain SM with an increased accuracy and higher spatial resolution.

In this study, we developed an SM retrieval framework based on the RF algorithm (RF-SM) by integrating in situ observations, Landsat 8 surface reflectance and temperature, a soil properties dataset, the ERA5 meteorological reanalysis dataset (containing relative humidity, atmospheric temperature at 2 m, wind speed at 10 m and 24-h total

precipitation) and four coarse-resolution SM products (SMAP, SMOS, ESA CCI and NCA-LDAS). Our objectives were to (1) evaluate the performance of the retrieval model against the in situ SM under different surface conditions; (2) construct an RF-SM framework based on comparisons between the RF algorithm and three other ML algorithms [extremely randomized trees (Extra-Trees), gradient boosting regression tree (GBRT) and extreme gradient boosting (XGBoost)] using in situ SM; and (3) map the SM spatial distribution at a 30-m resolution at the field scale.

## 2. Data

The four categories of datasets used to develop the SM retrieval framework are listed in Table 1. These datasets are categorized as follows: (1) SM products including an in-situ SM dataset from the International Soil Moisture Network (ISMN), three satellite-based SM products (SMAP Level-3 SM, SMOS-IC SM and ESA CCI V.6.1 SM) and NCA-LDAS SM based on the model, (2) Landsat 8 surface reflectance data from both the Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS), (3) soil properties from SoilGrids V.2.0, and (4) meteorological variables from ERA5.

### 2.1. In situ SM dataset

The ISMN collects and harmonizes in-situ SM datasets from global operational networks (Dorigo et al., 2011) that hold data from 786 sites operated by 6 different networks in the U.S. during 2016 (https://ismn.geo.tuwien.ac.at). In this study, considering the relatively shallow detection depth of Landsat 8 sensors, only the sites capable of measuring the surface SM (at a depth of<5 cm) were selected (Zhang et al., 2022). Moreover, the SM measurements from these stations were filtered based on the quality flags generated by the ISMN, retaining the records with the "G" flag (Dorigo et al., 2013). Finally, a total of 334 sites from 5 networks were used to retrieve the regional SM (Fig. 1).

### 2.2. Landsat 8 OLI and TIRS level-2 dataset

OLI and TIRS are the two sensors on board the Landsat 8 satellite. The OLI collects image data in nine shortwave spectral bands, while the TIRS collects image data in two thermal bands (Irons et al., 2012). In this study, we used Landsat 8 Collection 2 Level-2 datasets produced by the United States Geological Survey and selected two visible and near-infrared bands (i.e., SR_b4 and SR_b5), two shortwave-infrared bands (i.e., SR_b6 and SR_b7) and surface temperature data (i.e., SR_b10) obtained by thermal band processing. The Landsat 8 dataset provides image data of the above five bands at a spatial resolution of 30 m and the information of the cloud state produced by the C Function of Mask (CFMASK) algorithm, thereby providing convenience for data screening and use (Acharya and Yang 2015; Foga et al., 2017). With the help of the Google Earth Engine platform (https://earthengine.google.com), we extracted the band values of these datasets at the locations of selected ISMN sites and retained the records on clear-sky days based on quality control attributes.

**Table 1**
Multiple datasets used to develop SM retrieval framework.

| Dataset | Details | Spatial resolution | Temporal resolution |
|---|---|---|---|
| Soil moisture products | ISMN | Point scale | Hourly |
| | SMAP Level-3 | 36 km | Daily |
| | SMOS-IC | 25 km | Daily |
| | ESA CCI V.6.1 | 0.25° | Daily |
| | NCA-LDAS | 0.125° | Daily |
| Landsat 8 | OLI and TIRS | 30 m | 16 d |
| Soil properties | SoilGrids V.2.0 | 250 m | – |
| Meteorological variables | ERA5 | 0.25° | Hourly |

### 2.3. Soil properties dataset

SoilGrids is a Global Soil Information System that integrates approximately 240,000 soil observation locations worldwide and more than 400 auxiliary environmental covariates and is released by the International Soil Reference Information Centre (ISRIC). This system produces maps of global soil properties at a 250-m spatial resolution by using advanced ML methods (Hengl et al., 2017; Poggio et al., 2021). We used SoilGrids version 2.0 products as a source of soil property datasets to extract the mean clay, sand and silt contents at depths of 0–5 cm at the ISMN sites.

### 2.4. ERA5 reanalysis dataset

ERA5 is the fifth generation global climate and weather reanalysis dataset produced by the ECMWF (Hersbach et al., 2019). ERA5 provides hourly estimates for a large quantity of atmospheric, ocean-wave and land-surface variables from 1950 onward, at a higher spatiotemporal resolution than its predecessor, ERA-Interim (Hersbach et al., 2020). The dataset has been regridded to a regular grid of 0.25° and is freely available from https://cds.climate.copernicus.eu. In this study, we used the atmospheric temperature at 2 m (T2m) and 24-h total precipitation (Precipitation) on single levels and the relative humidity (RH) at various pressure levels and obtained the speed of the horizontal 10-m wind by combining the U component and V component of 10-m wind (Wind).

### 2.5. Soil moisture products

#### 2.5.1. Satellite-based SM products

(1) SMAP SM. The SMAP mission was launched as an L-band satellite on January 31st, 2015, by the National Aeronautics Space Administration (NASA) and scans the Earth's surface in both ascending (6:00 P.M.) and descending (6:00 A.M.) modes on a Sun-synchronous orbit. The SMAP SM product adopts a V-pol single channel algorithm (SCA-V) as a baseline retrieval algorithm, utilizing enhanced Radio Frequency Interference (RFI) mitigation methods to provide estimates of SM at a depth of 5 cm based on brightness temperature (O'Neill et al., 2010). Validation for 191 ground stations worldwide considering the conditions of climate zone, soil properties and land cover types revealed the high mean temporal correlation (0.667 and 0.651, respectively) and low mean unbiased root mean square error (ubRMSE) (0.055 and 0.054 $m^3/m^3$, respectively) for the descending and ascending products (Zhang et al., 2019). In this study, we acquired an SMAP passive Level-3 product posted on a 36-km grid using Equal-Area Scalable Earth Grid, Version 2.0 (EASE-Grid 2.0) from the National Snow and Ice Data Center (https://nsidc.org/data/smap).

(2) SMOS SM. The SMOS satellite is the first global observation satellite to obtain measurements in the microwave L-band, successfully launched by the European Space Agency (ESA) on November 2nd, 2009, in northern Russia (Kerr et al., 2001). SMOS has a 3-day revisit period at the equator while delivering on SM information over land and sea surface salinity over the oceans at 6:00 A.M. for the ascending mode and 6:00P.M. for the descending mode (Kerr et al., 2012). According to validation for 7 stations located in Northeast Brazil, SMOS-IC data conformed to the in situ measurements with the correlation coefficient (R) between 0.53 and 0.86, and ubRMSE between 0.0298 $m^3/m^3$ to 0.0545 $m^3/m^3$ (Araújo et al., 2018). In this study, we used the SMOS-IC product which is independent of auxiliary data such as the SM simulated from ECMWF and vegetation indices obtained from other remote sensing observations (Fernandez-Moran et al.,

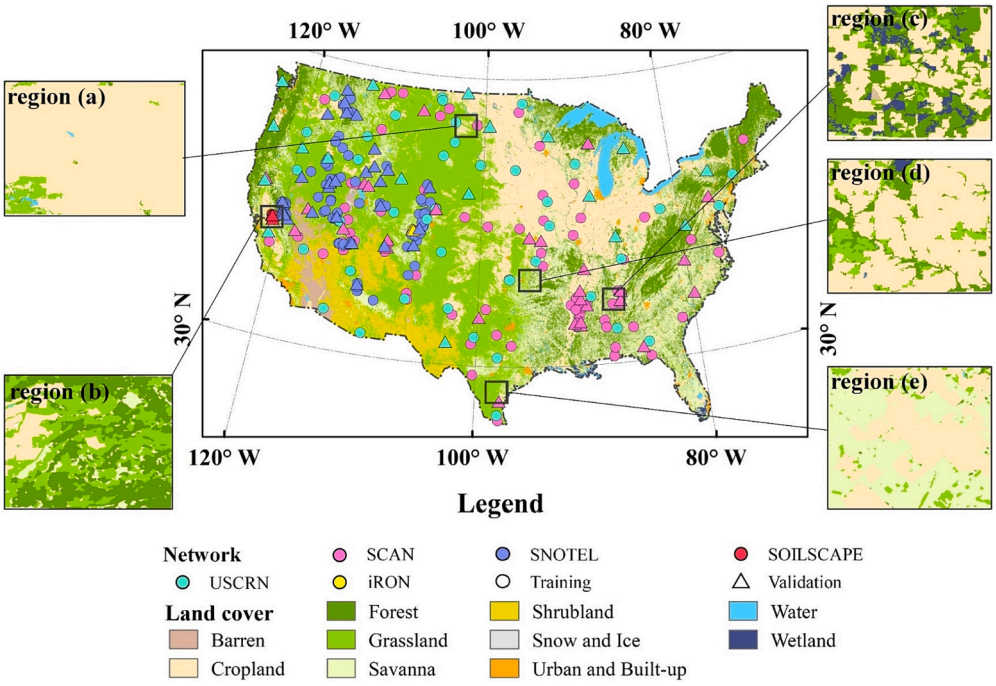



**Fig. 1.** Land cover map of U.S. and the location of 334 SM observation sites used in the study. The circles and triangles represent the training and validation sites, respectively, and the color of the dots indicates the observation network to which the site belongs. The main land cover types of the five substudy areas scattered across the U.S. are shown: region (a), cropland; region (b), grassland and forest; region (c), forest, cropland and wetland; region (d), cropland and grassland; region (e), savanna and cropland.

2017) and provides global daily SM data projected on EASE-Grid 2.0 25-km coordinates (Wigneron et al., 2021; Wigneron et al., 2018).

(3) ESA CCI SM. ESA CCI is a remote sensing monitoring project initiated by the European Space Agency Climate Change Association to monitor the global key climate variables that have feedback effects on climate change; SM was included in this project in 2010. The ESA CCI soil moisture production system provides a global daily SM datasets at a 0.25° spatial resolution (Liu et al., 2012a). With the introduction of new satellites, ESA CCI products have been continuously upgraded; version 06.1 was released in April 2021 and includes active, passive and combined products. The active dataset is derived from three active microwave products (i.e., ERS-1/2, AMI and ASCAT), and the passive dataset is derived from ten passive microwave products (e.g., SMMR, SSM/I, TMI, SMOS, SMAP and FengYun-3B). Based on their respective sensitivity to vegetation density, the combined dataset is produced by the blending weights from the error variance expressed by the signal-to-noise ratio (SNR) (Dorigo et al., 2017; Gruber et al., 2019). The product was evaluated at 596 stations distributed through 28 SM networks worldwide and the average R between ESA CCI data and ground-based SM observations was 0.46 (ubRMSE = 0.05 $m^3/m^3$) for all stations in the period of 1979–2012 (Dorigo et al., 2015). In this study, we used the combined ESA CCI SM product of version 06.1 for 2016 from https://www.esa-soilmoisture-cci.org.

#### 2.5.2. NCA-LDAS SM product

NCA-LDAS is a multisensor, multivariable land data assimilation product that assimilates SM, snow depth, snow cover and irrigation intensity data into the Noah land surface model for North America from 1979 to 2015 based on the NASA Land Information System (LIS) as an enabling tool for the U.S. National Climate Assessment (NCA) (Kumar et al., 2019). In the model configuration of NCA-LDAS, the sensitivities of seven microwave remote sensing sensors (i.e., SMMR, SSM/I, AMSR-E, ASCAT, ASMR-2, SMOS and SMAP) to SM in the low-frequency range are used to retrieval SM (Jasinski et al., 2019), which provides marginal improvements on the already established estimation of SM. Comparison of measured and NCA-LDAS product at five in situ SM networks demonstrated that the surface SM can be reasonably estimated with an average anomaly R of 0.76 and the average ubRMSE of 0.041 $m^3/m^3$ (Kumar et al., 2014). A daily NCA-LDAS Noah-3.3 LSM (Land Surface Model) dataset from https://disc.gsfc.nasa.gov with a spatial resolution of 0.125° was acquired, and the "SoilMoist0_10cm" layer was used to approximately represent SM at a depth of 5 cm in this study.

## 3. Methods

### 3.1. Overall retrieval framework and procedures

We developed a retrieval SM framework based on the RF algorithm (RF-SM) by integrating ground observations, optical/thermal infrared high-resolution data, soil properties, meteorological factors and multiple SM products (Fig. 2). The values of the multiple input variables from the satellite datasets were extracted based on the ground measurement data obtained from the ISMN sites and spatially and temporally collocated with the SM observations. First, we extracted the surface reflectance of two visible and near-infrared bands, two shortwave infrared bands and surface temperatures from the Landsat 8 dataset according to the latitude and longitude information of stations in 2016. Second, the auxiliary variables, including soil property variables (clay, sand and silt) from SoilGrids V.2.0 and meteorological reanalysis variables (RH, T2m, Wind and Precipitation) from ERA5, were extracted synchronously and jointly used as inputs in the ML algorithms. Finally, we introduced three satellite-based (SMAP, SMOS-IC and ESA CCI 6.1) and one model-based (NCA_LDAS) SM products into the model, and used the same method to extract the pixel values of the corresponding products at different scales.

On this basis, the RF and the three other ML algorithms (Extra-Trees, GBRT and XGBoost) were used to build the nonlinear relationships between multiple input variables and the estimated variable. To train models, a fivefold cross-validation method was adopted for model evaluation to reduce the bias caused by the division of the training and testing sets, and the parameters of each model were tuned by the grid search method. During the validation phase, we tested the developed model on sites with different land cover types to evaluate their applicability and robustness. Finally, we uniformly resampled all input datasets at different spatial resolutions to 30 m based on bilinear interpolation, and inputted them into the trained optimal retrieval

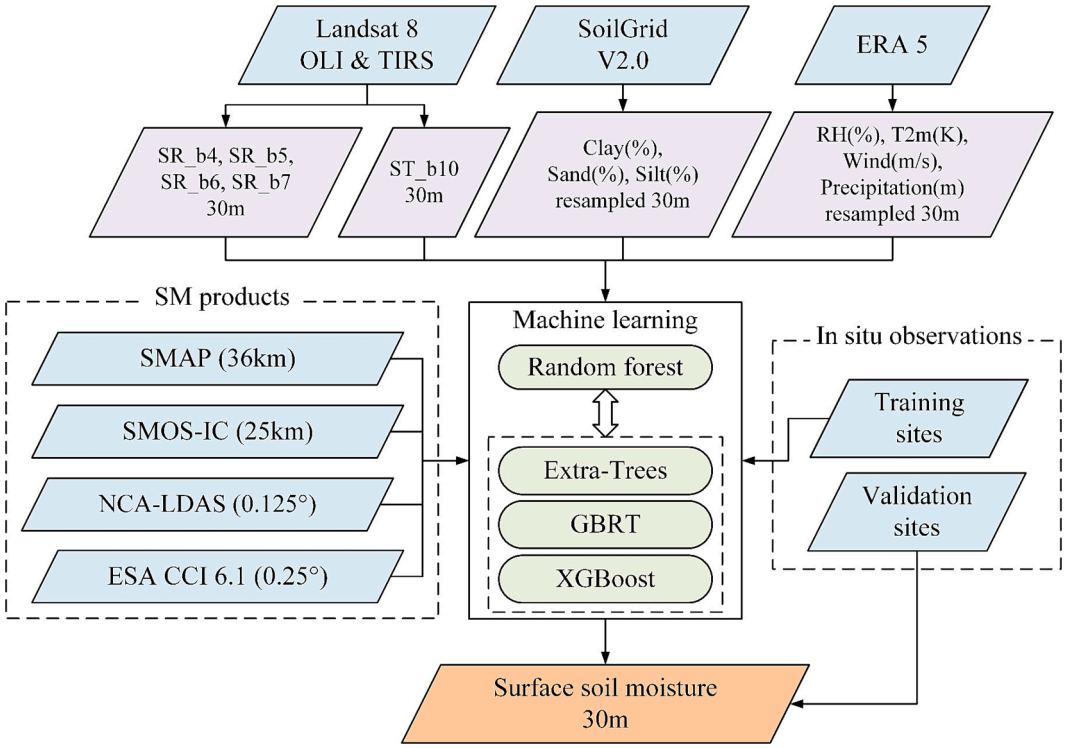



**Fig. 2.** SM retrieval framework and procedures adopted in this study.

model to map the SM spatial distribution at a 30-m resolution.

### 3.2. Random forest algorithm

The RF algorithm is an ML method proposed by Breiman in 2001 (Breiman 2001). Based on decision trees, the RF combines the idea of aggregation and bootstrapping (Hesterberg 2011), to obtain the optimal prediction result by averaging the outputs of several regression trees (Biau and Scornet 2016). The minimum mean square error principle is adopted in each regression tree. For any feature $A$, the two sides of node $s$ are divided into datasets $D_1$ and $D_2$, and the corresponding split node is the feature that minimizes the mean square error of $D_1$ and $D_2$ and minimizes the sum of the mean square error of both sets. This term can be expressed as follows,

$$\min_{A,s}\left[\min_{m_1}\sum_{x_i\in D_1(A,s)}(y_i-m_1)^2+\min_{m_2}\sum_{x_i\in D_2(A,s)}(y_i-m_2)^2\right] \quad (1)$$

where $m_1$ and $m_2$ are the sample output mean values of datasets $D_1$ and $D_2$, respectively.

The classifier combination method not only solves the problem of the low accuracy of a single classification model, but also effectively avoids the phenomenon of overfitting (Breiman 2001). Because the RF randomly selects samples and splits tree nodes in the process of generating a decision tree, it is relatively robust to outliers and noise and greatly improves the generalization ability of the learning system (Belgiu and Drăguţ 2016). In remote sensing applications, the RF can operate efficiently on large databases while performing nondeletion processing and importance assessments of input variables to generate internal unbiased estimates (Rodriguez-Galiano et al., 2012; Zhao et al., 2018).

### 3.3. Other machine learning algorithms

#### 3.3.1. Extremely randomized Trees

Extremely Randomized Trees (Extra-Trees) is a decision tree-based ML proposed by Geurts and other scholars after many experiments in 2006 (Geurts et al., 2006). Extra-Trees is a variant of the RF because it integrates multiple decision trees for scoring and votes based on the average value predicted by each decision tree. However, considering that the RF uses the random and replacement method to obtain the training set, causing repeated samples to be generated so that all samples cannot be guaranteed to be fully utilized and similarities may exist between each decision tree, Extra-Trees uses the entire training sample and splits nodes by choosing a cut point completely randomly (Geurts et al., 2006).

#### 3.3.2. Gradient boosting regression tree

The gradient boosting regression tree (GBRT) is an ML algorithm based on the boosting framework proposed by Friedman (Friedman 2001, 2002). The basic principle is to establish a new model in the gradient direction of the model residual reduction and continuously iterate to generate a combination of multiple weak classifiers. The resulting strong classifier has the advantage of minimizing the loss function. As an effective off-the-shelf nonlinear learning method (Johnson and Zhang, 2013), the GBRT has a strong generalization ability to flexibly handle various types of data and process skewed variables without requiring transformations (Wang et al., 2016). In addition, the model retains the advantages of the interpretability of the regression tree model, while using some robust loss functions that are very robust to outliers.

#### 3.3.3. XGBoost

Extreme gradient boosting (XGBoost) is another ML algorithm that is based on the GBDT (gradient boosting decision tree) and capable of gradient boosting “on steroids” (Chen and Guestrin 2016). XGBoost achieves accurate classification through the iterative calculation of weak classifiers, that is, it continuously adds feature splitting trees to the classifier. Each tree is equivalent to a weak classifier, and after these trees are integrated, the modeling results of all weak classifiers are aggregated to obtain a better classification performance than that of any single model. It is noteworthy that the regularization term is added to

the objective function to reduce the complexity of the model and avoid overfitting (Chen and Guestrin 2016).

### 3.4. Model evaluation methods

In this study, four statistical indicators were used to evaluate the accuracies of four SM products and the SM retrieval model as well as the performances of the four ML algorithms: the coefficient of determination ($R^2$), root-mean-square-error (RMSE), bias (Bias) and Kling-Gupta efficiency (KGE) (Gupta et al., 2009). $R^2$ represents the proportion of the ground measurement variation that can be explained by the control variable; RMSE measures the closeness between the predicted and observed SM values; and Bias reflects the error between the output of model on the sample and ground measurements. Some of these indicators can be calculated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_i - O_i)^2} \tag{2}$$

$$Bias = \frac{1}{N}\sum_{i=1}^{N}(S_i - O_i) \tag{3}$$

where $N$ is the total number of samples and $S_i$ and $O_i$ denote the model-predicted and ground-observed SM, respectively. KGE summarizes multiple metrics to achieve a comprehensive model evaluation, including the correlation ($r$), relative variability ($\alpha$) and ratio of the mean ($\beta$) in the simulated and observed values. Its decomposition formula and components are expressed as follows:

$$KGE = 1 - ED \tag{4}$$

$$ED = \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{5}$$

$$\alpha = \frac{\sigma_s}{\sigma_o} \tag{6}$$

$$\beta = \frac{\mu_s}{\mu_o} \tag{7}$$

where $ED$ is the Euclidian distance from the ideal point on the three-dimensional Pareto front; $\sigma_s$ and $\mu_s$ are the standard deviation and mean of the simulation of the model, respectively; $\sigma_o$ and $\mu_o$ are the standard deviation and mean of the observations, respectively. In the absence of simulation errors, all three components in KGE have a value of 1, meaning that the ideal KGE value is 1.

Moreover, Taylor diagrams (Taylor 2001) are another method used to comprehensively evaluate the performances of modeling algorithms. A Taylor diagram can present multiple performance evaluation metrics in a two-dimensional polar-style graph. The matching degree between the predicted and observed values is indicated by the correlation coefficient (r), the centered root-mean-square difference (RMSD) and the standard deviations (SD) and corresponds to a single point on the graph. Taylor diagrams allow one to measure the accuracy of a model in simulating the natural system and are ideal for evaluating the relative metrics of multiple models or sets of predictions.

### 3.5. Experimental setup

We built the RF and the three other ML algorithms (Extra-Trees, GBRT, and XGBoost) using the Scikit-learn package (Pedregosa et al., 2011) based on the Python platform. The remote sensing reflectance variables (SR_b4, SR_b5, SR_b6, SR_b7, and SR_b10), soil property variables (Clay, Sand, and Silt), meteorological reanalysis variables (RH, T2m, Wind, and Precipitation) and SM products (SMAP, SMOS, ESA CCI, and NCA-LDAS) were selected as the input variables, and the in situ SM observed at the sites was used as the output variable. For each training model, fivefold cross-validations were performed to evaluate the performances of the four models and find the optimal set of parameters.

To obtain better performances of the different ML algorithms, the parameters of the four models in the sklearn module were tuned separately, including n_estimators (the number of trees), max_depth (the maximum depth of a tree), eta (the learning rate), and subsample (the subsample ratio of columns for each level). Furthermore, the preset parameter combinations were looped through the GridSearchCV module to find the optimal parameter combination for each model. Although this parameter-tuning method is time-consuming, it can improve the model estimation accuracy while obtaining an SM retrieval framework with a high efficiency and good performance.

In the study, we used data from 334 ground measurement sites, covering six different land cover types: barren (7 sites), cropland (50 sites), forest (22 sites), grassland (145 sites), savanna (104 sites) and shrubland (6 sites). Among all records, 70% of the sites were used for training, and the remaining sites were used to validate the RF and the three other ML algorithms.

## 4. Results

### 4.1. Validation of four coarse SM products at site scale

To assess the performance of the four coarse SM products used in this study, we validated SMAP, SMOS, ESA CCI and NCA-LDAS based on ground measurements from 334 sites (Fig. 3). The results show that none of these products could accurately estimate SM due to the spatial mismatch or the fault of the satellite measurements. The SMAP product obtained the highest KGE value (0.46) but the lowest $R^2$ value (0.24, $p < 0.01$), while the NCA-LDAS product was had the highest $R^2$ value (0.41, $p < 0.01$) and the lowest KGE value (0.18). The RMSE of the ESA CCI product was the lowest at 0.098 $m^3/m^3$ and the RMSE order of the other three SM products ascended from SMOS (0.101 $m^3/m^3$) and SMAP (0.111 $m^3/m^3$) to NCA-LDAS (0.111 $m^3/m^3$). In addition, the bias values of the four products varied from −0.036 $m^3/m^3$ to 0.075 $m^3/m^3$, and only that of SMOS was negative. This means that the SMOS product underestimated of SM to some degree, as can also be seen from the scatter plot. The integrated results showing that each product has statistical strengths and weaknesses differing from one another lend themselves well to our ML algorithm here that capitalizes on their strengths while minimizing their weaknesses.

### 4.2. Evaluation of the RF algorithm for SM retrieval

#### 4.2.1. Model development based on the training sites

To better develop the SM retrieval framework, we trained the RF and three other ML algorithms based on the in situ SM observed at the 234 training sites. Fig. 4 shows the scatter plots obtained for the SM observations and SM training results using different algorithms. In terms of the SM predictions, the RF, Extra-Trees and XGBoost yielded similar results based on the $R^2$ (0.93, 0.93 and 0.93, $p < 0.01$, respectively), RMSE (0.033 $m^3/m^3$, 0.032 $m^3/m^3$ and 0.032 $m^3/m^3$, respectively) and Bias (-0.001 $m^3/m^3$, −0.001 $m^3/m^3$ and 0.004 $m^3/m^3$, respectively), and performed slightly superior to the results of the GBRT method in terms of the $R^2$ (0.92, $p < 0.01$) and RMSE (0.034 $m^3/m^3$). However, as far as KGE is concerned, due to the higher RMSE of the RF, its advantages were not outstanding compared to the other two algorithms, with a gap of 0.02.

#### 4.2.2. Model evaluation based on the validation sites

Fig. 5 presents the scatter plots of the SM observations recorded at the 100 validation sites and the SM results predicted using the RF and three other algorithms. The results show that the RF yielded the best performance with a high KGE value (0.69) and relatively low RMSE value (0.063 $m^3/m^3$) among the four methods. Although Extra-Trees had an acceptable accuracy based on its $R^2$ (0.66, $p < 0.01$) and RMSE (0.062 $m^3/m^3$), it performed the worst in terms of the KGE value (0.67) compared to the other methods. On the other hand, the GBRT had the

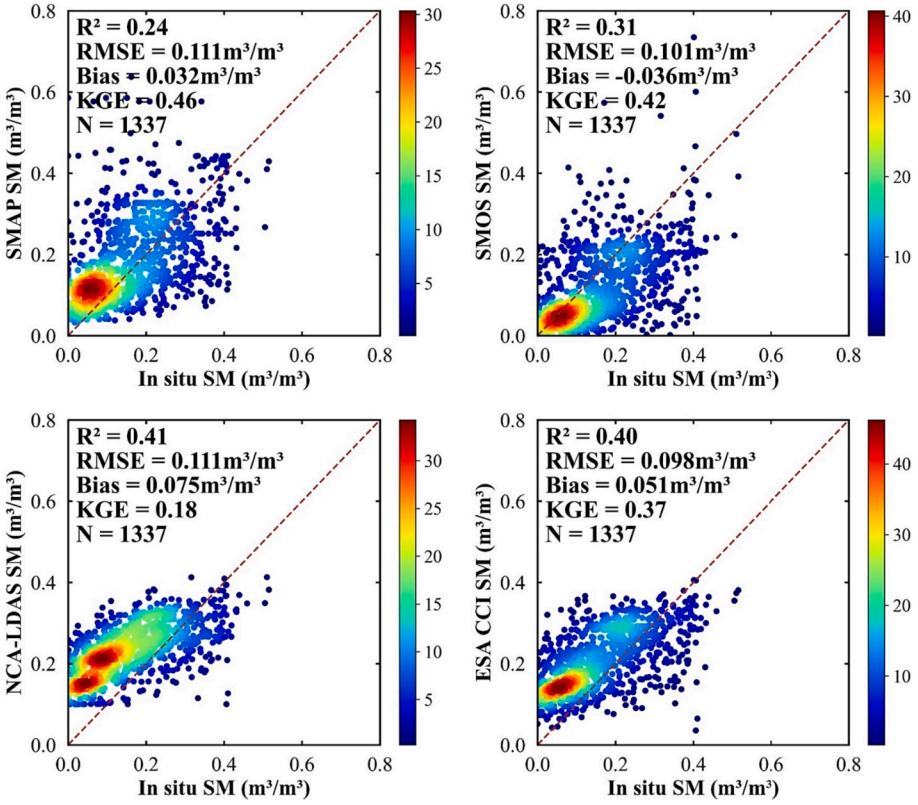



**Fig. 3.** Comparison of the four SM products and ground measurements. The color of the dots denotes the probability density of a particular location. The dark red dashed lines indicate the 1:1 line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

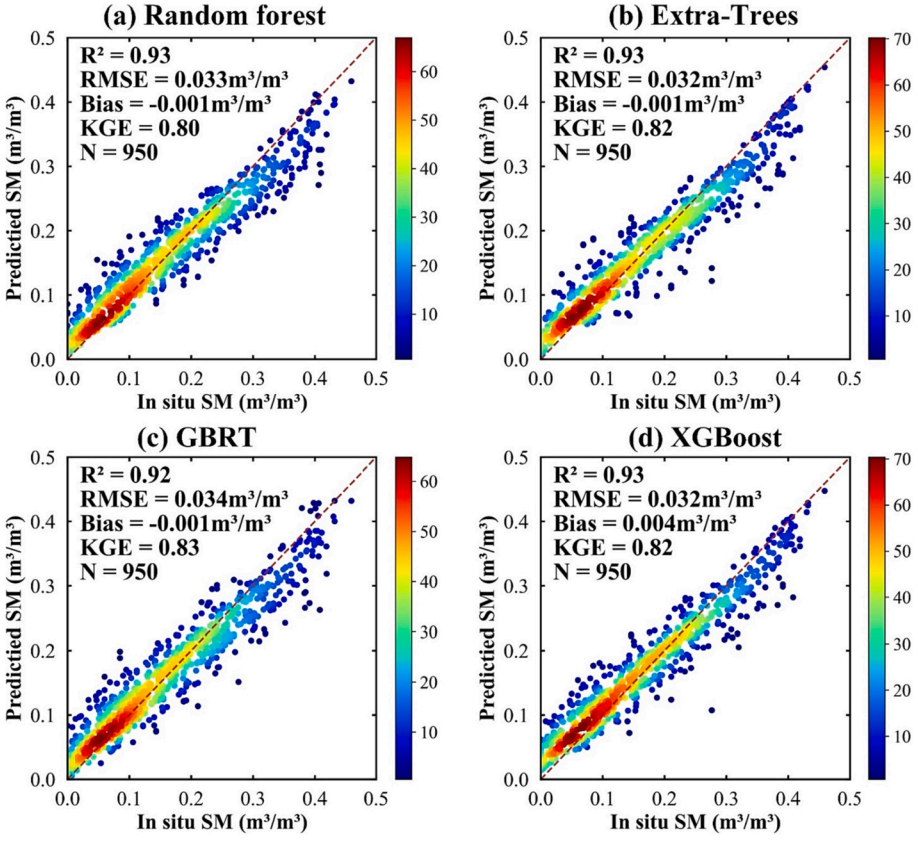



**Fig. 4.** Evaluation of the SM predicted by the (a) RF, (b) Extra-Trees, (c) GBRT and (d) XGBoost algorithms against the ground measurements at 234 training sites.

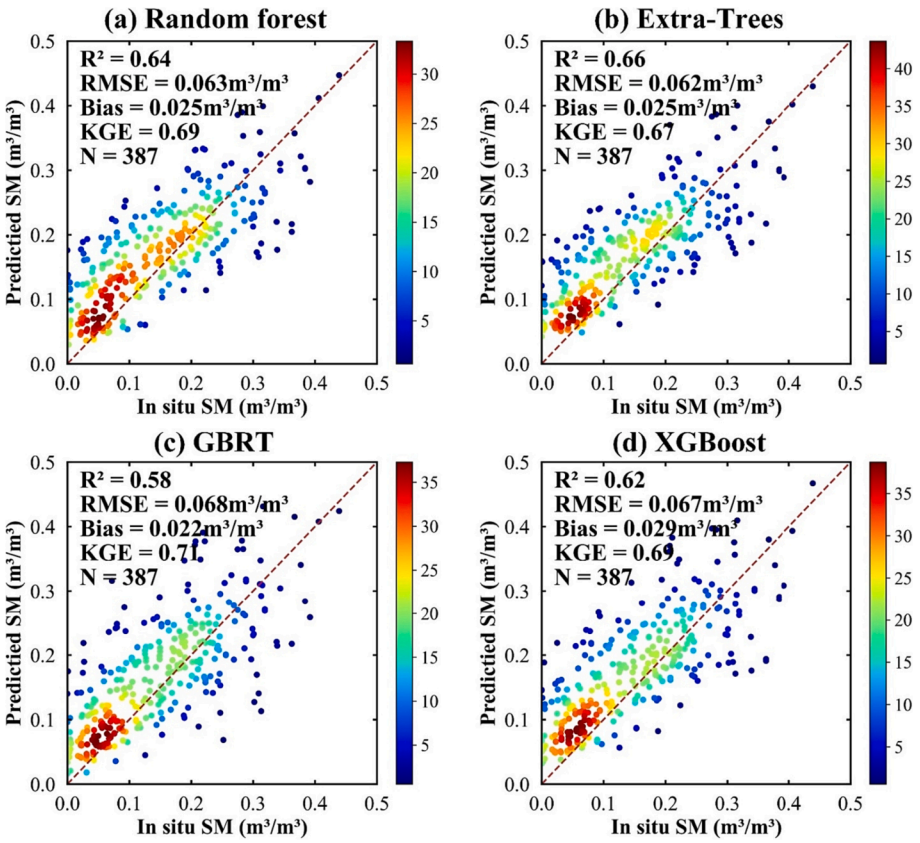



**Fig. 5.** Evaluation of the SM predicted by the (a) RF, (b) Extra-Trees, (c) GBRT and (d) XGBoost algorithms against ground measurements at 100 validation sites.

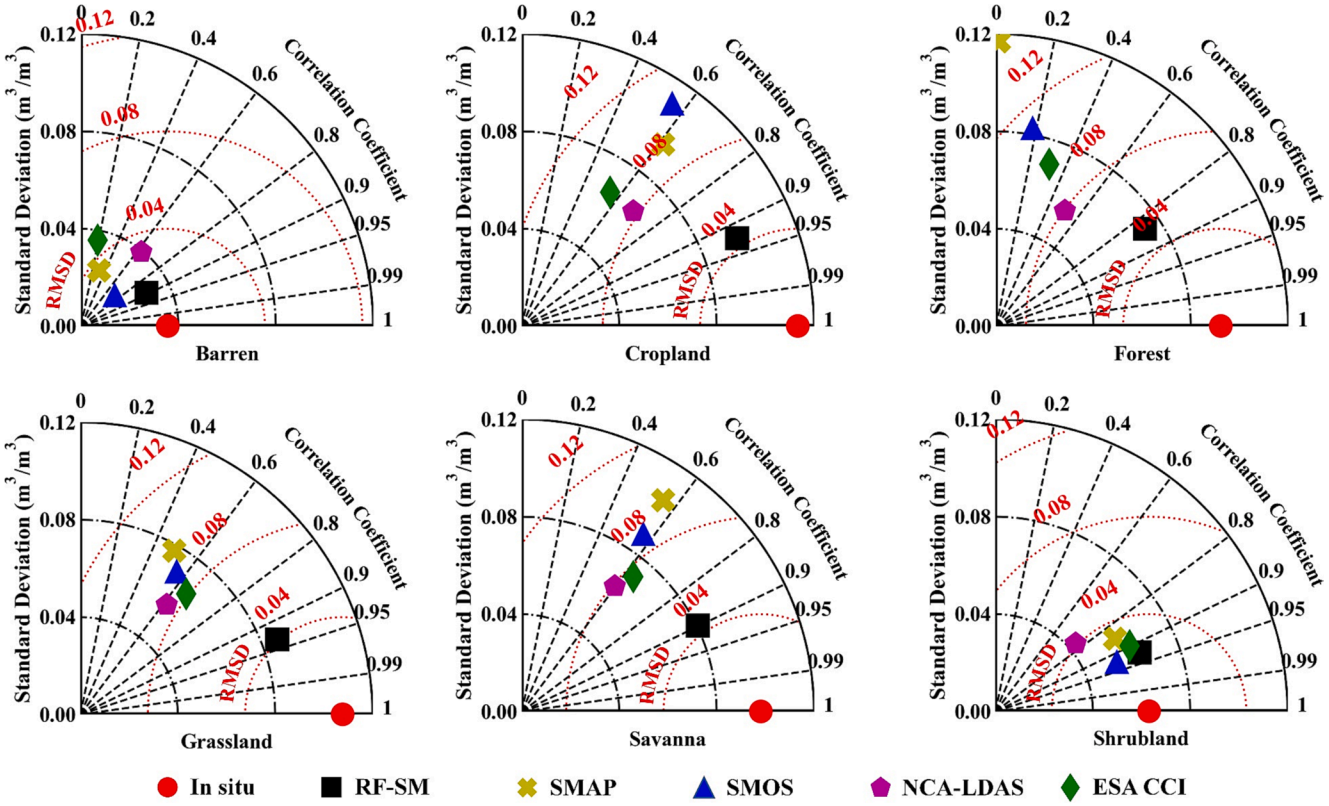



**Fig. 6.** Taylor diagrams showing the comparison between the RF-SM and multiple SM products against ground measurements take in sites with different land cover types.

highest KGE value (0.71), but was not considered the optimal algorithm for SM estimation due to it having the lowest $R^2$ value (0.58, $p < 0.01$) and highest RMSE value (0.068 $m^3/m^3$).

#### 4.2.3. Model evaluation for different land cover types

Based on our evaluation of the performances of different ML models, we developed the RF algorithm-based SM retrieval model (RF-SM) and continued to carry out follow-up experiments. Fig. 6 shows the Taylor diagrams for SM ground measurements with different land cover types and estimations from individual products, which indicates that the RF-SM predictions were significantly better than those of the other SM products and had a relatively stable accuracy, with r values between 0.84 and 0.94, and RMSD values between 0.017 $m^3/m^3$ and 0.051 $m^3/m^3$. In fact, differences in surface and climatic conditions prominently impact the SM estimation accuracies of these four products.

The ESA CCI demonstrated the best performance at the grassland and savanna sites with the highest r values (0.66 and 0.64, respectively) and the lowest RMSD values (0.080 $m^3/m^3$ and 0.078 $m^3/m^3$, respectively), but its performance on the barren sites was slightly less optimal than those of the other products. For the cropland and forest sites, NCA-LDAS had a better performance than the other three products, with the highest r values (0.69 and 0.53, respectively) and the lowest RMSD values (0.081 $m^3/m^3$ and 0.079 $m^3/m^3$, respectively). However, NCA-LDAS yielded the highest RMSD value (0.040 $m^3/m^3$) and the lowest r value (0.78), making it inferior to the other products on shrubland sites. The SMOS product had excellent performances at the barren and shrubland sites, with r values of 0.80 and 0.94, respectively, and RMSE values of 0.016 $m^3/m^3$ and 0.021 $m^3/m^3$, respectively.

Additionally, the SDs of the predictions and all SM products were lower than those of the ground measurements under corresponding surface conditions, revealing a potential underestimation trend of these models and products when making SM predictions. Overall, none of the individual products could accurately estimate SM, given that the four SM products showed significant discrepancies in different land cover types. However, this situation was greatly improved using the comprehensive RF-SM that incorporated multiple products and various auxiliary variables.

To further analyze the temporal variations in the RF-SM predictions under different surface conditions, we selected representative observation sites corresponding to each land cover type, simultaneously extracted the corresponding spatiotemporal pixel values on different SM products, and plotted these values against the predictions obtained by the RF-SM and the in situ SM recorded during 2016 (Fig. 7). The results showed that the RF-SM outperformed the other individual products in capturing the temporal variabilities in SM and precipitation events under the six land cover types. Furthermore, two SM products, the SMAP and SMOS products, produced SM estimations with higher consistencies with the ground observations, especially at cropland, savanna and shrubland sites, whereas NCA-LDAS and ESA CCI produced significant overestimations. However, overall, the SMOS products still suffered from underestimation, which was most pronounced at the forest and grassland sites. The ESA CCI product performed well in capturing the variations in SM and precipitation events, and the consistency of these estimations with the in situ SM was effectively validated at the grassland and savanna sites.

At the SCAN_Lind#1 site (cropland), the temporal trend of SM during 2016 showed a continuous decrease from April to October, and increases due to the occurrence of precipitation events from March to April and after October. The RF-SM predictions were slightly overestimated at this site, especially on March 23rd, which was affected by precipitation, and from June through October, when the area experienced a relatively long drought. At the USCRN_Newton-8-W site (savanna), the SM values were stable between 0.1 and 0.2 $m^3/m^3$ throughout the year, and two relatively strong precipitation events occurred on March 26th and September 2nd, but neither affected the SM trend. The RF-SM predictions also performed well, especially on September 9th, when a precipitation event occurred.

At the USCRN_Socorro-20-N site (shrubland), the RF-SM predictions were high consistent with the in situ SM; at the USCRN_Yuma-27-ENE site (barren), the SCAN_Shenandoah site (forest) and the USCRN_Stillwater-2-W site (grassland), RF-SM predictions all showed instability, suggesting that SM and its variation may be affected by meteorological and surface conditions other than precipitation at these sites.

### 4.3. Spatial patterns of RF-SM at the field scale

To better detect the performance of the spatial distribution in the retrieved SM at the field scale, Fig. 8 shows the spatial variations based on RF-SM with different background fields in five substudy areas with sizes of 9 km × 9 km during 2016. These spatial patterns clearly show the surface texture information such as fields, ridgelines and roads, and reflect the SM spatial variability at the 30-m × 30-m field scale. Therefore, one may reasonably come to the conclusion that the proposed framework can retrieve field-scale SM at a relatively high spatial resolution by exploiting the RF algorithm and the information contained in multiple satellite datasets.

Moreover, differences in the values and spatial trends of SM were identified among five substudy areas. Regions (a) and (b) are located in the northern and western U.S. and showed relatively low SM levels, especially in June and July, no higher than 0.15 $m^3/m^3$; In the central and eastern of regions (c) and (d), the SM in November still showed low values at the annual scale, but the actual surface SM was not lower than those in regions (a) and (b). Region (e), located on the south coast, showed low SM values throughout the year. In general, the SM values obtained in the same period showed spatial heterogeneities, and this was of great benefit for exploring the temporal and spatial variabilities in SM at the field scale.

## 5. Discussion

### 5.1. Performance of the RF-SM

#### 5.1.1. Ability of the RF-SM to retrieve SM

Based on the RF ensemble learning algorithm, we introduced satellite remote sensing datasets, soil properties, meteorological reanalysis products and multiple SM products into the model and constructed an SM retrieval framework to solve the limitations associated with the low spatial resolution. In our elucidation of the complex relationships between these multiple SM datasets and the auxiliary variables at different scales, the ML algorithms were found to play an important role, because their strength lies in their tolerance of multicollinearity among parameters under coupling and feedback effects (Im et al., 2016). According to the model evaluation results (Fig. 5), compared to the four products, the KGE value of the prediction model had increased by nearly 0.2, and the RMSE and Bias values also greatly reduced.

In this study, the RF and three other ML algorithms (Extra-Trees, GBRT and XGBoost) were used to predict SM. Both the RF and Extra-Trees algorithms had good model performances in terms of the independent verification site, with high $R^2$ value (0.64 and 0.66, respectively, $p < 0.01$) and low Bias value (0.025 $m^3/m^3$). These two algorithms have similar principles in their combination of weak classifiers, that is, the bagging method is based on the ensemble decision-making process (Altman and Krzywinski 2017). However, the decision tree division of the Extra-Trees algorithm is a more radical random selection method that differ from the optimal position of RF, leading to an increase scale of the generated decision tree (Geurts et al., 2006). At the same time, compared to the RF algorithm, the variance in the Extra-Trees results was further reduced, but the deviation was further increased, resulting in a lower KGE value (0.67) than that of the RF (0.69). The GBRT and XGBoost algorithms are both boosting methods, meaning they stack multiple weak classifiers layer by layer based on the

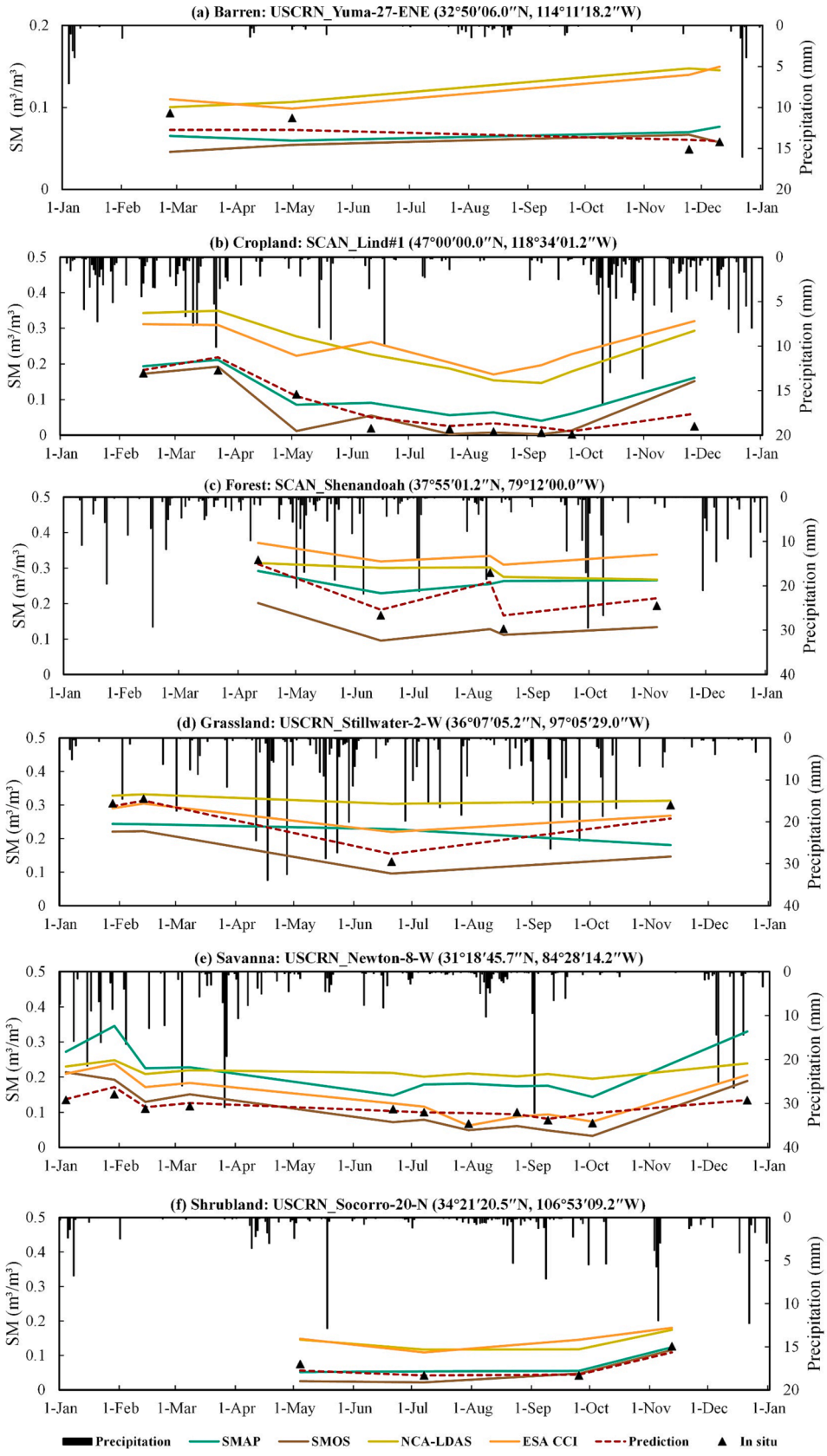



**Fig. 7.** Temporal variations in the in situ SM, multiple SM products, RF-SM and precipitation at the representative sites corresponding to each land cover type.

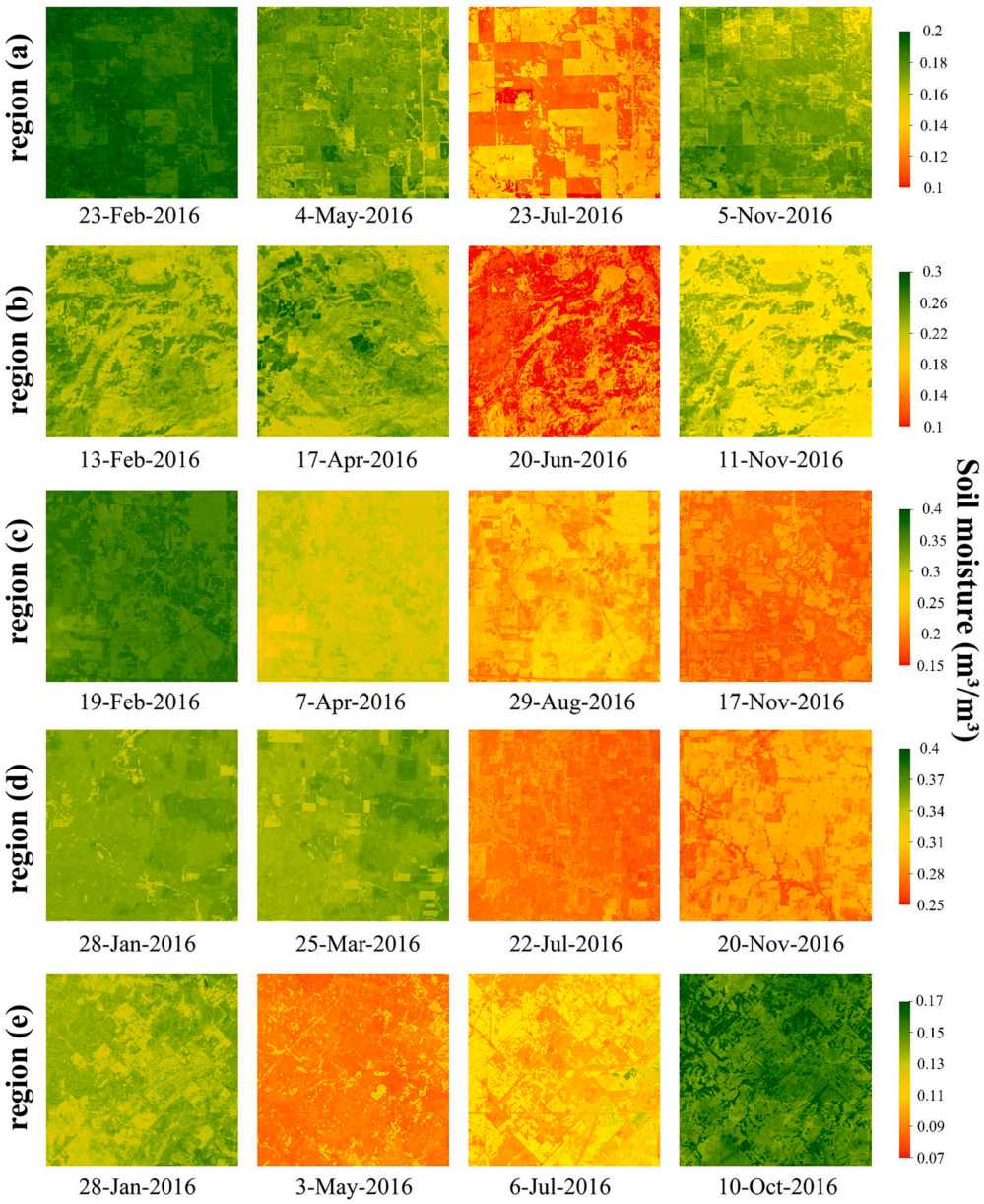



**Fig. 8.** Maps of the RF-SM results with a 30-m spatial resolution in the five substudy areas in 2016.

error rate and reducing the deviation to improve the model prediction accuracy (Ferreira and Figueiredo 2012). The difference is that XGBoost is optimized on the basis of the GBRT, and this model improves the calculation rate while suppressing overfitting (Chen and Guestrin 2016). In short, there are completely different ensemble strategies between bagging via the variance reduction scheme and boosting via the bias reduction scheme (Bühlmann 2012), but bagging technology often has a more robust performance in the face of data noise and imbalance (Barutçuoğlu and Alpaydın 2003; Khoshgoftaar et al., 2010). Considering the instability and obvious positively skewed distribution of experimental data in real scenarios, the RF-based prediction framework can obtain better SM estimations from multiple satellite datasets under complex surface conditions.

After further verification of the SM products based on different land cover types (Fig. 6), we found that the performances of the four SM products were similar at the cropland, grassland and savanna sites, but large errors appeared at the barren and forest sites. Since the sensitivity of the brightness temperature to SM might be significantly attenuated by the effects of vegetation and canopy inherent emissions (Cui et al., 2017), it is difficult for microwave products to accurately capture the SM trend corresponding to each land cover type. In areas with moderate vegetation, the L-band-based SM products (e.g., SMAP and SMOS) performed well; in areas with sparse or dense vegetation, these SM products exhibited relatively poor skills (Ma et al., 2019). At the site scale (Fig. 7), the data assimilation product (NCA-LDAS) and the active–passive combination product (ESA CCI) produced continuous overestimation of SM, while the SMOS product showed a trend of underestimation. The SMOS sensors are highly susceptible to RFI signals, which may also affect the ESA CCI products that integrate SMOS data to some extent (Al-Yaari et al., 2019). The validation results obtained in this study (Fig. 3) are consistent with the evaluation results of Al-Yaari et al. (2019) on SM products in the U.S. during 2015–2016. These biases depend largely on the model parameters in the satellite retrieval process and the least important metric in the algorithm calibration process (Fernandez-Moran et al., 2017; Jasinski et al., 2019).

After introducing multiple satellite datasets and auxiliary variables based on the RF-SM, the SM retrieval accuracy of the prediction model was also improved compared to the existing products, as was reflected in

the fact that the r values exceeded 0.8 and the SD value gap was narrowed to within 0.04 $m^3/m^3$ compared to the ground observations (Fig. 6). Therefore, the RF-SM provides a good strategy, that is, it predicts SM based on a situation with good numerical fitting and trend responses, thereby not only making up for the unstable observation results of the four products, but also successfully inheriting the advantages of a high sensitivity to precipitation capture.

#### 5.1.2. *Modeling contributions of the ancillary variables and single SM product*

To explore the contributions of the auxiliary variables to the RF-SM, we used the feature importance evaluation in the RF algorithm to attribute the contribution of each input variable to the retrieval framework predictions (Goldstein et al., 2011). A measure of feature importance can highlight how relevant input variables are to the target, thereby providing a better understanding of the physical mechanism of the RF algorithm (Breiman 2001). As shown in Fig. 9, the feature importance of the meteorological stress and surface properties accounted for 36.3% of all input variables, indicating that these auxiliary variables played important roles in the RF-SM retrieval framework.

The Landsat dataset made an important contribution to the RF-SM enabling the downscaling of coarse-resolution SM products. By using fine-scale optical images and features, including the vegetation index and thermal infrared surface temperature, as information sources, the response to SM variations and the disaggregation of coarse-resolution pixels can be realized in a triangular or trapezoidal feature space (Colliander et al., 2017; Peng et al., 2015). To fully exploit the potential spectral reflectance of the bands, we directly incorporated the red and near-infrared bands into the model, without considering the regional differences caused by the use of vegetation indices with specific physical significance. The results (Fig. 9) indicate that the red band (SR_b4), near-infrared band (SR_b5) and thermal-infrared band (SR_b10) all have high feature importance levels, and the contribution rate of SR_b5 was especially notable, accounting for 4.0% of the input variables. In addition, two shortwave-infrared bands (SR_b6 and SR_b7) also played considerable roles, and their effects on SM retrieval have been confirmed in previous studies and applications (Ngo Thi et al., 2019; Yue et al., 2019).

The soil properties and meteorological stress are two factors that characterize the spatial heterogeneity of SM (Chaney et al., 2015). In the RF-SM, the contribution of soil properties accounted for 11.6%. Since soil properties control water infiltration and thus drive SM dynamics, the spatial variation in SM is highly correlated with soil properties, especially under wet conditions (Baroni et al., 2013). Moreover, this finding provides the possibility to explore the SM temporal change process following heavy precipitation. The near-surface air variables (e.g. air temperature and air humidity) characterized by the variations of latent and sensible heat flux reflect the SM variation indirectly, and a feedback mechanism between the land and atmosphere is also realized through precipitation cycle (Entekhabi et al., 1996). Therefore, these meteorological variables will play crucial roles on SM retrieval accurately (Cai et al., 2019). In the study, the total contribution rate of the four meteorological reanalysis variables was only 9.8%, and their importance in the model was relatively weak. This may be due to the weakening of the spatial heterogeneity of meteorological parameters on the coarse grid or to the decrease in precision after the daily scale aggregation (Zhang et al., 2022). In addition, the feature importance of Precipitation was the lowest among these variables, because precipitation may show a weak or even negative correlation with SM, especially in humid or temperate regions with dense vegetation (Sehler et al., 2019).

The SM products had the highest feature importance proportions, among which the contribution rate of the active–passive combined ESA CCI product and the model-based NCA-LDAS product exceeded 20%. To evaluate the effect of different products on the RF-SM, we conducted 16 groups of experiments independently. Based on the four ML algorithms, we removed each individual product one by one and used the three remaining coarse-scale products as the background for ensemble learning (Fig. 10). The model evaluation performed based on the validation sites declared that the two SM products with outstanding feature importance levels played key roles in the retrieval process, because when they were removed in the experiment, all four ML algorithms showed significant declines in the model accuracy. Moreover, compared to the results of the other experiments, the removal of NCA-LDAS and ESA CCI led to greater reductions in $R^2$ and KGE values and increases in the RMSE and Bias values. In terms of the downscaling scheme, the

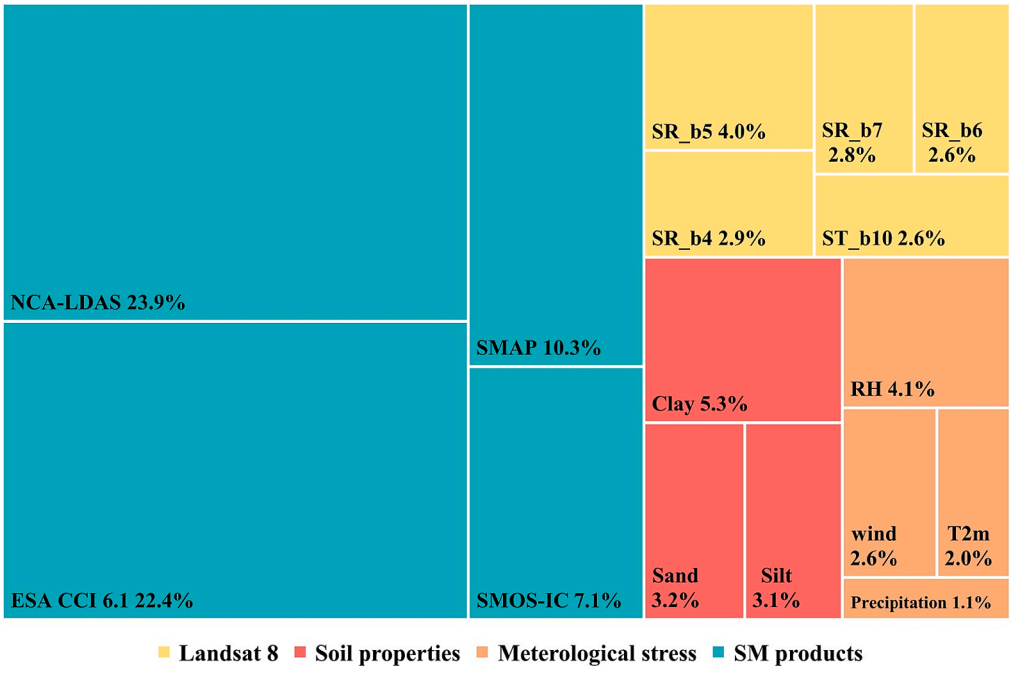



**Fig. 9.** Importance of input variables to the RF-SM. The contributions of the inputs to the RF-SM output were dominated by the SM products, followed by the Landsat 8 dataset. The soil properties and meteorological stress were less important.

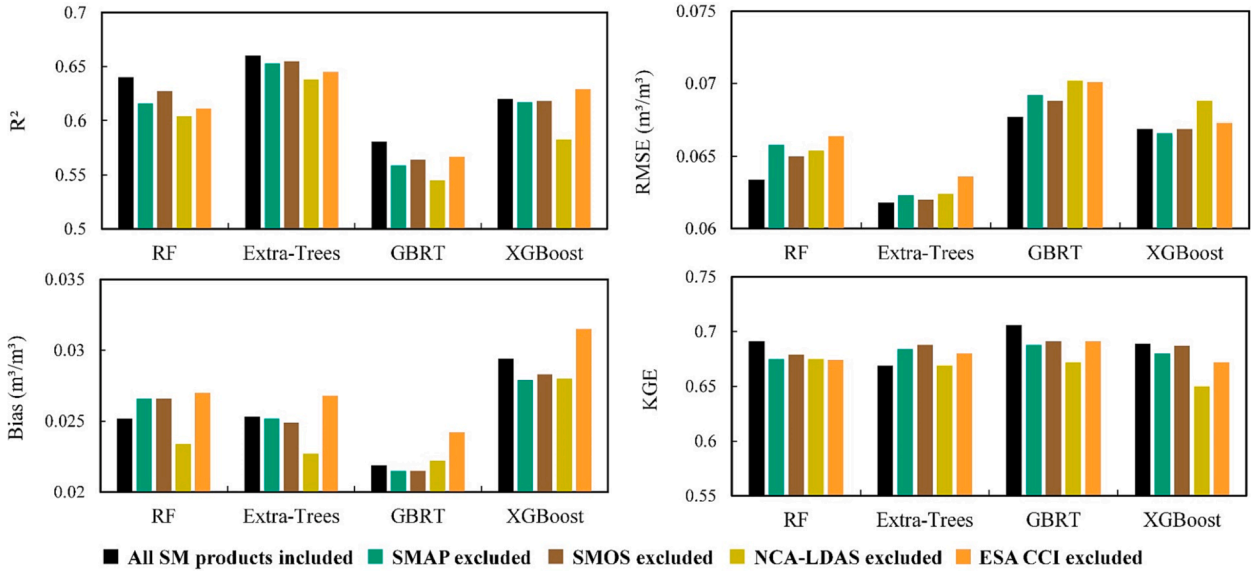



**Fig. 10.** Variations in the four statistical metrics of the ML algorithms on different sets of variables based on multiple SM products.

higher the resolution of the product was, the more the scheme can help achieve the fine-scale disaggregation of SM (Peng et al., 2021). However, the algorithms adopted for data assimilation or merging all incorporate various active and passive products, such as SMAP and SMOS; this indicates that although NCA-LDAS and ESA CCI can provide key scale references, the maintenance of the original satellite-derived retrieval state and the consideration of the results in the case of inconsistencies between multiple products may bring greater errors to the algorithm itself (Jasinski et al., 2019; Liu et al., 2012a). Therefore, combining multiple background fields can improve the spatial consistency and accuracy of the downscaled SM (Abowarda et al., 2021).

### 5.2. *Uncertainties in the RF-SM retrieval framework*

Within the framework of the RF-SM, we completed the 30-m resolution SM mapping of the substudy areas, and this greatly improved the representation of the spatial SM pattern. However, ML algorithms based on multiple satellite datasets inevitably have many uncertainties, including data sources errors, model structure uncertainties, and scale matching uncertainties (Abowarda et al., 2021; Guevara and Vargas, 2019; Shin and Mohanty 2013). Ground observations from the ISMN were used as the truth data, and the quality of these sites was generally controlled through measurement devices, sensor deployment and their calibration (Bogena et al., 2007). Previous studies have shown that, in addition to random noise errors in ground-measured SM readings (Dorigo et al., 2013), the sources of errors in the station measurements include typical soil wetting events caused by precipitation and continuous drying (Hillel 1998), alternating high and low SM due to soil freeze–thaw cycles (Hallikainen et al., 1985), and temperature sensitivities exhibited by measuring devices (Robinson et al., 2008). More than 90% of erroneous data can be screened out through quality-control methods, but errors in soil temperature and porosity still cause considerable omission or over flagging problems (Dorigo et al., 2013; Liao et al., 2019). At the same time, there are large deviations in SM products and meteorological reanalysis datasets (Dorigo et al., 2017; Quets et al., 2019; Tarek et al., 2020), including error accumulations under the constraints of radiative transfer model parameters and auxiliary information (Konings et al., 2011; Li et al., 2015; Reichle et al., 2004), uncertainty propagations related to the instrument configurations and algorithm assumptions (Parinussa et al., 2011).

The RF-SM approach structure introduces large uncertainties in the SM retrieval. The framework proposed herein made predictions by establishing the complex relationships between SM and different land surface and meteorological parameters and using certain ground observation data to carry out model training; this method also helped us obtain downscaled SM at a high accuracy and consistency. However, the algorithm was strongly influenced by the training data (Probst et al., 2019). On the one hand, the RF has a limited ability to predict conditions beyond the target range of the training set, and this increases the possibility of overfitting when processing specific noise data (Adab et al., 2020); on the other hand, when the decision tree is large, the storage and time required for model training increase, and this affects the efficiency and cost (Liu et al., 2012b).

Mismatches in spatial scale between ground measurements and satellite observations cause large uncertainties in modeling and validation results (Peng and Loew 2017). One single sensor has a limited spatial representativeness compared to large-scale satellite data (Gruber et al., 2013). Comparing SM datasets across scales, including SM products with several-kilometer grid cells and the RF-SM spatial pattern at 30-m spatial resolution, brings inevitable errors due to scale mismatches. In addition, SM networks tend to exhibit high spatial heterogeneity, yet the small number of observations collected under a satellite footprint can lead to falsely inflated retrieval errors (Miralles et al., 2010). Especially over complex underlying surfaces, there is a greater reliance on the quantity and quality of ground observation sites, both for coarse-resolution and fine-scale SM products (Zappa et al., 2020).

### 5.3. *Advantages and disadvantages of the RF-SM*

In this work, we developed the RF-SM retrieval framework to overcome the limitations of low-spatial-resolution passive microwave remote sensing products at the field scale, and achieved an overall consistency between the predicted SM and in situ SM datasets under different climate conditions and land cover types. First, we conducted rigorous model training using in situ measurements to provide accuracy guarantees for the establishment of model between fine-scale surface variables and coarse-scale SM and then took advantage of the original resolutions of the variables to produce SM at a relatively fine scale that enables the full potential of remote sensing data and in-situ measurements coordinated by RF algorithms (Abowarda et al., 2021; Zhang et al., 2022). Second, compared to the SM downscaling method developed for MODIS data (Im et al., 2016; Long et al., 2019), we selected the Landsat dataset with a finer scale as the input and improved the spatial resolution of the target SM while mitigating the scale mismatch between

the remote sensing observations and ground measurements (Zappa et al., 2020). Finally, our model used all available ISMN sites distributed in the U.S., covering different surface conditions, to provide a stronger universality and transferability for future SM retrievals and downscaling using more optical images with similar spectral ranges and spatial resolutions.

However, our study also has some limitations. First, there are correlations between auxiliary variables in the modeling process, as well as among SM products, and these correlations increase the model complexity and computational cost (Gregorutti et al., 2017). Second, errors are inevitable due to the uncertainties introduced by the four SM products at coarse spatial resolutions and the mismatched spatial scales with in situ measurements (Peng and Loew 2017). Finally, under the influence of cloud contamination on optical/thermal sensors, the temporal resolution of the downscaled SM is poor, making it impossible to analyze temporal heterogeneities within a substudy area.

## 6. Conclusions

This study proposed a framework (RF-SM) for high-resolution SM retrievals based on the RF algorithm and data-fusion strategies from ISMN in-situ SM measurements, Landsat 8 OLI and TIRS imageries, SM products (SMAP, SMOS, NCA-LDAS and ESA CCI), and auxiliary datasets including soil properties (clay, sand and silt) and meteorological reanalysis (RH, T2m, Wind and Precipitation). At both the site and field scales, the RF-SM model provided acceptable SM prediction results. The main conclusions of the study are summarized as follows:

(1) Based on multiple input parameters, the RF and the three other ML algorithms were trained on 234 sites and validated on 100 sites. The RF showed the best performance, with an RMSE value of 0.063 $m^3/m^3$ and a KGE value of 0.69 at the validation sites. Moreover, we obtained SM patterns with a spatial resolution of 30 m in five substudy areas, reflecting clear surface texture information.
(2) The performance of the RF-SM under the six land cover types of barren, cropland, forest, grassland, savanna and shrubland was better than those of the four existing SM products, indicating that by merging multiple datasets and auxiliary variables, the framework developed herein had a good prediction ability under complex surface conditions. On the studied timescale, the RF-SM also exhibited the ability to capture SM trends and precipitation events better than any single product.
(3) The Landsat dataset accounted for 14.9% of the feature importance in the RF-SM, and meteorological stress and soil properties as auxiliary variables accounted for 21.4%. The SM products account for the largest proportion (63.7%), especially NCA-LDAS and ESA CCI, which provided relatively rich SM disaggregation information with a relatively small grid range.

In summary, our study maximized the potential use of multiple satellite datasets and opened up broad application prospects for 30-m × 30-m SM retrievals in agricultural drought monitoring and water resource management at the field scale. With the further development of satellite observation technologies, scale transformation models and artificial intelligence algorithms, it will be possible to produce temporally continuous SM products covering larger regions and at finer scales in the near future.

## CRediT authorship contribution statement

**Jing Ning:** Software, Data curation, Validation, Writing – original draft. **Yunjun Yao:** Resources, Conceptualization, Data curation, Funding acquisition, Writing – review & editing. **Qingxin Tang:** Resources, Data curation, Formal analysis, Writing – review & editing. **Yufu Li:** Resources, Conceptualization, Data curation, Funding acquisition, Writing – review & editing. **Joshua B. Fisher:** Formal analysis, Writing – review & editing. **Xiaotong Zhang:** Formal analysis, Writing – review & editing. **Kun Jia:** Formal analysis, Writing – review & editing. **Jia Xu:** Software, Methodology, Investigation. **Ke Shang:** Software, Methodology, Investigation. **Junming Yang:** Software, Methodology, Investigation. **Ruiyang Yu:** Software, Methodology, Investigation. **Lu Liu:** Software, Methodology, Investigation. **Xueyi Zhang:** Software, Methodology, Investigation. **Zijing Xie:** Software, Investigation, Visualization. **Jiahui Fan:** Software, Investigation, Visualization.

## Declaration of Competing Interest

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.

## Data availability

Data will be made available on request.

## Acknowledgments

This work was supported by the Natural Science Fund of China (No.42192581 and No. 42192580 and No. 42171310). The in situ SM dataset was provided by International Soil Moisture Network (https://ismn.geo.tuwien.ac.at, accessed on Oct. 13th, 2021). Landsat 8 optical and thermal observations provided by United States Geological Survey were obtained online (https://earthexplorer.usgs.gov, accessed on Apr. 9th, 2021). The SoilGrids V.2.0 dataset were obtained from ISRIC (https://soilgrids.org, accessed on Mar. 5th, 2022). The ERA5 reanalysis dataset were obtained from Copernicus Climate Change Service (https://cds.climate.copernicus.eu, accessed on Dec. 7th, 2021). SMAP product provided by National Snow and Ice Data Center were obtained online (https://nsidc.org/data/smap, accessed on Dec. 12th, 2021). SMOS product provided by Centre Aval de Traitement des Données SMOS were obtained online (https://www.catds.fr, accessed on Jan. 1th, 2022). ESA CCI product provided by European Space Agency were obtained online (https://www.esa-soilmoisture-cci.org, accessed on Jan. 9th, 2022). NCA-LDAS product provided by NASA were obtained online (https://disc.gsfc.nasa.gov, accessed on Dec. 26th, 2021).

## References

Abowarda, A.S., Bai, L., Zhang, C., Long, D., Li, X., Huang, Q., Sun, Z., 2021. Generating surface soil moisture at 30 m spatial resolution using both data fusion and machine learning toward better water resources management at the field scale. Remote Sens. Environ. 255, 112301.
Acharya, T.D., Yang, I., 2015. Exploring landsat 8. Int. J. IT Eng. Appl. Sci. Res. (IJIEASR) 4, 4–10.
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, 3223.
Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., Notarnicola, C., 2015. Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. Remote Sens. (Basel) 7, 16398–16421.
Alizadeh, M.R., Nikoo, M.R., 2018. A fusion-based methodology for meteorological drought estimation using remote sensing data. Remote Sens. Environ. 211, 229–247.
Altman, N., Krzywinski, M., 2017. Ensemble methods: bagging and random forests. Nat. Methods 14 (10), 933–934.
Al-Yaari, A., Wigneron, J.-P., Dorigo, W., Colliander, A., Pellarin, T., Hahn, S., Mialon, A., Richaume, P., Fernandez-Moran, R., Fan, L., Kerr, Y.H., De Lannoy, G., 2019. Assessment and inter-comparison of recently developed/reprocessed microwave satellite soil moisture products using ISMN ground-based measurements. Remote Sens. Environ. 224, 289–303.
Araújo, D.C.D.S., Montenegro, S.M.G.L., Gusmão, A.C.V.E.L., Rodrigues, D.F.B., 2018. Validation of SMOS-IC Soil Moisture over Brazilian Semiarid Using in situ Measurements. EPiC Series in Engineering 3, 54–61.
Attila, N., Riczu, P., Gálya, B., Tamás, J., 2014. Spectral estimation of soil water content in visible and near infra-red range. Eurasian J. Soil Sci. 3, 163–171.
Baldocchi, D.D., Keeney, N., Rey-Sanchez, C., Fisher, J.B., 2022. Atmospheric humidity deficits tell us how soil moisture deficits down-regulate ecosystem evaporation. Adv. Water Resour. 159, 104100.

Baroni, G., Ortuani, B., Facchi, A., Gandolfi, C., 2013. The role of vegetation and soil properties on the spatio-temporal variability of the surface soil moisture in a maize-cropped field. J. Hydrol. 489, 148–159.
Barutçuoğlu, Z., & Alpaydın, E. (2003). A comparison of model aggregation methods for regression. *Artificial Neural Networks and Neural Information Processing—ICANN/ICONIP 2003* (pp. 76-83): Springer.
Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31.
Berg, A., Sheffield, J., 2018. Climate change and drought: the soil moisture perspective. Current Clim. Change Reports 4 (2), 180–191.
Biau, G., Scornet, E., 2016. A random forest guided tour. TEST 25 (2), 197–227.
Bogena, H.R., Huisman, J.A., Oberdörster, C., Vereecken, H., 2007. Evaluation of a low-cost soil water content sensor for wireless network applications. J. Hydrol. 344 (1-2), 32–42.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Brocca, L., Ciabatta, L., Massari, C., Camici, S., Tarpanelli, A., 2017. Soil moisture for hydrological applications: Open questions and new opportunities. Water 9, 140.
Bühlmann, P., 2012. Bagging, boosting and ensemble methods. *Handbook of computational statistics*. In: Gentle, J.E., Härdle, W.K., Mori, Y. (Eds.), Handbook of Computational Statistics. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 985–1022.
Cai, Y.u., Zheng, W., Zhang, X., Zhangzhong, L., Xue, X., Huk, M., 2019. Research on soil moisture prediction model based on deep learning. PLoS One 14 (4), e0214508.
Carlson, T.N., Gillies, R.R., Perry, E.M., 1994. A method to make use of thermal infrared temperature and NDVI measurements to infer surface soil water content and fractional vegetation cover. Remote Sens. Rev. 9 (1-2), 161–173.
Chan, S.K., Bindlish, R., O'Neill, P., Jackson, T., Njoku, E., Dunbar, S., Chaubell, J., Piepmeier, J., Yueh, S., Entekhabi, D., Colliander, A., Chen, F., Cosh, M.H., Caldwell, T., Walker, J., Berg, A., McNairn, H., Thibeault, M., Martínez-Fernández, J., Uldall, F., Seyfried, M., Bosch, D., Starks, P., Holifield Collins, C., Prueger, J., van der Velde, R., Asanuma, J., Palecki, M., Small, E.E., Zreda, M., Calvet, J., Crow, W.T., Kerr, Y., 2018. Development and assessment of the SMAP enhanced passive soil moisture product. Remote Sens. Environ. 204, 931–941.
Chaney, N.W., Roundy, J.K., Herrera-Estrada, J.E., Wood, E.F., 2015. High-resolution modeling of the spatial heterogeneity of soil moisture: Applications in network design. Water Resour. Res. 51 (1), 619–638.
Chauhan, N.S., Miller, S., Ardanuy, P., 2003. Spaceborne soil moisture estimation at high resolution: A microwave-optical/IR synergistic approach. Int. J. Remote Sens. 24 (22), 4599–4622.
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.
Coleman, M.L., Niemann, J.D., 2013. Controls on topographic dependence and temporal instability in catchment-scale soil moisture patterns. Water Resour. Res. 49 (3), 1625–1642.
Colliander, A., Fisher, J.B., Halverson, G., Merlin, O., Misra, S., Bindlish, R., Jackson, T. J., Yueh, S., 2017. Spatial downscaling of SMAP soil moisture using MODIS land surface temperature and NDVI during SMAPVEX15. IEEE Geosci. Remote Sens. Lett. 14 (11), 2107–2111.
Cui, C., Xu, J., Zeng, J., Chen, K.-S., Bai, X., Lu, H., Chen, Q., Zhao, T., 2017. Soil moisture mapping from satellites: An intercomparison of SMAP, SMOS, FY3B, AMSR2, and ESA CCI over two dense network regions at different spatial scales. Remote Sens. (Basel) 10, 33.
Dorigo, W., Xaver, A., Vreugdenhil, M., Gruber, A., Hegyiova, A., Sanchis-Dufau, A., Zamojski, D., Cordes, C., Wagner, W., & Drusch, M. (2013). Global automated quality control of in situ soil moisture data from the International Soil Moisture Network. *Vadose Zone Journal, 12*.
Dorigo, W.A., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Xaver, A., Gruber, A., Drusch, M., Mecklenburg, S., van Oevelen, P., Robock, A., Jackson, T., 2011. The International Soil Moisture Network: a data hosting facility for global in situ soil moisture measurements. Hydrol. Earth Syst. Sci. 15 (5), 1675–1698.
Dorigo, W.A., Gruber, A., De Jeu, R.A.M., Wagner, W., Stacke, T., Loew, A., Albergel, C., Brocca, L., Chung, D., Parinussa, R.M., Kidd, R., 2015. Evaluation of the ESA CCI soil moisture product using ground-based observations. Remote Sens. Environ. 162, 380–395.
Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., Haas, E., Hamer, P.D., Hirschi, M., Ikonen, J., de Jeu, R., Kidd, R., Lahoz, W., Liu, Y.Y., Miralles, D., Mistelbauer, T., Nicolai-Shaw, N., Parinussa, R., Pratola, C., Reimer, C., van der Schalie, R., Seneviratne, S.I., Smolander, T., Lecomte, P., 2017. ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions. Remote Sens. Environ. 203, 185–215.
Entekhabi, D., Rodriguez-Iturbe, I., Castelli, F., 1996. Mutual interaction of soil moisture state and atmospheric processes. J. Hydrol. 184 (1-2), 3–17.
Fernandez-Moran, R., Wigneron, J.-P., De Lannoy, G., Lopez-Baeza, E., Parrens, M., Mialon, A., Mahmoodi, A., Al-Yaari, A., Bircher, S., Al Bitar, A., Richaume, P., Kerr, Y., 2017. A new calibration of the effective scattering albedo and soil roughness parameters in the SMOS SM retrieval algorithm. Int. J. Appl. Earth Observ. Geoinformation 62, 27–38.
Ferreira, A.J., Figueiredo, M.A., 2012. Boosting algorithms: A review of methods, theory, and applications. Ensemble Mach. Learn. 35–85.
Foga, S., Scaramuzza, P.L., Guo, S., Zhu, Z., Dilley Jr, R.D., Beckmann, T., Schmidt, G.L., Dwyer, J.L., Hughes, M.J., Laue, B., 2017. Cloud detection algorithm comparison and validation for operational Landsat data products. Remote Sens. Environ. 194, 379–390.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.
Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38 (4), 367–378.
Garrido, F., Ghodrati, M., Chendorain, M., 1999. Small-scale measurement of soil water content using a fiber optic sensor. Soil Sci. Soc. Am. J. 63 (6), 1505–1512.
Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Extremely randomized trees. *Machine learning* 63 (1), 3–42.
Ghulam, A., Qin, Q., Teyip, T., Li, Z.-L., 2007. Modified perpendicular drought index (MPDI): a real-time drought monitoring method. ISPRS J. Photogramm. Remote Sens. 62 (2), 150–164.
Goldstein, B.A., Polley, E.C., Briggs, F.B., 2011. Random forests for genetic association studies. In: Statistical applications in genetics and molecular biology, p. 10.
Gregorutti, B., Michel, B., Saint-Pierre, P., 2017. Correlation and variable importance in random forests. Stat. Comput. 27 (3), 659–678.
Gruber, A., Dorigo, W.A., Zwieback, S., Xaver, A., Wagner, W., 2013. Characterizing coarse-scale representativeness of in situ soil moisture measurements from the International Soil Moisture Network. Vadose Zone J. 12 (vzj2012), 0170.
Gruber, A., Scanlon, T., van der Schalie, R., Wagner, W., Dorigo, W., 2019. Evolution of the ESA CCI Soil Moisture climate data records and their underlying merging methodology. Earth Syst. Sci. Data 11, 717–739.
Guevara, M., Vargas, R., 2019. Downscaling satellite soil moisture using geomorphometry and machine learning. PLoS One 14 (9), e0219639.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377 (1-2), 80–91.
Hallikainen, M., Ulaby, F., Dobson, M., El-rayes, M., Wu, L.-K., 1985. Microwave dielectric behavior of wet soil-part 1: Empirical models and experimental observations. IEEE Trans. Geosci. Remote Sens. GE-23 (1), 25–34.
Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., Bond-Lamberty, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One 12 (2), e0169748.
Hersbach, H., Bell, B., Berrisford, P., Horányi, A., Sabater, J.M., Nicolas, J., Radu, R., Schepers, D., Simmons, A., Soci, C., 2019. Global reanalysis: goodbye ERA-Interim, hello ERA5. ECMWF newsletter 159, 17–24.
Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.-N., 2020. The ERA5 global reanalysis. Q. J. R. Meteorolog. Soc. 146 (730), 1999–2049.
Hesterberg, T., 2011. Bootstrap. Wiley Interdiscip. Rev. Comput. Stat. 3 (6), 497–526.
Hillel, D. (1998). Environmental soil physics.
Holzman, M.E., Carmona, F., Rivas, R., Niclòs, R., 2018. Early assessment of crop yield from remotely sensed water stress and solar radiation data. ISPRS J. Photogramm. Remote Sens. 145, 297–308.
Im, J., Park, S., Rhee, J., Baik, J., Choi, M., 2016. Downscaling of AMSR-E soil moisture with MODIS products using machine learning approaches. Environ. Earth Sci. 75, 1–19.
Irons, J.R., Dwyer, J.L., Barsi, J.A., 2012. The next Landsat satellite: The Landsat data continuity mission. Remote Sens. Environ. 122, 11–21.
Jasinski, M.F., Borak, J.S., Kumar, S.V., Mocko, D.M., Peters-Lidard, C.D., Rodell, M., Rui, H., Beaudoing, H.K., Vollmer, B.E., Arsenault, K.R., Li, B., Bolten, J.D., Tangdamrongsub, N., 2019. NCA-LDAS: overview and analysis of hydrologic trends for the national climate assessment. J. Hydrometeorol. 20 (8), 1595–1617.
Johnson, R., Zhang, T., 2013. Learning nonlinear functions using regularized greedy forest. IEEE Trans. Pattern Anal. Mach. Intell. 36 (5), 942–954.
Kang, J., Jin, R., Li, X., Ma, C., Qin, J., Zhang, Y., 2017. High spatio-temporal resolution mapping of soil moisture by integrating wireless sensor network observations and MODIS apparent thermal inertia in the Babao River Basin, China. Remote Sens. Environ. 191, 232–245.
Kerr, Y.H., Waldteufel, P., Wigneron, J.-P., Martinuzzi, J., Font, J., Berger, M., 2001. Soil moisture retrieval from space: The Soil Moisture and Ocean Salinity (SMOS) mission. IEEE transactions on Geoscience and remote sensing 39, 1729–1735.
Kerr, Y.H., Waldteufel, P., Richaume, P., Wigneron, J.P., Ferrazzoli, P., Mahmoodi, A., Al Bitar, A., Cabot, F., Gruhier, C., Juglea, S.E., Leroux, D., Mialon, A., Delwart, S., 2012. The SMOS soil moisture retrieval algorithm. IEEE Trans. Geosci. Remote Sens. 50 (5), 1384–1403.
Kerr, Y.H., Al-Yaari, A., Rodriguez-Fernandez, N., Parrens, M., Molero, B., Leroux, D., Bircher, S., Mahmoodi, A., Mialon, A., Richaume, P., Delwart, S., Al Bitar, A., Pellarin, T., Bindlish, R., Jackson, T.J., Rüdiger, C., Waldteufel, P., Mecklenburg, S., Wigneron, J.-P., 2016. Overview of SMOS performance in terms of global soil moisture monitoring after six years in operation. Remote Sens. Environ. 180, 40–63.
Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A., 2011. Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 41 (3), 552–568.
Konings, A.G., Entekhabi, D., Chan, S.K., Njoku, E.G., 2011. Effect of radiative transfer uncertainty on L-band radiometric soil moisture retrieval. IEEE Trans. Geosci. Remote Sens. 49 (7), 2686–2698.

Kornelsen, K.C., Cosh, M.H., Coulibaly, P., 2015. Potential of bias correction for downscaling passive microwave and soil moisture data. J. Geophys. Res. Atmos. 120 (13), 6460–6479.
Kumar, S.V., Peters-Lidard, C.D., Mocko, D., Reichle, R., Liu, Y., Arsenault, K.R., Xia, Y., Ek, M., Riggs, G., Livneh, B., Cosh, M., 2014. Assimilation of remotely sensed soil moisture and snow depth retrievals for drought estimation. J. Hydrometeorol. 15 (6), 2446–2469.
Kumar, S.V., Jasinski, M., Mocko, D.M., Rodell, M., Borak, J., Li, B., Beaudoing, H.K., Peters-Lidard, C.D., 2019. NCA-LDAS land analysis: Development and performance of a multisensor, multivariate land data assimilation system for the National Climate Assessment. J. Hydrometeorol. 20, 1571–1593.
Li, D., Jin, R., Zhou, J., Kang, J., 2015. Analysis and reduction of the uncertainties in soil moisture estimation with the L-MEB model using EFAST and ensemble retrieval. IEEE Geosci. Remote Sens. Lett. 12, 1337–1341.
Li, Z.-L., Leng, P., Zhou, C., Chen, K.-S., Zhou, F.-C., Shang, G.-F., 2021. Soil moisture retrieval from remote sensing measurements: Current knowledge and directions for the future. Earth Sci. Rev. 218, 103673.
Liao, W., Wang, D., Wang, G., Xia, Y., Liu, X., 2019. Quality control and evaluation of the observed daily data in the north american soil moisture database. J. Meteorolog. Res. 33 (3), 501–518.
Lievens, H., Tomer, S.K., Al Bitar, A., De Lannoy, G.J.M., Drusch, M., Dumedah, G., Hendricks Franssen, H.-J., Kerr, Y.H., Martens, B., Pan, M., Roundy, J.K., Vereecken, H., Walker, J.P., Wood, E.F., Verhoest, N.E.C., Pauwels, V.R.N., 2015. SMOS soil moisture assimilation for improved hydrologic simulation in the Murray Darling Basin, Australia. Remote Sens. Environ. 168, 146–162.
Liu, Y., Dorigo, W.A., Parinussa, R., de Jeu, R.A., Wagner, W., McCabe, M.F., Evans, J., Van Dijk, A., 2012a. Trend-preserving blending of passive and active microwave soil moisture retrievals. Remote Sens. Environ. 123, 280–297.
Liu, Y., Wang, Y., Zhang, J., 2012b. New machine learning algorithm: Random forest. In: International Conference on Information Computing and Applications. Springer, pp. 246–252.
Long, D.i., Bai, L., Yan, L.a., Zhang, C., Yang, W., Lei, H., Quan, J., Meng, X., Shi, C., 2019. Generation of spatially complete and daily continuous surface soil moisture of high spatial resolution. Remote Sens. Environ. 233, 111364.
Lu, Y., Horton, R., Zhang, X., Ren, T., 2018. Accounting for soil porosity improves a thermal inertia model for estimating surface soil water content. Remote Sens. Environ. 212, 79–89.
Ma, H., Zeng, J., Chen, N., Zhang, X., Cosh, M.H., Wang, W., 2019. Satellite surface soil moisture from SMAP, SMOS, AMSR2 and ESA CCI: A comprehensive assessment using global ground-based observations. Remote Sens. Environ. 231, 111215.
Martínez-Fernández, J., González-Zamora, A., Sánchez, N., Gumuzzio, A., Herrero-Jiménez, C., 2016. Satellite soil moisture for agricultural drought monitoring: Assessment of the SMOS derived Soil Water Deficit Index. Remote Sens. Environ. 177, 277–286.
Merlin, O., Chehbouni, A.G., Kerr, Y.H., Njoku, E.G., Entekhabi, D., 2005. A combined modeling and multispectral/multiresolution remote sensing approach for disaggregation of surface soil moisture: Application to SMOS configuration. IEEE Trans. Geosci. Remote Sens. 43 (9), 2036–2050.
Merlin, O., Chehbouni, A., Walker, J.P., Panciera, R., Kerr, Y.H., 2008. A simple method to disaggregate passive microwave-based soil moisture. IEEE Trans. Geosci. Remote Sens. 46 (3), 786–796.
Merlin, O., Al Bitar, A., Walker, J.P., Kerr, Y., 2009. A sequential model for disaggregating near-surface soil moisture observations using multi-resolution thermal sensors. Remote Sens. Environ. 113 (10), 2275–2284.
Merlin, O., Al Bitar, A., Walker, J.P., Kerr, Y., 2010. An improved algorithm for disaggregating microwave-derived soil moisture based on red, near-infrared and thermal-infrared data. Remote Sens. Environ. 114 (10), 2305–2316.
Miralles, D.G., Crow, W.T., Cosh, M.H., 2010. Estimating spatial sampling errors in coarse-scale soil moisture estimates derived from point-scale observations. J. Hydrometeorol. 11, 1423–1429.
Mohanty, B.P., Cosh, M.H., Lakshmi, V., Montzka, C., 2017. Soil moisture remote sensing: State-of-the-science. Vadose Zone J. 16, 1–9.
Ngo Thi, D., Ha, N.T.T., Tran Dang, Q., Koike, K., Mai Trong, N., 2019. Effective band ratio of landsat 8 images based on VNIR-SWIR reflectance spectra of topsoils for soil moisture mapping in a tropical region. Remote Sens. (Basel) 11, 716.
Njoku, E.G., Jackson, T.J., Lakshmi, V., Chan, T.K., Nghiem, S.V., 2003. Soil moisture retrieval from AMSR-E. IEEE Trans. Geosci. Remote Sens. 41 (2), 215–229.
O'Neill, P., Entekhabi, D., Njoku, E., Kellogg, K., 2010. The NASA soil moisture active passive (SMAP) mission: Overview. In: 2010 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 3236–3239.
Parida, B.R., Collado, W.B., Borah, R., Hazarika, M.K., Samarakoon, L., 2008. Detecting drought-prone areas of rice agriculture using a MODIS-derived soil moisture index. GIScience & Remote Sensing 45 (1), 109–129.
Parinussa, R.M., Meesters, A.G.C.A., Liu, Y.Y., Dorigo, W., Wagner, W., de Jeu, R.A.M., 2011. Error estimates for near-real-time satellite soil moisture as derived from the land parameter retrieval model. IEEE Geosci. Remote Sens. Lett. 8 (4), 779–783.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Peng, J., Albergel, C., Balenzano, A., Brocca, L., Cartus, O., Cosh, M.H., Crow, W.T., Dabrowska-Zielinska, K., Dadson, S., Davidson, M.W.J., de Rosnay, P., Dorigo, W., Gruber, A., Hagemann, S., Hirschi, M., Kerr, Y.H., Lovergine, F., Mahecha, M.D., Marzahn, P., Mattia, F., Musial, J.P., Preuschmann, S., Reichle, R.H., Satalino, G., Silgram, M., van Bodegom, P.M., Verhoest, N.E.C., Wagner, W., Walker, J.P., Wegmüller, U., Loew, A., 2021. A roadmap for high-resolution satellite soil moisture applications–confronting product characteristics with user requirements. Remote Sens. Environ. 252, 112162.
Peng, J., Loew, A., 2017. Recent advances in soil moisture estimation from remote sensing. Water 9, 530.
Peng, J., Loew, A., Zhang, S., Wang, J., Niesel, J., 2015. Spatial downscaling of satellite soil moisture data using a vegetation temperature condition index. IEEE Trans. Geosci. Remote Sens. 54 (1), 558–566.
Peng, J., Loew, A., Merlin, O., Verhoest, N.E.C., 2017. A review of spatial downscaling of satellite remotely sensed soil moisture. Rev. Geophys. 55 (2), 341–366.
Poggio, L., De Sousa, L.M., Batjes, N.H., Heuvelink, G., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7, 217–240.
Price, J.C., 1985. On the analysis of thermal infrared imagery: The limited utility of apparent thermal inertia. Remote Sens. Environ. 18 (1), 59–73.
Probst, P., Wright, M.N., Boulesteix, A.L., 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery 9, e1301.
Purdy, A.J., Fisher, J.B., Goulden, M.L., Colliander, A., Halverson, G., Tu, K., Famiglietti, J.S., 2018. SMAP soil moisture improves global evapotranspiration. Remote Sens. Environ. 219, 1–14.
Quets, J., De Lannoy, G.J., Al Yaari, A., Chan, S., Cosh, M.H., Gruber, A., Reichle, R.H., Van der Schalie, R., Wigneron, J.-P., 2019. Uncertainty in soil moisture retrievals: An ensemble approach using SMOS L-band microwave data. Remote Sens. Environ. 229, 133–147.
Rahman, M.S., Di, L., Shrestha, R., Eugene, G.Y., Lin, L., Zhang, C., Hu, L., Tang, J., & Yang, Z. (2017). Agriculture flood mapping with Soil Moisture Active Passive (SMAP) data: A case of 2016 Louisiana flood. In, *2017 6th International Conference on Agro-Geoinformatics* (pp. 1-6): IEEE.
Ranney, K.J., Niemann, J.D., Lehman, B.M., Green, T.R., Jones, A.S., 2015. A method to downscale soil moisture to fine resolutions using topographic, vegetation, and soil data. Adv. Water Resour. 76, 81–96.
Reichle, R.H., Entekhabi, D., McLaughlin, D.B., 2001. Downscaling of radio brightness measurements for soil moisture estimation: A four-dimensional variational data assimilation approach. Water Resour. Res. 37 (9), 2353–2364.
Reichle, R.H., Koster, R.D., 2005. Global assimilation of satellite surface soil moisture retrievals into the NASA Catchment land surface model. Geophys. Res. Lett. 32.
Reichle, R.H., Koster, R.D., Dong, J., Berg, A.A., 2004. Global soil moisture from satellite observations, land surface models, and ground data: Implications for data assimilation. J. Hydrometeorol. 5 (3), 430–442.
Robinson, D.A., Campbell, C.S., Hopmans, J.W., Hornbuckle, B.K., Jones, S.B., Knight, R., Ogden, F., Selker, J., Wendroth, O., 2008. Soil moisture measurement for ecological and hydrological watershed-scale observatories: A review. Vadose Zone J. 7 (1), 358–389.
Robock, A., Vinnikov, K.Y., Srinivasan, G., Entin, J.K., Hollinger, S.E., Speranskaya, N.A., Liu, S., Namkhai, A., 2000. The global soil moisture data bank. Bull. Am. Meteorol. Soc. 81 (6), 1281–1299.
Rodell, M., Houser, P.R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J.K., Walker, J.P., Lohmann, D., Toll, D., 2004. The global land data assimilation system. Bull. Am. Meteorol. Soc. 85 (3), 381–394.
Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J. Photogramm. Remote Sens. 67, 93–104.
Sabaghy, S., Walker, J.P., Renzullo, L.J., Jackson, T.J., 2018. Spatially enhanced passive microwave derived soil moisture: Capabilities and opportunities. Remote Sens. Environ. 209, 551–580.
Sadeghi, M., Ebtehaj, A., Crow, W.T., Gao, L., Purdy, A.J., Fisher, J.B., Jones, S.B., Babaeian, E., Tuller, M., 2020. Global estimates of land surface water fluxes from SMOS and SMAP satellite soil moisture data. J. Hydrometeorol. 21, 241–253.
Sahoo, A.K., De Lannoy, G.J., Reichle, R.H., Houser, P.R., 2013. Assimilation and downscaling of satellite observed soil moisture over the Little River Experimental Watershed in Georgia, USA. Adv. Water Resour. 52, 19–33.
Sehler, R., Li, J., Reager, J.T., Ye, H., 2019. Investigating relationship between soil moisture and precipitation globally using remote sensing observations. J. Contemporary Water Res. Educ. 168 (1), 106–118.
Seneviratne, S.I., Corti, T., Davin, E.L., Hirschi, M., Jaeger, E.B., Lehner, I., Orlowsky, B., Teuling, A.J., 2010. Investigating soil moisture–climate interactions in a changing climate: A review. Earth Sci. Rev. 99 (3-4), 125–161.
Shin, Y., Mohanty, B.P., 2013. Development of a deterministic downscaling algorithm for remote sensing soil moisture footprint using soil and vegetation classifications. Water Resour. Res. 49 (10), 6208–6228.
Tarek, M., Brissette, F.P., Arsenault, R., 2020. Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America. Hydrol. Earth Syst. Sci. 24, 2527–2544.
Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. Atmos. 106 (D7), 7183–7192.
Tomer, S.K., Al Bitar, A., Sekhar, M., Zribi, M., Bandyopadhyay, S., Kerr, Y., 2016. MAPSM: A spatio-temporal algorithm for merging soil moisture from active and passive microwave remote sensing. Remote Sens. (Basel) 8, 990.
Topp, G.C., Reynolds, W.D., 1998. Time domain reflectometry: A seminal technique for measuring mass and energy in soil. Soil Tillage Res. 47, 125–132.
Valente, A., Morais, R., Tuli, A., Hopmans, J.W., Kluitenberg, G.J., 2006. Multi-functional probe for small-scale simultaneous measurements of soil thermal properties, water content, and electrical conductivity. Sens. Actuators, A 132 (1), 70–77.

van der Molen, M.K., Dolman, A.J., Ciais, P., Eglin, T., Gobron, N., Law, B.E., Meir, P., Peters, W., Phillips, O.L., Reichstein, M., Chen, T., Dekker, S.C., Doubková, M., Friedl, M.A., Jung, M., van den Hurk, B.J.J.M., de Jeu, R.A.M., Kruijt, B., Ohta, T., Rebel, K.T., Plummer, S., Seneviratne, S.I., Sitch, S., Teuling, A.J., van der Werf, G. R., Wang, G., 2011. Drought and ecosystem carbon cycling. Agric. For. Meteorol. 151 (7), 765–773.

Vergopolan, N., Chaney, N.W., Beck, H.E., Pan, M., Sheffield, J., Chan, S., Wood, E.F., 2020. Combining hyper-resolution land surface modeling with SMAP brightness temperatures to obtain 30-m soil moisture estimates. Remote Sens. Environ. 242, 111740.

Verrelst, J., Camps-Valls, G., Muñoz-Marí, J., Rivera, J.P., Veroustraete, F., Clevers, J.G., Moreno, J., 2015. Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties–A review. ISPRS J. Photogramm. Remote Sens. 108, 273–290.

Wang, Y., Feng, D., Li, D., Chen, X., Zhao, Y., Niu, X., 2016. A mobile recommendation system based on logistic regression and gradient boosting decision trees. In: 2016 international joint conference on neural networks (IJCNN). IEEE, pp. 1896–1902.

Wang, X., Xie, H., Guan, H., Zhou, X., 2007. Different responses of MODIS-derived NDVI to root-zone soil moisture in semi-arid and humid regions. J. Hydrol. 340 (1-2), 12–24.

Wigneron, J.-P., Mialon, A., De Lannoy, G., Fernández-Morán, R., Al-Yaari, A., Ebrahimi, M., Rodriguez-Fernandez, N., Kerr, Y., Quets, J., Pellarin, T., 2018. SMOS-IC: Current status and overview of soil moisture and VOD applications. In: IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 1451–1453.

Wigneron, J.-P., Li, X., Frappart, F., Fan, L., Al-Yaari, A., De Lannoy, G., Liu, X., Wang, M., Le Masson, E., Moisy, C., 2021. SMOS-IC data record of soil moisture and L-VOD: Historical development, applications and perspectives. Remote Sens. Environ. 254, 112238.

Yue, J., Tian, J., Tian, Q., Xu, K., Xu, N., 2019. Development of soil moisture indices from differences in water absorption between shortwave-infrared bands. ISPRS J. Photogramm. Remote Sens. 154, 216–230.

Zappa, L., Woods, M., Hemment, D., Xaver, A., Dorigo, W., 2020. Evaluation of remotely sensed soil moisture products using crowdsourced measurements. In: Eighth international conference on remote sensing and geoinformation of the environment (RSCy2020). SPIE, pp. 660–672.

Zhang, R., Kim, S., Sharma, A., 2019. A comprehensive validation of the SMAP Enhanced Level-3 Soil Moisture product using ground measurements over varied climates and landscapes. Remote Sens. Environ. 223, 82–94.

Zhang, R., Kim, S., Sharma, A., Lakshmi, V., 2021. Identifying relative strengths of SMAP, SMOS-IC, and ASCAT to capture temporal variability. Remote Sens. Environ. 252, 112126.

Zhang, Y., Liang, S., Zhu, Z., Ma, H., He, T., 2022. Soil moisture content retrieval from Landsat 8 data using ensemble learning. ISPRS J. Photogramm. Remote Sens. 185, 32–47.

Zhao, W., Sánchez, N., Lu, H., Li, A., 2018. A spatial downscaling approach for the SMAP passive surface soil moisture product using random forest regression. J. Hydrol. 563, 1009–1024.

Zhou, Q.Y., Shimada, J., Sato, A., 2001. Three-dimensional spatial and temporal monitoring of soil water content using electrical resistivity tomography. Water Resour. Res. 37 (2), 273–285.