land

MDPI

Article

A Framework for High-Spatiotemporal-Resolution Soil Moisture Retrieval in China Using Multi-Source Remote Sensing Data

Zhuangzhuang Feng [1,2], Xingming Zheng [1,3,*], Xiaofeng Li [1,3], Chunmei Wang [4], Jinfeng Song [1,2], Lei Li [1], Tianhao Guo [1,2] and Jia Zheng [1,2]

[1] State Key Laboratory of Black Soils Conservation and Utilization, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun 130102, China; fengzhuangzhuang@iga.ac.cn (Z.F.); lixiaofeng@iga.ac.cn (X.L.); songjinfeng23@mails.ucas.ac.cn (J.S.); lilei@iga.ac.cn (L.L.); guotianhao@iga.ac.cn (T.G.); zhengjia@iga.ac.cn (J.Z.)
[2] College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
[3] Changchun Jingyuetan Remote Sensing Experiment Station, Chinese Academy of Sciences, Changchun 130102, China
[4] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; wangcm@aircas.ac.cn
* Correspondence: zhengxingming@iga.ac.cn; Tel.: +86-0431-8554-2227

Abstract: High-spatiotemporal-resolution and accurate soil moisture (SM) data are crucial for investigating climate, hydrology, and agriculture. Existing SM products do not yet meet the demands for high spatiotemporal resolution. The objective is to develop and evaluate a retrieval framework to derive SM estimates with high spatial (100 m) and temporal (<3 days) resolution that can be used on a national scale in China. Therefore, this study integrates multi-source data, including optical remote sensing (RS) data from Sentinel-2 and Landsat-7/8/9, synthetic aperture radar (SAR) data from Sentinel-1, and auxiliary data. Four machine learning and deep learning algorithms are applied, including Random Forest Regression (RFR), Extreme Gradient Boosting (XGBoost), Long Short-Term Memory (LSTM) networks, and Ensemble Learning (EL). The integrated framework (IF) considers three feature scenarios (SC1: optical RS + auxiliary data, SC2: SAR + auxiliary data, SC3: optical RS + SAR + auxiliary data), encompassing a total of 33 features. The results are as follows: (1) The correlation coefficients (r) between auxiliary data (such as sand fraction, r = −0.48; silt fraction, r = 0.47; and evapotranspiration, r = −0.42), SAR features (such as the backscatter coefficients for VV-pol (σ^0_{vv}), r = 0.47), and optical RS features (such as Shortwave Infrared Band 2 (SWIR2) reflectance data from Sentinel-2 and Landsat-7/8/9, r = −0.39) with observed SM are significant. This indicates that multi-source data can provide complementary information for SM monitoring. (2) Compared to XGBoost and LSTM, RFR and EL demonstrate superior overall performance and are the preferred models for SM prediction. Their R^2 for the training and test sets exceed 0.969 and 0.743, respectively, and their ubRMSE are below 0.022 and 0.063 m^3/m^3, respectively. (3) The SM prediction accuracy is highest for the scenario of optical + SAR + auxiliary data, followed by SAR + auxiliary data, and finally optical + auxiliary data. (4) With an increasing Normalized Difference Vegetation Index (NDVI) and SM values, the trained models exhibit a general decrease in prediction performance and accuracy. (5) In 2021 and 2022, without considering cloud cover, the IF theoretically achieved an SM revisit time of 1–3 days across 95.01% and 96.53% of China's area, respectively. However, SC1 was able to achieve a revisit time of 1–3 days over 60.73% of China's area in 2021 and 69.36% in 2022, while the area covered by SC2 and SC3 at this revisit time accounted for less than 1% of China's total area. This study validates the effectiveness of combining multi-source RS data with auxiliary data in large-scale SM monitoring and provides new methods for improving SM retrieval accuracy and spatiotemporal coverage.

Keywords: remote sensing; high spatiotemporal resolution; surface soil moisture; machine learning; deep learning; China

check for updates

Citation: Feng, Z.; Zheng, X.; Li, X.; Wang, C.; Song, J.; Li, L.; Guo, T.; Zheng, J. A Framework for High-Spatiotemporal-Resolution Soil Moisture Retrieval in China Using Multi-Source Remote Sensing Data. *Land* **2024**, *13*, 2189. https://doi.org/10.3390/land13122189

Academic Editor: Paolo Nasta

Received: 30 October 2024
Revised: 6 December 2024
Accepted: 9 December 2024
Published: 15 December 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Soil moisture (SM) is a critical parameter in climate, hydrology, and agriculture, significantly influencing the water and energy exchanges between soil, plants, and the atmosphere [1]. High-spatiotemporal-resolution SM data can support refined assessments of climate impacts and significantly enhance the accuracy of regional and global climate change simulations [2]. It can also capture hydrological dynamics at various scales, aiding in the identification of localized hydrological processes and improving the development and calibration of watershed hydrological models [3]. Additionally, such data can be utilized to guide precision irrigation in agriculture and optimize water resource use [4].

Over the past two decades, the capability to measure SM across different scales (from local to global measurements) has significantly improved [5]. Currently, there are three primary methods for obtaining SM: in situ measurements, Land Surface Model (LSM) simulations, and remote sensing (RS)-based retrievals. In situ measurements provide the most accurate and continuous SM information. Numerous regional-scale SM monitoring networks have been established globally, such as the International Soil Moisture Network (ISMN) [6]. However, the deployment of dense networks is time-consuming and costly, making large-scale and high-spatial-resolution SM monitoring challenging [7,8]. SM products generated by LSMs offer high temporal and spatial consistency. For example, historical SM data from the ECMWF Reanalysis 5th Generation for Land (ERA 5-Land), Global Land Data Assimilation System with the Noah Land Surface Model (GLDAS-Noah), and Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2), are widely used in global- or regional-scale long-term climate change studies [7,9]. However, LSM-generated SM products have low spatial resolution and cannot capture the spatial heterogeneity of SM at the field scale, limiting their applicability in field-scale agricultural production management.

Satellite RS, including microwave, optical, and thermal infrared RS, is a powerful tool for monitoring SM from regional to global scales [5]. Microwave RS (both passive and active) is particularly suitable for global SM data acquisition due to its all-weather, all-time capability, and sensitivity to SM [10]. Passive microwave RS effectively monitors SM through brightness temperature (TB) measurements obtained by radiometers. Various SM products based on passive microwave satellites, such as Soil Moisture and Ocean Salinity (SMOS) [11], Soil Moisture Active Passive (SMAP) [12], Advanced Microwave Scanning Radiometer 2 (AMSR2) [13], and the European Space Agency Climate Change Initiative (ESA CCI) [14], are freely available. Notably, the ESA CCI product is a merged dataset coordinated from multiple satellite-based SM datasets, including Advanced SCATterometer (ASCAT) satellite data (active microwave) [15]. While passive microwave RS has more developed algorithms and higher temporal resolution than active microwave RS, it suffers from lower spatial resolution (25–40 km) [8]. To address this issue, the SMAP satellite was equipped with both radar and radiometer to enhance spatial resolution. Unfortunately, the radar sensor ceased operations in July 2015 [16]. Subsequently, researchers have used data fusion techniques to combine SMAP and Sentinel-1 (which substitutes for SMAP L-band radar data) to provide moderate-spatial-resolution SM products (3 km or 1 km) [17–19]. However, in agricultural research, the spatial resolution of these SM products is relatively coarse, limiting their applicability at the field and farmland scales; thus, they have not been widely used for farmland-scale agricultural management [5,20].

Synthetic aperture radar (SAR), an active microwave technology, is highly regarded for its sensitivity to SM and its high spatial resolution (1 m–100 m) [21,22]. Although SAR can provide high-resolution data, it comes at the cost of long revisit time for early SAR sensors [19]. Operational and upcoming public or commercial SAR satellites, such as the Advanced Land Observing Satellite (ALOS) series [23], Sentinel-1 series [24], GF-3 series [25], as well as Biomass [26], the NASA-ISRO Synthetic Aperture Radar (NISAR) [27], and the Radar Observation System for Europe–L-band (Rose-L) [28], are enhancing SAR's Earth observation capabilities. However, the current SAR satellite coverage remains limited, with long revisit cycles for acquiring continuous data on a global or continental scale, which

is challenging for agricultural production needs. For instance, Sentinel-1A has a revisit time of 12 days. With the addition of Sentinel-1B, the revisit time of the Sentinel-1 constellation was significantly reduced to 6 days [8]. Unfortunately, Sentinel-1B ceased operations in December 2021, which caused the revisit time of the Sentinel-1 constellation to revert to 12 days, leading to data gaps in some regions, such as Northeast China. SM retrieval using microwave RS is constrained by either coarse spatial resolution (passive microwave) or a long revisit time after the damage to Sentinel-1B (SAR). Optical RS data offer fine spatial resolution and a short revisit time, making it an effective alternative for improving the spatiotemporal monitoring frequency of SM [29,30]. Numerous studies have proven the functional relationship between the spectral characteristics of optical RS and SM [5,31–33]. However, optical RS lacks a direct physical relationship with SM, only capturing changes in the surface SM of bare soil or indirectly estimating SM status through spectral information and vegetation indices [34]. Another limitation is its susceptibility to cloud cover, which can prevent adequate optical data acquisition in certain areas, leading to fragmented satellite observations or reduced temporal resolution [5].

Due to the respective advantages and limitations of microwave and optical sensors, their synergistic use is expected to overcome these constraints [35]. Numerous studies have proven the effectiveness of integrating multi-source RS data for estimating SM. These studies primarily focus on the synergy between passive and active microwave RS [36] and the combination of microwave and optical RS [37–39]. It is important to note that most of the current research on the synergy of multi-source RS data has been conducted on a small scale (local scale) and has utilized a limited number of multi-source RS image pairs (ranging from a few to a dozen). The applicability of these findings to other regions remains questionable. Due to the coverage of RS data and limitations in revisit time, the SM products derived from these studies are challenging to effectively guide agricultural production activities at the field scale on a national level [40].

High-temporal (weekly and sub-weekly)- and spatial ($\leq$1 km)-resolution SM datasets are crucial for precise irrigation management at the farmland scale [41]. However, in practical applications, achieving daily-scale matching of multi-source RS data at a national scale in China is extremely challenging. Besides the limitations of existing sensors, weather factors also significantly hinder the synergistic monitoring of SM using multi-source RS data. Therefore, it is essential to explore a high-spatiotemporal-resolution multi-source RS SM monitoring framework suitable for different scenarios at a national scale in China. This new framework should be extendable to regional scales to produce SM products and meet the demand for SM data in precision agriculture and other fields. The proposed framework in this study designs three scenarios: SAR-optical-auxiliary data, SAR-auxiliary data, and optical-auxiliary data. These scenarios integrate multi-source RS observations from different platforms for SM estimation through machine/deep learning methods. Machine/deep learning, with its powerful nonlinear modeling capabilities, can effectively utilize complementary information from multi-source data for SM retrieval [18,29].

This study utilizes multi-source RS data, including SAR, optical, and auxiliary data, and applies four machine/deep learning algorithms to achieve accurate estimation of high-spatiotemporal-resolution SM at a national scale in China. In this context, the specific objectives of this study are as follows: (1) to develop a high-spatiotemporal-resolution multi-source RS SM retrieval framework with a spatial resolution of 100 m and a temporal resolution better than 3 days, suitable for application at a national scale in China; (2) to compare the accuracy of different machine/deep learning algorithms in predicting SM and determine the optimal model; and (3) to evaluate the potential of different scenarios of multi-source RS data in predicting SM.

2. Study Area and Datasets

2.1. Study Area and In Situ SM Measurements

The 0–5 cm surface SM data used in this study were sourced from the Soil Moisture Observation Network of Typical Ecosystems in China (SONTE-China) [42]. SONTE-China

consists of 17 sites distributed across 13 provinces in China (Figure 1). The network spans a longitude range from 81.18° E to 128.96° E and a latitude range from 23.24° N to 49.34° N, covering various landscapes such as grasslands, farmlands, deserts, and forests. Table 1 provides detailed information on the geographical locations, land cover, soil characteristics, and topographical conditions of the 17 sites. It is important to note that the land cover in the Jingyuetan and Yucheng sites is subject to human agricultural activities.



Figure 1. The spatial distribution of the SONTE-China 17 sites within the study area.

Table 1. List of 17 ground-based SM sites from SONTE-China.

Name of Sites	ID	Longitude	Latitude	Number of Nodes	Land Cover	Soil Texture	Elevation (m)
Hulunber	HLBE	119.9894	49.3320	10	Grassland	Silt loam	634
Jingyuetan	JYT	125.6222	44.7914	10	Corn	Silt loam	187
Xilinhaote	XLHT	116.3303	44.1367	10	Grassland	Sandy loam	1100
Jiangshanjiao	JSJ	128.9523	43.8554	10	Grassland	Silt loam	423
Xitianshan	XTS	81.1725	43.7435	10	Apple tree	Silt loam	733
Guyuan	GY	115.6809	41.7631	10	Grassland	Sandy loam	1384
Minqin	MQ	102.9182	38.6296	10	Shrubland	Sand	1374
Haibei	HB	101.3131	37.6108	10	Grassland	Silt	3193
Yucheng	YC	116.5700	36.8285	10	Corn-Wheat	Silt loam	23
Qingdao2	QD2	120.1727	35.9446	5	Grassland	Silt loam	4
Hefei	HF	117.1698	31.9036	10	Grassland	Silt	30
Naqu	NQ	92.0118	31.6432	10	Grassland	Loam	4593
Nanjing	NJ	119.2128	31.5020	10	Tea plant	Silt loam	18
Dongtinghu	DTH	113.1684	29.3152	10	Camellia oleifera	Silt loam	56
Qiyang	QY	111.8711	26.7600	10	Tea plant	Silt loam	150
Qianyanzhou	QYZ	115.0719	26.7451	10	Shrubland	Silt	72
Guangzhou	GZ	113.6342	23.2447	10	Lawn grass	Silt loam	22

Each site is equipped with 10 observation nodes, and each node measures SM at four depths (5 cm, 10 cm, 20 cm, and 40 cm). SM and soil temperature are automatically measured every 30 min using four Meter 5TM sensors. After calibration, the average root mean square error (RMSE) for SM at each site is 0.027 m^3/m^3 [43]. This study uses the calibrated daily average data as the true values of surface SM, with the specific measurement period illustrated in Table 2.

Table 2. Surface SM measurement period and satellite image overview for 17 SONTE-China sites.

Sites	Time Span	Number of SAR Images	Number of Optical Images	Training Set Size	Test Set Size
DTH	2020-08-29–2022-07-03	102	54	117	28
GY	2018-07-21–2022-06-26	132	121	172	72
GZ	2018-12-14–2021-11-12	79	100	110	60
HB	2019-07-25–2022-07-03	106	183	177	93
HF	2019-04-24–2022-07-03	96	223	213	87
HLBE	2019-08-29–2022-07-03	50	148	113	65
JSJ	2019-08-01–2022-07-03	104	99	119	75
JYT	2020-08-07–2022-05-26	35	101	95	34
MQ	2019-06-19–2022-06-26	181	222	232	115
NJ	2019-12-28–2022-07-03	114	157	152	107
NQ	2019-08-18–2022-06-26	145	75	123	78
QD2	2019-03-26–2022-07-03	281	212	295	148
QYZ	2019-11-08–2022-07-03	74	104	104	64
QY	2019-11-16–2022-06-27	66	67	80	46
XLHT	2019-05-18–2022-07-03	177	199	229	112
XTS	2019-08-22–2022-06-27	191	148	198	109
YC	2019-03-21–2022-07-03	147	212	234	100
ALL sites	2018-07-21–2022-07-03	2080	2425	2763	1393

2.2. Remote Sensing Data

2.2.1. Sentinel-1 Data

Sentinel-1, an Earth observation satellite by the European Space Agency, comprises two satellites: Sentinel-1A and Sentinel-1B. They use C-band SAR (5.405 GHz) to provide all-weather, all-time high-resolution imagery, with a combined revisit time of 6 days. The Sentinel-1 data used in this study are the Interferometric Wide Swath mode (IW) Ground Range Detected High resolution (GRDH) product. These data were downloaded and processed via Google Earth Engine (GEE) (https://earthengine.google.com/platform/, accessed on 1 January 2024). The preprocessing steps included thermal noise removal, radiometric calibration, terrain correction, filtering, and resampling to 100 m. After data preprocessing, the output includes the backscatter coefficients for VV-pol and VH-pol (σ^0_{vv} and σ^0_{vh}), as well as the incidence angle. σ^0_{vv} and σ^0_{vh} were then used to calculate four SAR indices: the Polarization Ratio (PR) [44], Polarization Difference (PD) [20], Radar Vegetation Index (RVI) [45], and Normalized Difference Polarization Index (NDPI) [46].

2.2.2. Landsat-7/8/9 and Sentinel-2 Data

This study utilized optical satellite data from the Landsat series (including Landsat-7, Landsat-8, and Landsat-9) and the Sentinel-2 series (including Sentinel-2A and Sentinel-2B). To ensure data consistency and reliability, we conducted several preprocessing steps on the imagery data, including radiometric calibration, atmospheric correction, cloud and cloud shadow removal, resampling to 100 m, and index calculation. This study utilized reflectance data from common multispectral bands of the Landsat and Sentinel-2 series, including blue, green, red, Near-Infrared (NIR), Shortwave Infrared Band 1 (SWIR1), and Shortwave Infrared Band 2 (SWIR2). These data were used along with their derived indices as predictor datasets for SM estimation. The indices included the Normalized Difference Water Index (NDWI) [47], Soil Adjusted Vegetation Index (SAVI) [48], Enhanced Vegetation Index (EVI) [49], Shortwave Infrared Soil Moisture Index (SIMI) [50], and Normalized Difference Vegetation Index (NDVI) [51].

2.3. Field and Ancillary Datasets

The soil texture data at the SONTE-China sites were obtained through laboratory analysis, based on the percentages of sand and silt. The 5 cm daily average soil temperature (DAST) data were collected using 5TM soil sensors. Precipitation and evapotranspiration

data were sourced from the high-resolution climate reanalysis dataset ERA 5-Land, which has a spatial resolution of 0.1° and a temporal resolution of daily averages. These data were accessed and processed via the GEE platform. Topographic data were derived from the Shuttle Radar Topography Mission Version 3 (SRTM V3) product provided by National Aeronautics and Space Administration Jet Propulsion Laboratory (NASA JPL), with a resolution of 1 arc-second (approximately 30 m). Based on these data, information on elevation, slope, aspect, and the Topographic Wetness Index (TWI) was obtained. Other auxiliary data included the site's longitude, latitude, Day of Year (DOY), and land cover. The above data were used as the third data source to enhance the model's performance.

3. Methods

3.1. The Research Framework and Process of SM Estimation

Figure 2 provides a comprehensive overview of the SM estimation process, with the research framework comprising four main steps:

Data Collection and Preprocessing:

This study utilized three types of data sources (Table 3). Data source 1 was provided by Sentinel-2 and Landsat-7/8/9 series satellites, including reflectance data from six bands (blue, green, red, NIR, SWIR1, and SWIR2) and five indices: NDWI, SAVI, EVI, SIMI, and NDVI. Data source 2 was provided by Sentinel-1 series satellites, including three parameters (σ^0_{vv}, σ^0_{vh}, and incidence angle), four indices: PR, PD, RVI, and NDPI. It also includes the direction of orbit (DOO) and daily NDVI (NDVI_Daily) derived from the original NDVI data using S-G filtering and interpolation as auxiliary information. DOO indicates whether the data acquisition is on an ascending or descending orbit. Data source 3 comprised field and ancillary datasets, including three sets of soil data (sand fraction, silt fraction, and DAST), two sets of meteorological data (precipitation and evapotranspiration), four sets of topographical data (elevation, slope, aspect, and TWI), and four other sets of ancillary data (longitude, latitude, DOY, and land cover).

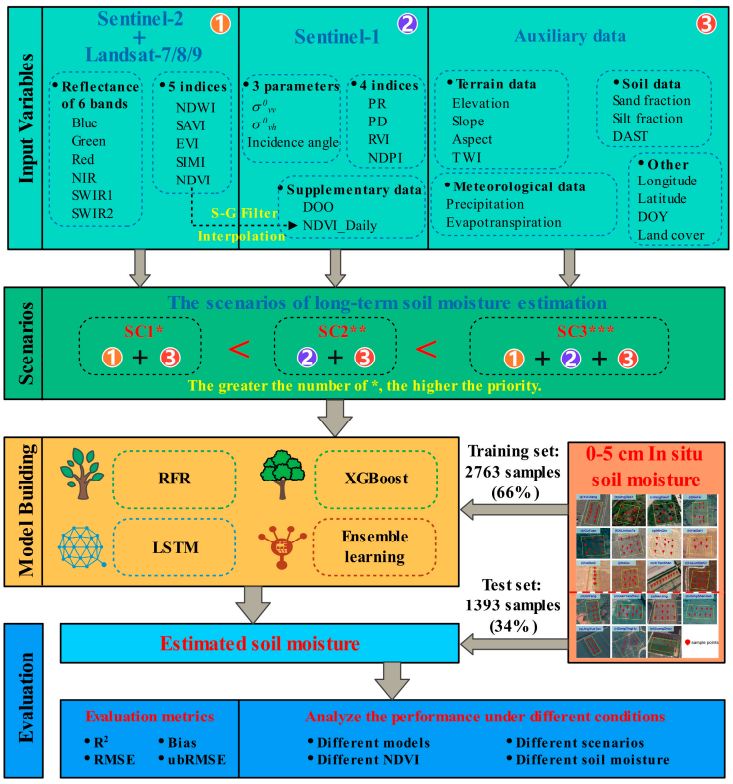


Figure 2. A framework for estimating SM based on multi-source RS data. *** represents the first priority, ** represents the second priority, and * represents the third priority.

Table 3. A list of predictive features from different data sources.

Type (Source)	Input Data	Type (Source)	Input Data
Data source 1: Multispectral observations (Sentinel-2A/2B; Landsat-7/8/9)	Blue; Green; Red; NIR; SWIR1; SWIR2	Data source 3: Auxiliary data	Sand fraction [%]
	NDWI		Silt fraction [%]
	SAVI		DAST [°C]
	EVI		Elevation [m]
	SIMI		Slope [°]
	NDVI		Aspect [°]
Data source 2: SAR observations (Sentinel-1A/1B)	NDVI_Daily		TWI
	σ^0_{vv} [dB]		Precipitation [mm/d]
	σ^0_{vh} [dB]		Evapotranspiration [mm/d]
	Incidence angle [°]		Longitude [°]
	PR: $\sigma^0_{vh}/\sigma^0_{vv}$ [linear scale]		Latitude [°]
	PD: $\sigma^0_{vv} - \sigma^0_{vh}$ [linear scale]		DOY
	RVI: $4\,\sigma^0_{vh}/(\sigma^0_{vv} + \sigma^0_{vh})$ [linear scale]		Land cover
	NDPI: $(\sigma^0_{vv} - \sigma^0_{vh})/(\sigma^0_{vv} + \sigma^0_{vh})$ [linear scale]		
	DOO		

Feature Development and Scenario Construction (SC):

A time-independent evaluation strategy was adopted in this study. If the data from 2021 are used as the test set, the proportion of the test set for DTH and JYT sites will be greater than that of the training set. Therefore, the data from the 17 sites in 2020 were integrated to form an independent test set (1393 samples), while the data from the remaining years at these 17 sites were integrated to form the training set (2763 samples). It is important to note that when the 5 cm soil temperature at a site is $\leq$0 °C, the dielectric properties of the soil change; as a result, the corresponding SM data were excluded. Additionally, the number of satellite images (optical or SAR) in Table 2 refers to those that can be matched with the available in situ SM measurements. Some sites have a small amount of missing SM measurements due to sensor malfunctions.

Based on the collected data, three scenarios were constructed with different features. SC3 includes 33 features generated from data sources 1, 2, and 3. It is required that the optical and SAR data, along with the dynamic auxiliary data (e.g., DAST, precipitation, and evapotranspiration), be acquired on the same day. SC2 includes 22 features generated from data sources 2 and 3. SC1 includes 24 features generated from data sources 1 and 3. The integration framework (IF) includes SC1, SC2, and SC3. Additionally, inversion strategies were set up for different situations as follows:

(1) When data sources meet SC3, the complete inversion framework in Figure 2 is used.
(2) When data source 1 is unavailable, SC2 observation data are used, skipping input from data source 1, and following a nearly identical inversion framework to Figure 2.
(3) When data source 2 is unavailable, SC1 observation data are used, skipping input from data source 2, and following a nearly identical inversion framework to Figure 2.
(4) If both data source 1 and data source 2 are unavailable, SM retrieval is not performed.

Model Description:

Different machine/deep learning methods were used, including Random Forest Regression (RFR), Extreme Gradient Boosting (XGBoost), Long Short-Term Memory (LSTM) networks, and Ensemble Learning (EL). The EL framework integrates RFR, XGBoost, and LSTM models using linear regression. Overall, 66% of the SM measurement data was used for training, and 34% for validation.

Model Evaluation and Performance Analysis:

The impact of different feature combinations and models on SM dynamics prediction was evaluated. The detailed calculation and definitions of these metrics are provided in Section 3.3.

3.2. Model Description

This study employed four models to estimate SM: RFR, XGBoost, LSTM, and EL. These models are capable of handling the complex nonlinear relationships between input features and measured SM. To ensure a fair and uniform comparison of the four models, 66% of the input features and SM data (2763 samples) were selected as the training set, while the remaining 34% (1393 samples) were used as the test set to evaluate the accuracy of the SM estimates from these models. Additionally, grid search combined with 5-fold cross-validation was utilized to optimize the parameters during model training and to assess the models' stability and generalization capabilities.

RFR is an ensemble machine learning method that enhances prediction performance and stability by constructing multiple decision trees [52]. Key parameters include the number of trees, features per split, and minimum samples per leaf. Due to its inherent randomness, RFR exhibits strong regression and generalization capabilities. XGBoost is an advanced machine learning algorithm based on the gradient boosting framework that incrementally optimizes the objective function [53]. Its performance was enhanced in this study using grid search to adjust hyperparameters such as maximum depth, learning rate, γ, and λ [10]. LSTM is an enhanced variant of Recurrent Neural Networks (RNNs) that addresses gradient vanishing and explosion issues [54]. It controls information flow through input, forget, and output gates, making it suitable for time series data such as SM or leaf area index predictions [55]. EL is a machine learning approach that enhances model performance by combining multiple base learners [56]. In this study, RFR, XGBoost, and LSTM were used as base models in a stacking ensemble framework. The base models were independently trained, and their predictions were used as new features for a meta-model (e.g., linear regression), which was trained on this new dataset. A 5-fold cross-validation was employed to mitigate overfitting risk.

3.3. Validation Metrics

The performance of the models was evaluated using five commonly used statistical metrics: correlation coefficient (r), the coefficient of determination (R^2), root mean squared error (RMSE), bias, and unbiased RMSE (ubRMSE). The metrics are calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \quad (4)$$

$$ubRMSE = \sqrt{RMSE^2 - Bias^2} \quad (5)$$

x_i represents the input data values; $\overline{x}$ represents the mean of the input data values; y_i represents the observed values; $\hat{y}_i$ represents the model-estimated values; $\overline{y}$ represents the mean of the observed values; n represents the number of observations.

4. Results

4.1. Correlation Analysis of Prediction Indicators and Observed SM

This study conducted a correlation analysis between training set features from different data sources (optical, SAR, and auxiliary data) and observed SM (Table 4). The more stars (*) there are, the greater the significance of the correlation present.

Table 4. Correlation analysis of input data and observed SM.

Input Data	Correlation Coefficient (r)	Input Data	Correlation Coefficient (r)
Blue	−0.23	NDPI	0.10
Green	−0.31	DOO	−0.13
Red	−0.36	NDVI_Daily	0.26
NIR	−0.09	Sand fraction	−0.48 ****
SWIR1	−0.28	Silt fraction	0.47 ***
SWIR2	−0.39 *	DAST	−0.03
NDWI	0.12	Precipitation	0.13
SAVI	0.22	Evapotranspiration	−0.42 **
EVI	0.22	Elevation	−0.08
SIMI	−0.34	Slope	0.01
NDVI	0.25	Aspect	−0.17
σ^0_{vv}	0.47 ***	TWI	−0.28
σ^0_{vh}	0.34	Longitude	0.22
Incidence angle	−0.10	Latitude	−0.35
PR	−0.09	DOY	−0.06
PD	0.34	Land cover	−0.15
RVI	−0.10		

❶ green; ❷ blue; ❸ yellow.

Among the optical RS features, SWIR2 reflectance data showed the strongest negative correlation (−0.39), followed by red band reflectance data (−0.36), SIMI (−0.34), and green band reflectance data (−0.31). This indicates that these features are important in predicting SM using optical RS. Among the SAR features, σ^0_{vv} exhibited the strongest positive correlation (0.47), indicating its high potential in SM prediction. Both σ^0_{vh} and PD also showed significant correlations (0.34). Additionally, NDVI_Daily, used as auxiliary data for SAR, demonstrated a considerable correlation (0.26). Due to its ability to penetrate vegetation and the soil surface, SAR data have shown high value in SM retrieval, especially under conditions where optical RS is limited. Among the ancillary data features, sand fraction and silt fraction exhibited the most significant correlations, with r values of −0.48 and 0.47, respectively. This indicates that soil particle composition has a substantial impact on moisture content. Evapotranspiration showed a high negative correlation (−0.42), making it one of the most influential auxiliary data. Additionally, other auxiliary data such as latitude (−0.35) and TWI (−0.28) also showed certain negative correlations.

Overall, the top five features with the highest correlations were sand fraction, σ^0_{vv}, silt fraction, evapotranspiration, and SWIR2 reflectance data, with r values of −0.48, 0.47, 0.47, −0.42, and −0.39, respectively. These results provide strong support for further optimization of SM retrieval models and highlight the importance of integrating multi-source RS data with auxiliary data.

4.2. Evaluation and Comparison of Different Models

Figure 3 illustrates the SM prediction results of the four models (RFR, XGBoost, LSTM, and EL) based on the SONTE-China training and test sets across 17 sites. These results specifically reflect the outcomes from IF encompassing SC1, SC2, and SC3. Comparing the performance on the training set, EL and RFR showed similar accuracy. EL achieved an R^2 of 0.973, RMSE and ubRMSE of 0.021 m^3/m^3. RFR followed closely with an R^2 of 0.969, RMSE and ubRMSE of 0.022 m^3/m^3, demonstrating high fitting accuracy. LSTM performed the worst with an R^2 of only 0.818 and showed slight underestimation (bias = −0.006 m^3/m^3). On the test set, the R^2 of all models decreased, exhibiting varying degrees of overestimation for low values and underestimation for high values. RFR, XGBoost, and EL demonstrated similar performances. RFR achieved an R^2 of 0.758, with RMSE and ubRMSE values of 0.061 m^3/m^3, indicating good generalization ability. XGBoost attained an R^2 of 0.751, with RMSE and ubRMSE values of 0.062 m^3/m^3. EL showed an R^2 of 0.743, with RMSE and

ubRMSE values of 0.063 m^3/m^3. The R^2 of LSTM decreased to 0.633, indicating a weaker generalization ability.

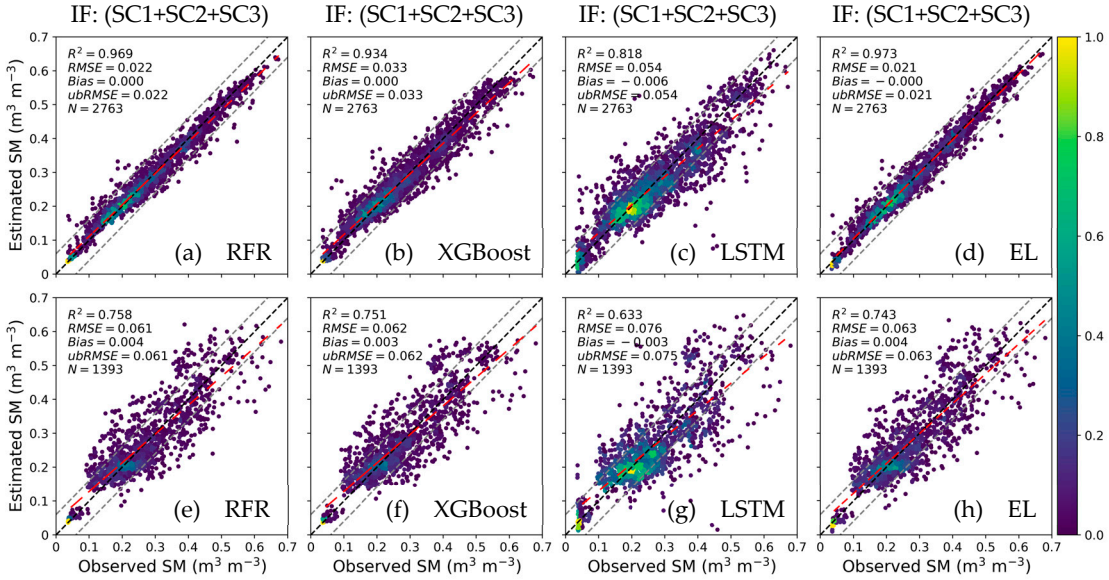


Figure 3. The training (**top**) and test (**bottom**) results of four models from IF at SONTE-China (17 sites). The red dotted line is the trend line. The gray dotted line represents the error line at 0.06 m^3/m^3.

Considering the performance on both the training and test sets, RFR demonstrated high stability and accuracy in SM prediction, showing consistency across different datasets. EL exhibited outstanding performance on the training set, its performance declined on the test set, potentially due to the choice of the meta-model. In this study, linear regression was used as the meta-model, but selecting a more suitable meta-model for this dataset is beyond the scope of this paper. Overall, RFR and EL demonstrated good performance in this study and are the preferred models for SM prediction.

4.3. Evaluation and Comparison of Different Scenarios

Figures 4 and 5 show the SM prediction results of the four models under different feature scenarios (SC1, SC2, SC3) for the training and test sets, respectively.

In the training set, under SC1, RFR and EL performed the best, both with an R^2 of 0.984 and RMSE and ubRMSE of 0.016 m^3/m^3. XGBoost had slightly lower accuracy (R^2 of 0.930, ubRMSE of 0.034 m^3/m^3) compared to RFR and EL, while LSTM showed relatively poor performance (R^2 of 0.849, ubRMSE of 0.047 m^3/m^3). Under SC2, EL and RFR demonstrated similar performance. EL achieved an R^2 of 0.959 and a ubRMSE of 0.026 m^3/m^3, while RFR attained an R^2 of 0.953 and a ubRMSE of 0.028 m^3/m^3. LSTM performed the worst, with an R^2 of 0.821 and a ubRMSE of 0.054 m^3/m^3. Under SC3, EL, XGBoost, and RFR demonstrated similar performance, with $R^2 \geq 0.956$ and ubRMSE ≤ 0.026 m^3/m^3. LSTM showed a pronounced trend of underestimating high values and overestimating low values under SC3, with an R^2 of 0.575 and a ubRMSE of 0.078 m^3/m^3.

In the test set, RFR, XGBoost, and EL showed similar performance under the same configurations (SC1, SC2, or SC3). Under SC1, RFR and XGBoost had nearly identical R^2 and ubRMSE values, while LSTM performed poorly. Under SC2, the accuracy metrics for RFR (R^2 of 0.773 and ubRMSE of 0.061 m^3/m^3) were slightly higher than those for XGBoost (R^2 of 0.758 and ubRMSE of 0.063 m^3/m^3) and EL (R^2 of 0.755 and ubRMSE of 0.063 m^3/m^3). Under SC3, RFR again demonstrated slightly better accuracy than XGBoost and EL. LSTM demonstrated significant overestimation for low values and underestimation for high values, with an R^2 of only 0.466 and a ubRMSE of 0.093 m^3/m^3.

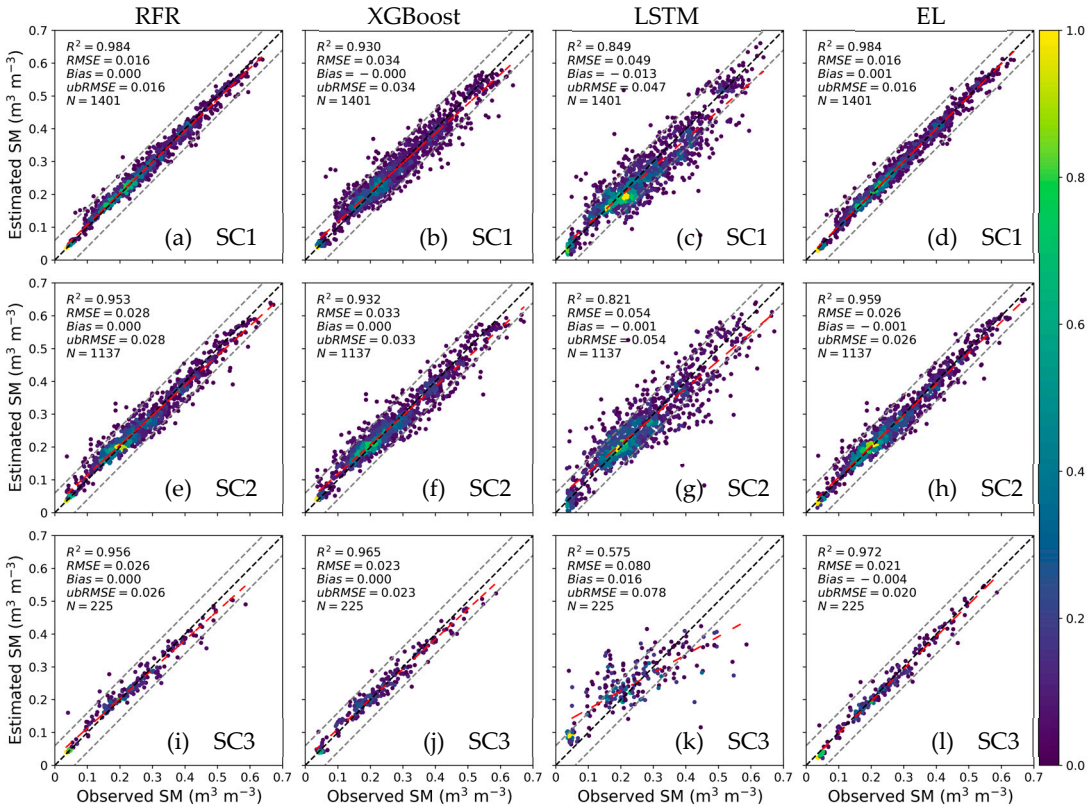


Figure 4. The training results of four models at SONTE-China (17 sites). The red dotted line is the trend line. The gray dotted line represents the error line at 0.06 m^3/m^3.

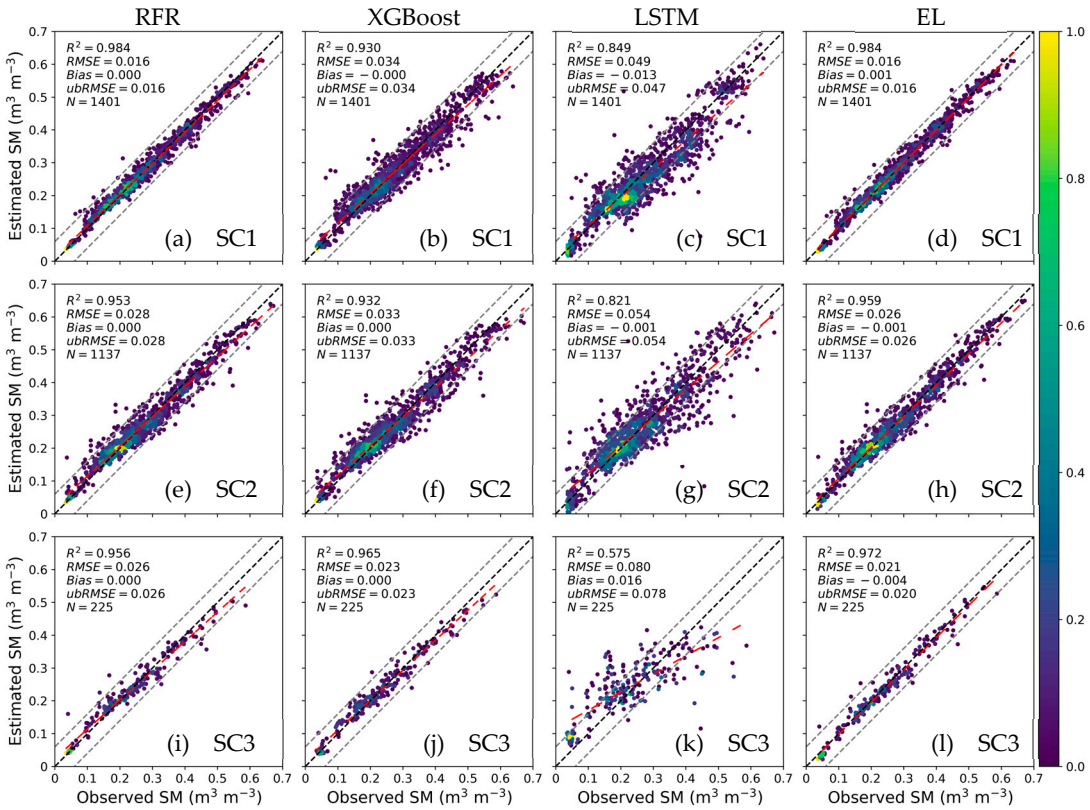


Figure 5. The test results of four models at SONTE-China (17 sites). The red dotted line is the trend line. The gray dotted line represents the error line at 0.06 m^3/m^3.

Overall, all models performed better on the training set compared to the test set. RFR, XGBoost, and EL all showed excellent performance. Additionally, due to the significant differences in sample sizes across different scenarios (especially SC3), the accuracy comparison of the same model under different scenarios may not be entirely representative. Therefore, when evaluating model performance under different scenarios, the differences in sample sizes should be taken into consideration (see Section 5.1 for details).

4.4. Comparison of Temporal Variations Between Estimated and Observed SM

Figure 6 presents a comparison of estimated and observed SM time series at three representative sites with different SM levels: NQ (high SM), JYT (medium SM), and MQ (low SM) under the three scenarios. Overall, the estimated and observed SM responses were relatively consistent in the training set for representative sites with different SM levels. However, there were differences in performance across the three sites in the test set. At the NQ site, the estimated SM was initially underestimated and then overestimated at the end. For the JYT site, the observed SM in the training set showed good consistency with the estimated SM. In the test set, frequent precipitation causes the observed SM to experience rapid and significant changes over short periods. This makes it challenging for the model to capture these rapid continuous changes, resulting in underestimation. This indicates the model's limitations in handling rapidly changing medium SM conditions. The estimated SM at the MQ site was able to accurately reflect the observed SM trend, indicating that the model has high estimation accuracy under low SM conditions.



Figure 6. The time series of estimated and observed SM from three scenarios at NQ, JYT, and MQ sites. The blue solid line represents the observed SM at 0–5 cm. The green solid line represents the daily NDVI. The red, green, and purple squares represent the estimated SM for SC1, SC2, and SC3, respectively. The blue bars indicate daily precipitation. The red dashed vertical lines distinguish between the training and test sets.

Although the estimated SM in the test set shows a slight bias, the models constructed using the IF demonstrated a good ability to represent observed SM trends in both the training and test sets. This indicates that the method has significant feasibility and application potential for generating high-spatiotemporal-resolution SM products at a national scale in China. Further optimization of the models, particularly in handling extreme and rapidly changing SM conditions, will help enhance estimation accuracy.

4.5. Spatial Comparison of Revisit Time Between SC1, SC2, SC3, and IF

Figures 7 and 8 illustrate the revisit time in 2021 and 2022 for different scenarios in China for SM monitoring. Different colors represent the revisit times for monitoring SM: red indicates a revisit time of ≤1 day, orange indicates a revisit time of (1–2] days, and blue indicates a revisit time of (31–365] days. SC1, which only uses optical RS and auxiliary data, has a shorter revisit time but is less sensitive to SM. SC2, which uses SAR and auxiliary data, is more sensitive to SM but has a longer revisit time, particularly after the Sentinel-1B sensor malfunctioned in December 2021, significantly reducing revisit time in the northeastern region. SC3 combines optical, SAR, and auxiliary data, providing multi-source information. Although the frequency of synergistic observations is limited, SC3 performs well in SM monitoring. The IF leverages the advantages of SC3, SC2, and SC1, prioritizing SC3, followed by SC2, and finally SC1. This approach significantly enhances the spatiotemporal coverage and reliability of monitoring.

Additionally, the IF effectively addresses the issue of SM monitoring gaps in the northeastern region caused by the Sentinel-1B sensor malfunction (Figure 8b). It is important to note that the revisit time statistics for optical RS data did not account for the impact of cloud cover and cloud shadows, which are major obstacles to the usability of optical RS images. Overall, the IF configuration, by combining multiple data sources, provides more frequent and reliable monitoring data, which is crucial for large-scale, high-frequency SM monitoring. For detailed statistics and analysis of the revisit time for different scenarios, refer to Section 5.4.

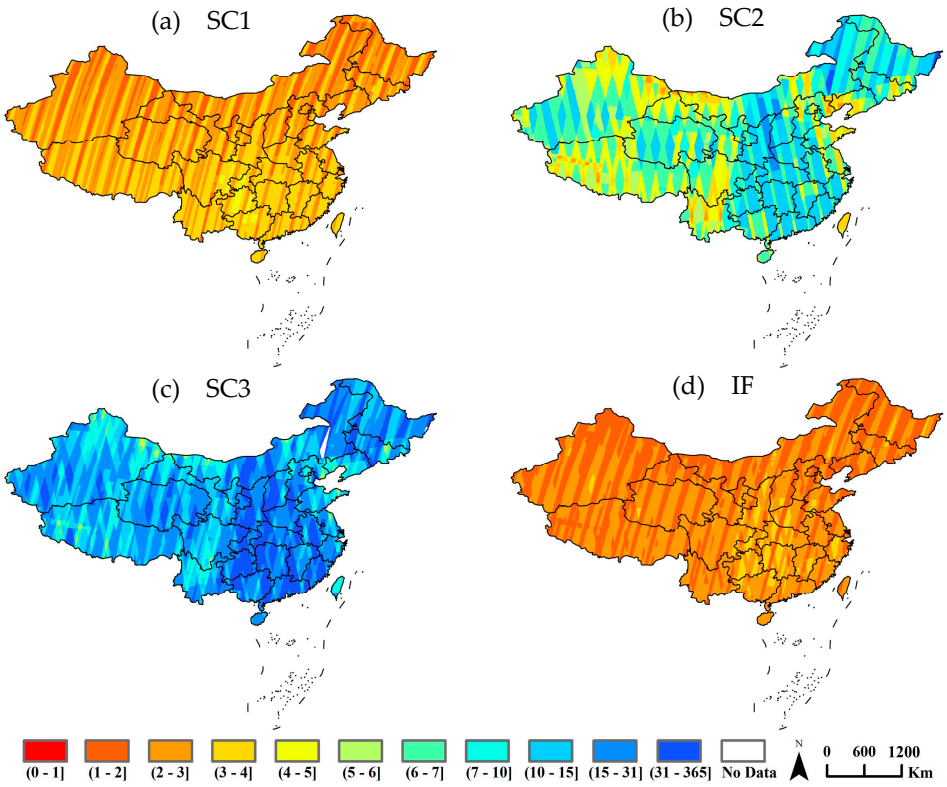


Figure 7. Revisit time between SC1, SC2, SC3, and IF for monitoring SM in China (2021). (**a**) SC1: Optical RS + auxiliary data only; (**b**) SC2: SAR + auxiliary data only; (**c**) SC3: optical RS + SAR + auxiliary data; (**d**) IF: combined SC3, SC2, and SC1 scenarios.

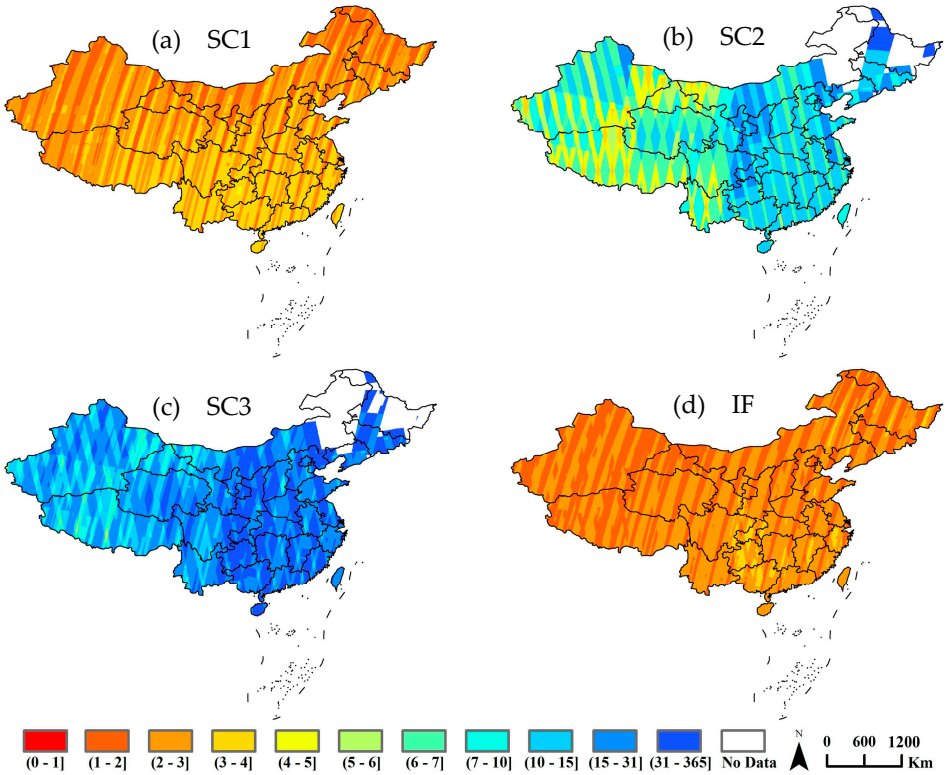


Figure 8. Revisit time between SC1, SC2, SC3, and IF for monitoring SM in China (2022). (**a**) SC1: Optical RS + auxiliary data only; (**b**) SC2: SAR + auxiliary data only; (**c**) SC3: optical RS + SAR + auxiliary data; (**d**) IF: combined SC3, SC2, and SC1 scenarios.

5. Discussion

5.1. Analysis of SM Estimation Accuracy for Different Data Source Combinations Using the Same Sample Set

Based on the findings in Section 4.2, RFR demonstrated excellent performance. Therefore, RFR was used for SM prediction under different RS data conditions (Figure 9). To better compare the effects of optical RS + auxiliary data, SAR + auxiliary data, and optical RS + SAR + auxiliary data on SM retrieval accuracy using the same sample set (with consistent sample size and distribution), we further divided the SC3 dataset. Specifically, the SC3 dataset, which includes optical RS + SAR + auxiliary data, was split into three categories: optical RS + auxiliary data, SAR + auxiliary data, and optical RS + SAR + auxiliary data. This allowed us to investigate the impact of different RS data combinations on SM retrieval accuracy under the same conditions. Unlike the priority-based retrieval strategy discussed in Section 4, this analysis focuses on comparing how different RS data source combinations influence SM retrieval accuracy under identical sample sets.

In the training set, the combinations of optical RS + SAR + auxiliary data and SAR + auxiliary data demonstrated similar accuracy. Among them, the accuracy metrics of optical RS + SAR + auxiliary data were slightly better than those of SAR + auxiliary data, with an R^2 of 0.956 and an ubRMSE of 0.0256 m^3/m^3 (Figure 9c). In contrast, optical RS + auxiliary data showed slightly lower accuracy, with an R^2 of 0.888 and an ubRMSE of 0.0409 m^3/m^3. In the test set, the combinations of optical RS + SAR + auxiliary data, SAR + auxiliary data, and optical RS + auxiliary data demonstrated similar accuracy. Among these, optical RS + SAR + auxiliary data achieved the highest accuracy metrics, while optical RS + auxiliary data showed the lowest reliability.

Overall, the accuracy of SM retrieval improved in both the training and test sets with the addition of multi-source RS data (from optical RS + auxiliary data to optical RS + SAR + auxiliary data), except for the bias metrics for the training set. This validates the

effectiveness of combining multi-source RS data for SM monitoring, although further data samples are needed for validation.

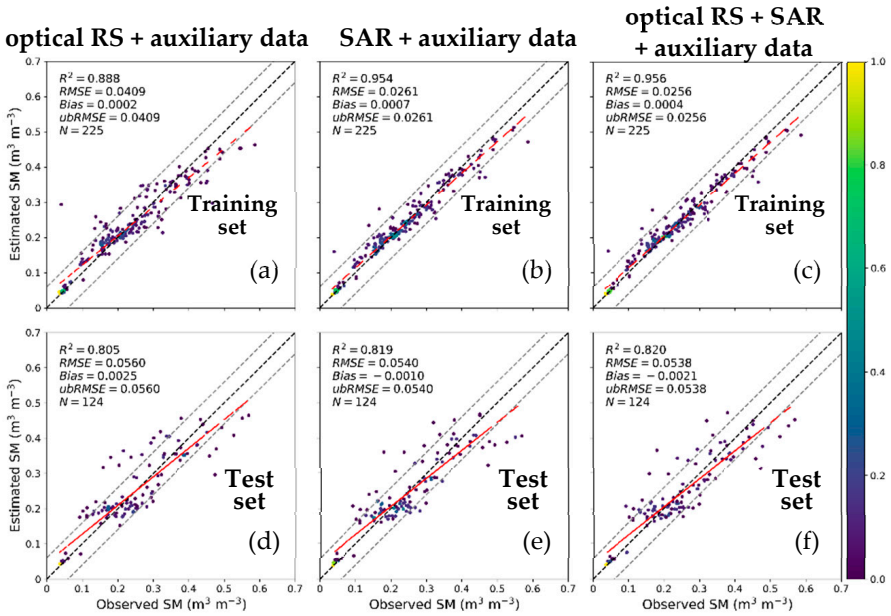


Figure 9. Training (**top**) and test (**bottom**) results of three categories using the RFR based on the SC3 dataset at SONTE-China (17 sites). The red dotted line is the trend line. The gray dotted line represents the error line at 0.06 m^3/m^3.

5.2. Comparison of SM Estimation Accuracy for Different NDVI

Figure 10 illustrates the accuracy and error of four models (RFR, XGBoost, LSTM, EL) in the training and test sets under different NDVI conditions. NDVI is divided into five levels: (0–0.2], (0.2–0.4], (0.4–0.6], (0.6–0.8], and (0.8–1]. The colored dot lines represent R^2, while the bar charts represent ubRMSE, with different colors corresponding to different models.

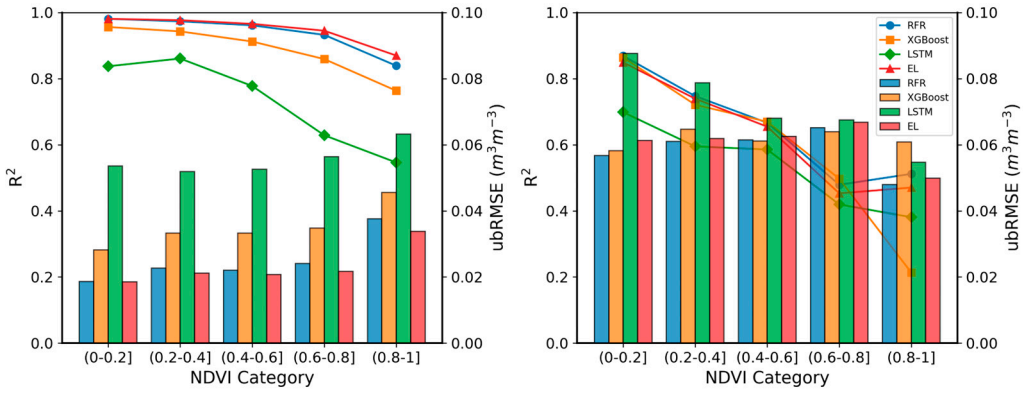


Figure 10. Performance of different models under various NDVI categories in the training set (**left**) and test set (**right**). The colored dot lines represent R^2, and the bar charts represent ubRMSE.

In the training set, EL and RFR exhibited high R^2 across all NDVI levels, with XGBoost following closely. These models demonstrated good stability and consistency. LSTM showed lower R^2 across all NDVI levels, performing particularly poorly in the (0.6–0.8] and (0.8–1] levels. The R^2 of all four models tended to decrease as NDVI increased. In the test set, the R^2 for RFR, XGBoost, and EL significantly decreased compared to the training set. XGBoost had the lowest R^2 in the (0.8–1] level. RFR, XGBoost, and EL outperformed LSTM in terms of R^2. EL had the lowest ubRMSE across all NDVI levels in the training set, indicating the strong fitting ability. In the test set, RFR and LSTM showed the smallest and

largest ubRMSE, respectively. Overall, RFR and EL demonstrated high accuracy and low error across different NDVI conditions.

5.3. Comparison of SM Estimation Accuracy for Different SM Categories

Figure 11 displays the prediction accuracy of four models (RFR, XGBoost, LSTM, EL) under different SM levels, with the left panel representing the training set and the right panel representing the test set. SM is divided into six levels: (0–0.1], (0.1–0.2], (0.2–0.3], (0.3–0.4], (0.4–0.5], and (0.5–0.7]. The bar charts show the ubRMSE for different models, and the red dot line indicates the average ubRMSE of the four models.

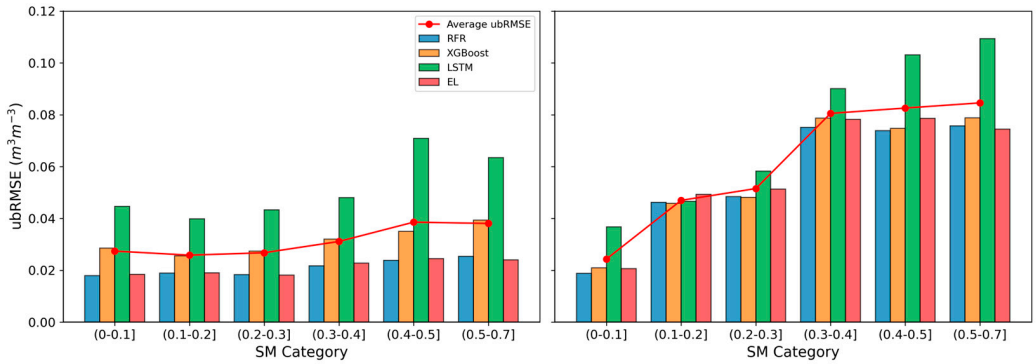


Figure 11. Performance of different models under various SM categories in the training set (**left**) and test set (**right**). The bar charts represent ubRMSE, and the red dot line represents the average ubRMSE.

In the training set, the average ubRMSE of all models generally increases with the SM levels. LSTM has the highest ubRMSE across all SM levels, indicating the largest errors. In contrast, RFR and EL have lower ubRMSE, indicating higher prediction accuracy, especially in the (0–0.3] range. In the test set, the trend is similar to that of the training set, with the average ubRMSE gradually increasing as the SM levels rise. LSTM's ubRMSE are the highest across most SM levels, except for the (0.1–0.2] range. RFR exhibits the lowest overall ubRMSE in the test set, indicating the highest prediction accuracy. Although EL and XGBoost also perform well, their prediction accuracy is slightly lower than that of RFR.

In summary, RFR and EL demonstrate high prediction accuracy and low error across different SM levels, making them the preferred model for SM prediction. The LSTM model shows lower prediction accuracy and higher error across various SM levels. The average ubRMSE of the four models increases with higher SM levels, indicating a decrease in prediction accuracy as SM levels rise.

5.4. Revisit Time Analysis of SM Monitoring Under Different Scenarios

This study analyzed the revisit time of multi-source RS data for monitoring SM at a national scale in China under different scenarios (Figure 12). The results show that SC1 predominantly had a revisit time of (2–3] days in 2021 and 2022, accounting for 43.78% and 46.66% of the total area in China, respectively. The total proportion of areas with a revisit time of (1–4] days was 98.75% in 2021 and 99.90% in 2022, indicating a significant advantage of SC1 in high-frequency monitoring. In 2021, SC2 had the highest proportion of areas with a revisit time in the (6–7] days range, accounting for 30.24% of the total area. However, in 2022, due to the Sentinel-1B sensor malfunction, the proportion of high revisit time monitoring significantly decreased across the monitored area in China. The proportion of areas with a revisit time of (4–5] days shifted from third place in 2021 to the "Other" category in 2022. The proportion of areas with a revisit time of (6–7] days proportion decreased from 30.24% to 24.23%, while the proportion of areas with a revisit time of (10–15] days proportion increased from 22.77% to 26.55%. This change indicates that SC2's revisit time became more dispersed, with an increase in low revisit time monitoring. For SC3, the primary revisit time across the monitored area in China in 2021 was (15–31] days, accounting for 42.51% of the total area, with areas having a revisit time of (10–15] days

and (31–365] days accounting for 18.84% and 25.07%, respectively. In 2022, the proportion of areas with a revisit time of (15–31] days slightly decreased to 40.82%, the proportion of areas with a revisit time of (10–15] days decreased to 16.68%, and the proportion of areas with a revisit time of (31–365] days increased to 26.66%. This shift reflects a move towards lower revisit time monitoring in SC3 during 2022. The IF, which integrates the advantages of SC1, SC2, and SC3, mainly had a revisit time of (2–3] days in 2021 and 2022, accounting for 52.39% and 54.43% of the total monitored area in China, respectively. The total proportion of areas with a revisit time of (1–3] days was 95.01% in 2021 and 96.53% in 2022.

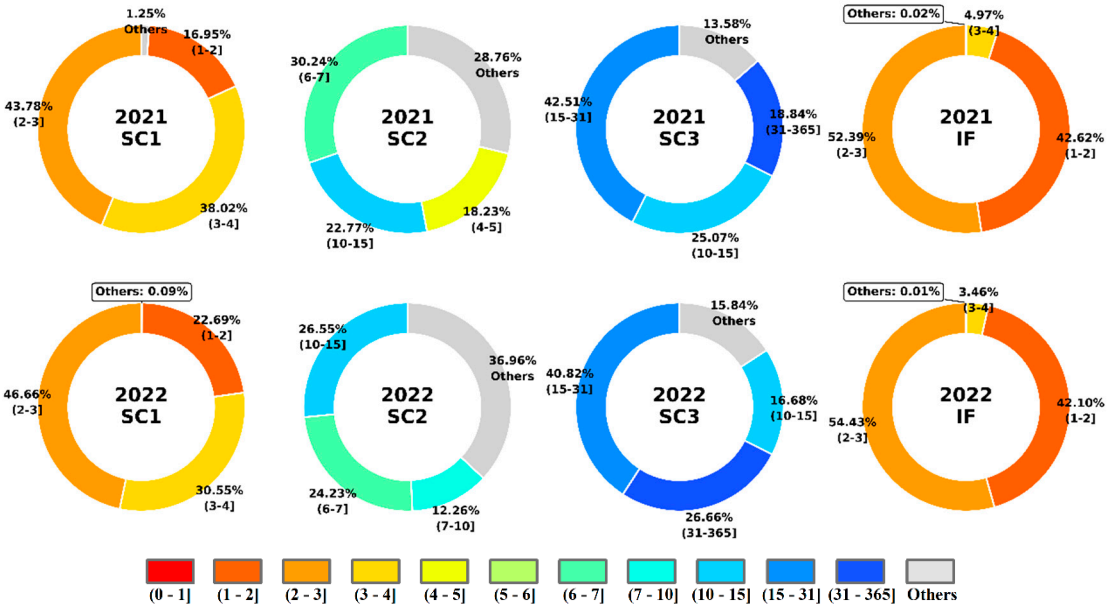


Figure 12. Revisit time distribution for multi-source RS monitoring of SM under different scenarios (2021–2022).

In summary, there are differences in the revisit time of SM monitoring among the different scenarios. The IF, by integrating the advantages of SC1, SC2, and SC3, provides higher revisit time and effectively enhances the monitoring of SM dynamics. This finding offers important insights for optimizing future RS strategies for SM monitoring. In particular, in cases of sensor malfunction or data gaps, the rational configuration and combination of different RS data sources can significantly improve the timeliness and coverage of monitoring.

5.5. Limitations and Future Prospects

In recent years, RS-based SM data products, such as SMOS, SMAP, and ASCAT, along with model simulation products like ERA 5-Land, have generally faced issues related to coarse spatial resolution. These products often fail to meet the demands of fine-scale research, particularly in scenarios that require high spatial and temporal resolution. Although many scholars have employed machine learning techniques to downscale these coarse-resolution SM products to approximately 1 km [57–60], the demand for higher resolution remains unmet. This is especially true in high-latitude and topographically complex areas, where the performance of downscaling algorithms is often unsatisfactory [61].

This study proposes a new framework that directly integrates high-resolution multisource RS data to produce SM products with a high spatial resolution of 100 m and a temporal resolution that is better than 3 days. The primary advantage of this framework lies in its ability to significantly reduce revisit times for SM monitoring using high-resolution satellites. This enables our SM products to better capture short-term anomalies in hydrological signals [62], thereby aiding the implementation of precision agriculture irrigation. However, high-resolution optical data still face challenges, particularly related to cloud

contamination, which can lead to data gaps and signal interference. Additionally, the coverage of high-resolution satellites is limited, with revisit times that cannot achieve 1 day. These factors pose challenges to generating continuous SM products based on the current framework. Overall, the framework we present lays the foundation for developing high-quality SM products, but further modifications and optimizations are necessary to achieve integrated characteristics such as high resolution, seamless spatial distribution, long time series, and low error.

Future research could focus on augmenting in situ SM observation data from various sources across different locations, climates, and surface environments [63]. Incorporating existing coarse-resolution SM data with seamless spatial distribution into our framework may also serve as an effective solution for producing temporally and spatially continuous SM products.

6. Conclusions

To meet the demand for high-spatiotemporal-resolution and accurate SM data in climate, hydrology, and agriculture, this study integrates multi-source data including optical RS, SAR, and auxiliary data, and applies four machine/deep learning algorithms to construct a high-spatiotemporal-resolution multi-source RS SM retrieval framework suitable for application on a national scale in China. This framework meets the needs for high-precision and high-spatiotemporal-resolution SM data for precision agriculture and environmental monitoring. The conclusions are as follows:

(1) The R values between auxiliary data (such as sand fraction, r = −0.48; silt fraction, r = 0.47; and evapotranspiration, r = −0.42), SAR features (such as σ^0_{vv}, r = 0.47), and optical RS features (such as SWIR2 reflectance data, r = −0.39) with observed SM are significant. This indicates that multi-source RS data can provide complementary information for SM monitoring. (2) Compared to XGBoost and LSTM, RFR and EL demonstrate superior overall performance and are the preferred models for SM prediction. Their R^2 for the training and test sets exceed 0.969 and 0.743, respectively, and their ubRMSE are below 0.022 and 0.063 m^3/m^3, respectively. (3) The SM prediction accuracy is highest for the scenario of optical + SAR + auxiliary data, followed by SAR + auxiliary data, and finally optical + auxiliary data. With increasing NDVI and SM values, the trained models exhibit a general decrease in prediction performance and accuracy. (4) In 2021 and 2022, without considering cloud cover, the IF theoretically achieves an SM monitoring revisit time of 1–3 days in 95.01% and 96.53% of the total area in China, respectively.

This study integrates multi-source RS data and innovatively provides a reference framework for developing SM products with high spatiotemporal resolution.

Author Contributions: Methodology, Z.F.; Software, J.Z.; Validation, T.G.; Investigation, L.L.; Writing—original draft, Z.F.; Writing—& editing, Z.F., X.Z. and X.L.; Visualization, J.S. and J.Z.; Supervision, X.L. and C.W.; Funding acquisition, X.Z. and L.L. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the National Natural Science Foundation of China (No. 42371381), National Key Research and Development Project of China (No. 2021YFD1500103), Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDA28100500), Science and Technology Development Plan Project of Jilin Province (No. 20210201044GX; No. YDZJ202401484ZYTS), Land Observation Satellite Supporting Platform of National Civil Space Infrastructure Project (No. CASPLOS-CCSI).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. He, B.; Jia, B.; Zhao, Y.; Wang, X.; Wei, M.; Dietzel, R. Estimate soil moisture of maize by combining support vector machine and chaotic whale optimization algorithm. *Agric. Water Manag.* **2022**, *267*, 107618. [CrossRef]

2. Piao, S.L.; Ciais, P.; Huang, Y.; Shen, Z.H.; Peng, S.S.; Li, J.S.; Zhou, L.P.; Liu, H.Y.; Ma, Y.C.; Ding, Y.H.; et al. The impacts of climate change on water resources and agriculture in China. *Nature* **2010**, *467*, 43–51. [CrossRef] [PubMed]
3. Robinson, D.A.; Campbell, C.S.; Hopmans, J.W.; Hornbuckle, B.K.; Jones, S.B.; Knight, R.; Ogden, F.; Selker, J.; Wendroth, O. Soil moisture measurement for ecological and hydrological watershed-scale observatories: A review. *Vadose Zone J.* **2008**, *7*, 358–389. [CrossRef]
4. Karthikeyan, L.; Chawla, I.; Mishra, A.K. A review of remote sensing applications in agriculture for food security: Crop growth and yield, irrigation, and crop losses. *J. Hydrol.* **2020**, *586*, 124905. [CrossRef]
5. Babaeian, E.; Paheding, S.; Siddique, N.; Devabhaktuni, V.K.; Tuller, M. Estimation of root zone soil moisture from ground and remotely sensed soil information with multisensor data fusion and automated machine learning. *Remote Sens. Environ.* **2021**, *260*, 112434. [CrossRef]
6. Dorigo, W.; Himmelbauer, I.; Aberer, D.; Schremmer, L.; Petrakovic, I.; Zappa, L.; Preimesberger, W.; Xaver, A.; Annor, F.; Ardö, J.; et al. The International Soil Moisture Network: Serving Earth system science for over a decade. *Hydrol. Earth Syst. Sci.* **2021**, *25*, 5749–5804. [CrossRef]
7. Fan, L.; Xing, Z.; Lannoy, G.D.; Frappart, F.; Peng, J.; Zeng, J.; Li, X.; Yang, K.; Zhao, T.; Shi, J.; et al. Evaluation of satellite and reanalysis estimates of surface and root-zone soil moisture in croplands of Jiangsu Province, China. *Remote Sens. Environ.* **2022**, *282*, 113283. [CrossRef]
8. Wang, L.; Fang, S.; Pei, Z.; Wu, D.; Zhu, Y.; Zhuo, W. Developing machine learning models with multisource inputs for improved land surface soil moisture in China. *Comput. Electron. Agric.* **2022**, *192*, 106623. [CrossRef]
9. Zhang, P.; Zheng, D.; van der Velde, R.; Zeng, J.; Wang, X.; Wang, Z.; Zeng, Y.; Wen, J.; Li, X.; Su, Z. Assessment of long-term multisource surface and subsurface soil moisture products and estimate methods on the Tibetan Plateau. *J. Hydrol.* **2024**, *640*, 131713. [CrossRef]
10. Jamei, M.; Karbasi, M.; Malik, A.; Jamei, M.; Kisi, O.; Yaseen, Z.M. Long-term multi-step ahead forecasting of root zone soil moisture in different climates: Novel ensemble-based complementary data-intelligent paradigms. *Agric. Water Manag.* **2022**, *269*, 107679. [CrossRef]
11. Kerr, Y.H.; Waldteufel, P.; Wigneron, J.P.; Martinuzzi, J.M.; Font, J.; Berger, M. Soil moisture retrieval from space: The Soil Moisture and Ocean Salinity (SMOS) mission. *IEEE Trans. Geosci. Remote Sens.* **2001**, *39*, 1729–1735. [CrossRef]
12. Entekhabi, D.; Njoku, E.G.; O'Neill, P.E.; Kellogg, K.H.; Crow, W.T.; Edelstein, W.N.; Entin, J.K.; Goodman, S.D.; Jackson, T.J.; Johnson, J.; et al. The Soil Moisture Active Passive (SMAP) Mission. *Proc. IEEE* **2010**, *98*, 704–716. [CrossRef]
13. Imaoka, K.; Kachi, M.; Fujii, H.; Murakami, H.; Hori, M.; Ono, A.; Igarashi, T.; Nakagawa, K.; Oki, T.; Honda, Y.; et al. Global Change Observation Mission (GCOM) for Monitoring Carbon, Water Cycles, and Climate Change. *Proc. IEEE* **2010**, *98*, 717–734. [CrossRef]
14. Dorigo, W.; Wagner, W.; Albergel, C.; Albrecht, F.; Balsamo, G.; Brocca, L.; Chung, D.; Ertl, M.; Forkel, M.; Gruber, A.; et al. ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions. *Remote Sens. Environ.* **2017**, *203*, 185–215. [CrossRef]
15. Wagner, W.; Hahn, S.; Kidd, R.; Melzer, T.; Bartalis, Z.; Hasenauer, S.; Figa-Saldaña, J.; de Rosnay, P.; Jann, A.; Schneider, S.; et al. The ASCAT Soil Moisture Product: A Review of its Specifications, Validation Results, and Emerging Applications. *Meteorol. Z.* **2013**, *22*, 5–33. [CrossRef]
16. Chan, S.K.; Bindlish, R.; O'Neill, P.; Jackson, T.; Njoku, E.; Dunbar, S.; Chaubell, J.; Piepmeier, J.; Yueh, S.; Entekhabi, D.; et al. Development and assessment of the SMAP enhanced passive soil moisture product. *Remote Sens. Environ.* **2018**, *204*, 931–941. [CrossRef]
17. Das, N.N.; Entekhabi, D.; Dunbar, R.S.; Chaubell, M.J.; Colliander, A.; Yueh, S.; Jagdhuber, T.; Chen, F.; Crow, W.; O'Neill, P.E.; et al. The SMAP and Copernicus Sentinel 1A/B microwave active-passive high resolution surface soil moisture product. *Remote Sens. Environ.* **2019**, *233*, 111380. [CrossRef]
18. Ma, H.; Zeng, J.; Zhang, X.; Peng, J.; Li, X.; Fu, P.; Cosh, M.H.; Letu, H.; Wang, S.; Chen, N.; et al. Surface soil moisture from combined active and passive microwave observations: Integrating ASCAT and SMAP observations based on machine learning approaches. *Remote Sens. Environ.* **2024**, *308*, 114197. [CrossRef]
19. Thomas, J.; Gupta, M.; Srivastava, P.K.; Pandey, D.K.; Bindlish, R. Development of High-Resolution Soil Hydraulic Parameters with Use of Earth Observations for Enhancing Root Zone Soil Moisture Product. *Remote Sens.* **2023**, *15*, 706. [CrossRef]
20. Nguyen, T.T.; Ngo, H.H.; Guo, W.; Chang, S.W.; Nguyen, D.D.; Nguyen, C.T.; Zhang, J.; Liang, S.; Bui, X.T.; Hoang, N.B. A low-cost approach for soil moisture prediction using multi-sensor data and machine learning algorithm. *Sci. Total Environ.* **2022**, *833*, 155066. [CrossRef]
21. Balenzano, A.; Mattia, F.; Satalino, G.; Lovergine, F.P.; Palmisano, D.; Peng, J.; Marzahn, P.; Wegmüller, U.; Cartus, O.; Dabrowska-Zielinska, K.; et al. Sentinel-1 soil moisture at 1 km resolution: A validation study. *Remote Sens. Environ.* **2021**, *263*, 112554. [CrossRef]
22. Pathe, C.; Wagner, W.; Sabel, D.; Doubkova, M.; Basara, J.B. Using ENVISAT ASAR Global Mode Data for Surface Soil Moisture Retrieval Over Oklahoma, USA. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 468–480. [CrossRef]
23. Rosenqvist, A.; Shimada, M.; Ito, N.; Watanabe, M. ALOS PALSAR: A Pathfinder mission for global-scale monitoring of the environment. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 3307–3316. [CrossRef]

24. Torres, R.; Snoeij, P.; Geudtner, D.; Bibby, D.; Davidson, M.; Attema, E.; Potin, P.; Rommen, B.; Floury, N.; Brown, M.; et al. GMES Sentinel-1 mission. *Remote Sens. Environ.* **2012**, *120*, 9–24. [CrossRef]
25. Yang, L.; Shi, L.; Sun, W.D.; Yang, J.; Li, P.X.; Li, D.R.; Liu, S.W.; Zhao, L.L. Radiometric and Polarimetric Quality Validation of Gaofen-3 over a Five-Year Operation Period. *Remote Sens.* **2023**, *15*, 1605. [CrossRef]
26. Banda, F.; Giudici, D.; Le Toan, T.; d'Alessandro, M.M.; Papathanassiou, K.; Quegan, S.; Riembauer, G.; Scipal, K.; Soja, M.; Tebaldini, S.; et al. The BIOMASS Level 2 Prototype Processor: Design and Experimental Results of Above-Ground Biomass Estimation. *Remote Sens.* **2020**, *12*, 985. [CrossRef]
27. Albinet, C.; Whitehurst, A.S.; Jewell, L.A.; Bugbee, K.; Laur, H.; Murphy, K.J.; Frommknecht, B.; Scipal, K.; Costa, G.; Jai, B.; et al. A Joint ESA-NASA Multi-mission Algorithm and Analysis Platform (MAAP) for Biomass, NISAR, and GEDI. *Surv. Geophys.* **2019**, *40*, 1017–1027. [CrossRef]
28. Davidson, M.; Kilbanek, J.; Lannini, L.; Furtiell, R.; Di Casino, G.; Gebert, N.; Petrolati, D.; Geiidtner, D.; Osborne, S. ROSE-L—The Copernicus L-Band Synthetic Aperture Radar Imaging Mission. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Pasadena, CA, USA, 16–21 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 4568–4571.
29. Tao, S.; Zhang, X.; Feng, R.; Qi, W.; Wang, Y.; Shrestha, B. Retrieving soil moisture from grape growing areas using multi-feature and stacking-based ensemble learning modeling. *Comput. Electron. Agric.* **2023**, *204*, 107537. [CrossRef]
30. Nie, Y.; Tan, Y.; Deng, Y.Q.; Yu, J. Suitability Evaluation of Typical Drought Index in Soil Moisture Retrieval and Monitoring Based on Optical Images. *Remote Sens.* **2020**, *12*, 2587. [CrossRef]
31. Zhu, S.; Cui, N.; Guo, L.; Jin, H.; Jin, X.; Jiang, S.; Wu, Z.; Lv, M.; Chen, F.; Liu, Q.; et al. Enhancing precision of root-zone soil moisture content prediction in a kiwifruit orchard using UAV multi-spectral image features and ensemble learning. *Comput. Electron. Agric.* **2024**, *221*, 108943. [CrossRef]
32. Cheng, M.; Jiao, X.; Liu, Y.; Shao, M.; Yu, X.; Bai, Y.; Wang, Z.; Wang, S.; Tuohuti, N.; Liu, S.; et al. Estimation of soil moisture content under high maize canopy coverage from UAV multimodal data and machine learning. *Agric. Water Manag.* **2022**, *264*, 107530. [CrossRef]
33. Zhang, Y.; Han, W.; Zhang, H.; Niu, X.; Shao, G. Evaluating soil moisture content under maize coverage using UAV multimodal data by machine learning algorithms. *J. Hydrol.* **2023**, *617*, 129086. [CrossRef]
34. Cheng, M.; Li, B.; Jiao, X.; Huang, X.; Fan, H.; Lin, R.; Liu, K. Using multimodal remote sensing data to estimate regional-scale soil moisture content: A case study of Beijing, China. *Agric. Water Manag.* **2022**, *260*, 107298. [CrossRef]
35. Cho, E.; Choi, M.; Wagner, W. An assessment of remotely sensed surface and root zone soil moisture through active and passive sensors in northeast Asia. *Remote Sens. Environ.* **2015**, *160*, 166–179. [CrossRef]
36. Jagdhuber, T.; Baur, M.; Akbar, R.; Das, N.N.; Link, M.; He, L.; Entekhabi, D. Estimation of active-passive microwave covariation using SMAP and Sentinel-1 data. *Remote Sens. Environ.* **2019**, *225*, 458–468. [CrossRef]
37. Guo, J.; Bai, Q.; Guo, W.; Bu, Z.; Zhang, W. Soil moisture content estimation in winter wheat planting area for multi-source sensing data using CNNR. *Comput. Electron. Agric.* **2022**, *193*, 106670. [CrossRef]
38. Mohamed, E.S.; Ali, A.; El-Shirbeny, M.; Abutaleb, K.; Shaddad, S.M. Mapping soil moisture and their correlation with crop pattern using remotely sensed data in arid region. *Egypt. J. Remote Sens. Space Sci.* **2020**, *23*, 347–353. [CrossRef]
39. Holtgrave, A.K.; Förster, M.; Greifeneder, F.; Notarnicola, C.; Kleinschmit, B. Estimation of Soil Moisture in Vegetation-Covered Floodplains with Sentinel-1 SAR Data Using Support Vector Regression. *PFG J. Photogramm. Remote Sens. Geoinf. Sci.* **2018**, *86*, 85–101. [CrossRef]
40. Zhao, X.; Huang, N.; Niu, Z.; Raghavan, V.; Song, X. Soil moisture retrieval in farmland using C-band SAR and optical data. *Spat. Inf. Res.* **2017**, *25*, 431–438. [CrossRef]
41. 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.; et al. A roadmap for high-resolution satellite soil moisture applications—Confronting product characteristics with user requirements. *Remote Sens. Environ.* **2021**, *252*, 112162. [CrossRef]
42. Wang, C.M.; Gu, X.F.; Zhou, X.; Yang, J.; Yu, T.; Tao, Z.; Gao, H.L.; Zhan, Y.L.; Wei, X.Q.; Li, J.; et al. Chinese Soil Moisture Observation Network and Time Series Data Set for High Resolution Satellite Applications. *Sci. Data* **2023**, *10*, 424. [CrossRef] [PubMed]
43. Li, B.Z.; Wang, C.M.; Ma, M.; Li, L.; Feng, Z.Z.; Ding, T.Y.; Li, X.F.; Jiang, T.; Li, X.J.; Zheng, X.M. Accuracy calibration and evaluation of capacitance-based soil moisture sensors for a variety of soil properties. *Agric. Water Manag.* **2022**, *273*, 107913. [CrossRef]
44. Blaes, X.; Defourny, P.; Wegmüller, U.; Della Vecchia, A.; Guerriero, L.; Ferrazzoli, P. C-band polarimetric indexes for maize monitoring based on a validated radiative transfer model. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 791–800. [CrossRef]
45. Trudel, M.; Charbonneau, F.; Leconte, R. Using RADARSAT-2 polarimetric and ENVISAT-ASAR dual-polarization data for estimating soil moisture over agricultural fields. *Can. J. Remote Sens.* **2012**, *38*, 514–527.
46. Pan, G.; Wang, C.; Zhang, W.; Wang, H.; Tian, G. Analysis of Seasonal Change of Land Cover Characteristics with SSM/I Data in China. *J. Remote Sens.* **2003**, *7*, 498–503.
47. Gao, B.-C. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sens. Environ.* **1996**, *58*, 257–266. [CrossRef]
48. Huete, A.R. A Soil-Adjusted Vegetation Index (SAVI). *Remote Sens. Environ.* **1988**, *25*, 295–309. [CrossRef]

49. Huete, A.; Didan, K.; Miura, T.; Rodriguez, E.P.; Gao, X.; Ferreira, L.G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sens. Environ.* **2002**, *83*, 195–213. [CrossRef]
50. Yao, Y.J.; Qin, Q.M.; Zhao, S.H.; Yuan, W.L. Retrieval of soil moisture based on MODIS shortwave infrared spectral feature. *J. Infrared Millim. Waves* **2011**, *30*, 9. [CrossRef]
51. Tucker, C.J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.* **1979**, *8*, 127–150. [CrossRef]
52. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
53. Chen, T.Q.; Guestrin, C.; Assoc Comp, M. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
54. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
55. Ma, H.; Liang, S. Development of the GLASS 250-m leaf area index product (version 6) from MODIS data using the bidirectional LSTM deep learning model. *Remote Sens. Environ.* **2022**, *273*, 112985. [CrossRef]
56. Qian, J.; Yang, J.; Sun, W.; Zhao, L.; Shi, L.; Dang, C. Evaluation and improvement of temporal robustness and transfer performance of surface soil moisture estimated by machine learning regression algorithms. *Comput. Electron. Agric.* **2024**, *217*, 108518. [CrossRef]
57. Li, Q.L.; Shi, G.S.; Shangguan, W.; Nourani, V.; Li, J.D.; Li, L.; Huang, F.N.; Zhang, Y.; Wang, C.Y.; Wang, D.G.; et al. A 1 km daily soil moisture dataset over China using in situ measurement and machine learning. *Earth Syst. Sci. Data* **2022**, *14*, 5267–5286. [CrossRef]
58. Han, Q.Q.; Zeng, Y.J.; Zhang, L.J.; Wang, C.; Prikaziuk, E.; Niu, Z.G.; Su, B.B. Global long term daily 1 km surface soil moisture dataset with physics informed machine learning. *Sci. Data* **2023**, *10*, 101. [CrossRef]
59. Shangguan, Y.L.; Min, X.X.; Wang, N.; Tong, C.; Shi, Z. A long-term, high-accuracy and seamless 1km soil moisture dataset over the Qinghai-Tibet Plateau during 2001–2020 based on a two-step downscaling method. *GISci. Remote Sens.* **2024**, *61*, 2290337. [CrossRef]
60. Zheng, C.L.; Jia, L.; Zhao, T.J. A 21-year dataset (2000–2020) of gap-free global daily surface soil moisture at 1-km grid resolution. *Sci. Data* **2023**, *10*, 139. [CrossRef]
61. Fang, B.; Lakshmi, V.; Cosh, M.; Liu, P.W.; Bindlish, R.; Jackson, T.J. A global 1-km downscaled SMAP soil moisture product based on thermal inertia theory. *Vadose Zone J.* **2022**, *21*, e20182. [CrossRef]
62. Song, P.L.; Zhang, Y.Q.; Guo, J.P.; Shi, J.C.; Zhao, T.J.; Tong, B. A 1 km daily surface soil moisture dataset of enhanced coverage under all-weather conditions over China in 2003–2019. *Earth Syst. Sci. Data* **2022**, *14*, 2613–2637. [CrossRef]
63. Abowarda, A.S.; Bai, L.; Zhang, C.; Long, D.; Li, X.; Huang, Q.; Sun, Z. 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.* **2021**, *255*, 112301. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.