


## Article

# Sensitive Feature Evaluation for Soil Moisture Retrieval Based on Multi-Source Remote Sensing Data with Few In-Situ Measurements: A Case Study of the Continental U.S.

Ling Zhang<sup>1,2</sup>, Zixuan Zhang<sup>3</sup>, Zhaohui Xue<sup>2</sup>  and Hao Li<sup>2,\*</sup>

<sup>1</sup> School of Naval Architecture & Ocean Engineering, Jiangsu Maritime Institute, Nanjing 211100, China; zhangling\_jmi@163.com

<sup>2</sup> School of Earth Sciences and Engineering, Hohai University, Nanjing 211100, China; zhaohui.xue@hhu.edu.cn

<sup>3</sup> Suzhou Surveying & Mapping Institute, Suzhou 215006, China; zixuan970209@163.com

\* Correspondence: lihao@hhu.edu.cn

**Abstract:** Soil moisture (SM) plays an important role for understanding Earth's land and near-surface atmosphere interactions. Existing studies rarely considered using multi-source data and their sensitiveness to SM retrieval with few in-situ measurements. To solve this issue, we designed a SM retrieval method (Multi-MDA-RF) using random forest (RF) based on 29 features derived from passive microwave remote sensing data, optical remote sensing data, land surface models (LSMs), and other auxiliary data. To evaluate the importance of different features to SM retrieval, we first compared 10 filter or embedded type feature selection methods with sequential forward selection (SFS). Then, RF was employed to establish a nonlinear relationship between the in-situ SM measurements from sparse network stations and the optimal feature subset. The experiments were conducted in the continental U.S. (CONUS) using in-situ measurements during August 2015, with only 5225 training samples covering the selected feature subset. The experimental results show that mean decrease accuracy (MDA) is better than other feature selection methods, and Multi-MDA-RF outperforms the back-propagation neural network (BPNN) and generalized regression neural network (GRNN), with the R and unbiased root-mean-square error (ubRMSE) values being 0.93 and 0.032 cm<sup>3</sup>/cm<sup>3</sup>, respectively. In comparison with other SM products, Multi-MDA-RF is more accurate and can well capture the SM spatial dynamics.

**Keywords:** soil moisture retrieval; random forest; multi-source remote sensing; feature selection; the continental U.S.



**Citation:** Zhang, L.; Zhang, Z.; Xue, Z.; Li, H. Sensitive Feature Evaluation for Soil Moisture Retrieval Based on Multi-Source Remote Sensing Data with Few In-Situ Measurements: A Case Study of the Continental U.S. *Water* **2021**, *13*, 2003. <https://doi.org/10.3390/w13152003>

Academic Editors: Kebiao Mao, Chunxiang Shi and Shibo Fang

Received: 11 June 2021

Accepted: 16 July 2021

Published: 21 July 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 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 usually defined as a volume of water stored within the unsaturated zone [1,2], and surface (0–5 cm) SM is an important variable associated with global terrestrial water, energy, and carbon cycles [3,4]. Therefore, it is necessary to obtain accurate and timely SM data.

Although traditional in-situ SM acquiring methods can provide accurate data each day, they only obtain scattered and limited point data [5]. In addition, it is impractical to deploy intensive monitoring stations around the world. Therefore, the in-situ measurements cannot well describe the spatial variability at large scale, especially when the measurements are sparse. Satellite microwave remote sensing is a more advocated method with the advantages of large-scale observation and high temporal resolution, which was proven to be a more effective way to estimate SM. Microwave remote sensing data, optical remote sensing data, and land surface models (LSMs) all provide products to estimate SM values.

For the microwave remote sensing data, several satellites were launched in the past two decades, carrying on board radiometer (passive), radar (active), or both sensors in various frequency bands with different spatial and temporal resolutions. Compared

with the active microwave products, passive microwave products have lower spatial resolutions, but the revisit periods are shorter. Therefore, passive microwave products are more conducive to large-scale SM retrieval and obtain timely SM values. Microwave remote sensing products also have different bands. Various low frequencies (X, C, and L bands) have been used to detect SM content [6]. Some multi-band sensors mainly include Advanced Microwave Scanning Radiometer-Earth Observing System (AMSR-E), the Advanced Microwave Scanning Radiometer 2 (AMSR2), the Advanced Scatterometer (ASCAT), and Fengyun-3B Microwave Radiation Imager (FY-3B), etc. L-band radiometers and radars mainly include Soil Moisture and Ocean Salinity (SMOS) and the Soil Moisture Active Passive (SMAP), which are more widely used these years since the L band is considered as the optimal band for SM monitoring due to its strong penetration ability of both soil and vegetable [7].

For the optical remote sensing data, related studies mainly used Moderate Resolution Imaging Spectroradiometer (MODIS), which provides many land products related to SM variations [8–14]. For example, Cui et al. [9] reconstructed the FY-3B SM product based on three MODIS products including 16-day NDVI, daily land surface temperature, and 16-day albedo. Kim et al. [8] used the MODIS-based EVI product as an indication for the vegetation condition at each site to assess the remote sensing SM products.

As for the LSMs, there are the European Center for Medium-range Weather Forecasts Re-Analysis Interim (ERA-Interim) [5,15–17], Global Land Data Assimilation System (GLDAS) [18–22], and Modern Era Retrospective Analysis for Research and Applications (MERRA) [23–25], which are also beneficial to estimating SM values. For example, Qu et al. [22] used the GLDAS Noah Land Surface Model L4 data as the reference to estimate SM values. Ge et al. [5] used the ERA-Interim product to train the deep convolutional neural network and neural network, the results suggest that the simulated SM values and the ERA-Interim SM agree relatively well at a global scale.

SM retrieval methods can use those multi-source data to estimate SM values, and different methods can be categorized into two classes: physical-driven models and data-driven models. Traditional SM retrieval methods are based on different physical models [26–37]. Although these physical models can estimate SM values by fitting geophysical parameters and in-situ measurements, they are lack of extendibility and flexibility. On the one hand, the model parameters are usually directly obtained from limited measurement values, which cannot extend to large areas. On the other hand, the complexity of physical models makes it difficult to flexibly construct the relationship between physical parameters and SM values.

Some SM retrieval methods are equipped with machine learning (ML) algorithms. ML-based models can handle a large amount of nonlinear data and flexibly combine information from multiple sources without explicit physical relationships. According to the usage of training samples, they can be divided into three types. The first type uses LSMs as the reference for the SM retrieval model [5,38]. The advantage of this method is that it is more suitable for large-scale SM retrieval since the LSMs have global space-time coverage, inducing sufficient training samples. The disadvantage is that each LSM has some uncertainties due to the model parameter estimation errors, which can be transferred to the SM retrieval model. The second type uses the satellite SM products as the training data [9,22,39]. This method has been widely used in SM downscaling studies to obtain high spatial resolution SM. Although high spatial resolution SM products can enhance ecological and hydrological applications, the accuracy is relatively low compared with in-situ measurements. The third type uses in-situ measurements as the reference. Xu et al. [40] designed a method based on generalized regression neural network (GRNN) to train SMAP products using in-situ measurements from five networks, and they found that GRNN has a good potential for retrieving SM. Eroglu et al. [41] used the in-situ measurements as a reference based on artificial neural network, which was capable of generating sub-daily and high-resolution SM predictions. The advantage of this method is that the in-situ measurements are closer to the true value than using LSMs or satellite SM products as

the reference, which is beneficial to improving the accuracy for SM retrieval. However, it is hard to obtain large-scale, adequate, and evenly distributed in-situ SM measurements. In addition, we should consider the problem of spatial matching between in-situ point measurements and satellite remote sensing data.

SM retrieval is a complex process, which depends on many interactive factors, such as soil texture, soil structure, the organic matter content, surface roughness, topography, and vegetation coverage [42,43]. It poses great challenges for accurate SM retrieval when facing with multi-source data, especially for multiple feature selection and fusion based on those factors. ML-based models have big potentials for SM retrieval considering those challenges. Among most of existing ML-based studies, the involved features are mainly from a single-source, and they are usually selected based on prior experience [5,9,22,40,44]. The features may contain noisy or redundant information without feature selection, which may decrease accuracy and increase computational cost [45]. Actually, feature selection methods can help us to understand the impact of different features on retrieval accuracy, which can also help to reduce noisy or redundant information, and avoid over-fitting in the SM retrieval model. To the best of our knowledge, very few ML-based studies have considered feature selection in SM retrieval based on multi-source data.

In this study, we designed a novel SM retrieval method (Multi-MDA-RF) using random forest (RF) based on 29 features derived from passive microwave remote sensing data, optical remote sensing data, LSMs, and other auxiliary data. The Multi-MDA-RF model is examined in a serial of comprehensive experiments: (1) we compared 10 filter or embedded type feature selection methods combined with sequential forward selection (SFS) to find the optimal feature subset for SM retrieval; (2) we analyzed the impacts of RF parameters on accuracy, including *mtry* and *ntree*; (3) our model was compared to back-propagation neural network (BPNN) and GRNN; (4) our product was compared with five popular SM products, including SMAP, AMSR2, SMOS, FY-3B, and ERA-Interim; (5) we analyzed the applicability of our model using in situ measurements from seven networks; (6) we visually inspected our product on three U.S. states with similar latitudes in the east, central, and west of CONUS; (7) we resampled the optimal features according to the lowest spatial resolution, to improve the spatial resolution of the input features and obtain a higher spatial resolution product.

It is worth noting that, some studies [40,44,46] are closely related to our work. However, there are some essential differences. Firstly, we considered multi-source data and generated 29 features as the inputs of the SM retrieval model, nevertheless other studies mainly used a single microwave remote sensing product. Secondly, the importance of different features was evaluated by using 10 feature selection methods, while existing studies did not exploit feature selection in SM retrieval. Thirdly, we used fewer training samples (i.e., a total of 5225), which was around one-third to a half of that reported in other studies, whereas achieving more accurate results with  $R = 0.93$ , which was around 0.03–0.26 higher than other studies. Finally, we produced a SM product with higher spatial resolution of  $0.125^\circ$ , which is around one-third of that reported in other studies. In this context, the main contribution and novelty of our work lie in that we proposed a novel Multi-MDA-RF model for SM retrieval, where multi-source data and 29 features were used as the inputs, and the importance of different features are evaluated. In addition, fewer training samples were used to produce more accurate results with higher spatial resolution. To the best of our knowledge, the proposed model is unique in the literature.

The rest of this paper is organized as follows. Section 2 describes the study area and the multi-source data used for the SM retrieval model. Section 3 illustrates the feature selection methods and presents the Multi-MDA-RF procedure for SM estimation. Section 4 reports the experimental results with a comprehensive comparison. Section 5 compares our model with other published state-of-the-art methods. Section 6 provides a summary and puts forward an outlook for future work.

## 2. Study Area and Data

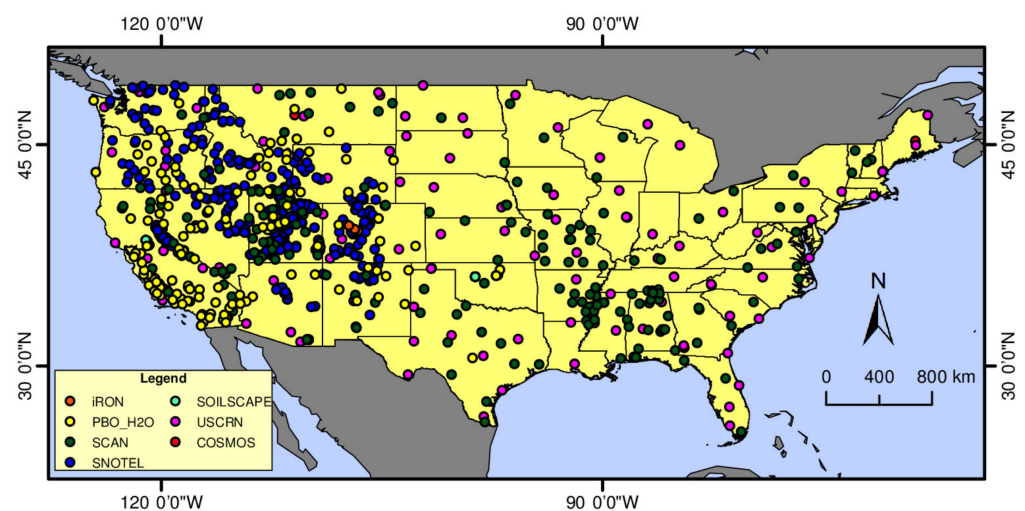
### 2.1. Study Area

The study area is the continental U.S. (CONUS), which borders on the North Atlantic and North Pacific [44]. The terrain of CONUS is complex and diverse. The east is composed of hills and low mountains, and the center is a vast plain. The west is the most complex area, consisting of the Colorado Plateau, Wyoming Plateau, Columbia Plateau, Grand Canyon, the Sierra Nevada Mountains, and the Cascade Mountains. There are many rivers and lakes in this area. The central and eastern regions have the largest freshwater lakes in the world, which is known as the North American Mediterranean. The climate of CONUS is mostly temperate and subtropical. The southeast is a subtropical climate zone with an average annual rainfall of 1500 mm, which is relatively humid. The climate in the western plateau is dry with large temperature differences, and the annual average rainfall is below 500 mm.

### 2.2. In-Situ Measurements

CONUS is an ideal study area for SM retrieval, because it has large-scale and long-term SM observation networks covering almost the entire continent. These abundant in-situ measurements are evenly distributed, which is conducive to the construction of SM retrieval models.

The in-situ measurements come from seven networks, which are spread across the whole CONUS: the Cosmic-ray Soil Moisture Observing System (COSMOS) [47–49], the Interactive Roaring Fork Observation Network (iRON) [50], PBO\_H2O [51], the Soil Climate Analysis Network (SCAN) [52,53], the Snow Telemetry (SNOTEL) [54], the Soil moisture Sensing Controller and oPtimal Estimator (SOILSCAPE) [55], and the U.S. Climate Reference Network (USCRN) [56]. All in-situ measurements can be downloaded from the International Soil Moisture Network (ISMN) (<http://ismn.geo.tuwien.ac.at/> accessed on 20 February 2021). ISMN has assembled over 50 operational and experimental SM networks around the world [57,58], providing uniform data format and pre-processing quality flags for global SM database [45]. Since the remote sensing satellite can only detect the surface SM, the in-situ measurements should be screened by retaining the surface SM measurements with a depth of 0–10 cm for COSMOS and 0–5 cm for others. The spatial distribution and the characteristics of the in-situ measurements are presented in Figure 1 and Table 1, respectively. The sites (total) in Table 1 refer to the total number provided by ISMN, and the sites (used) refer to the remaining number after the screening of depth.



**Figure 1.** Study area and the spatial distribution of the in-situ measurements.

**Table 1.** The characteristics of the in-situ measurements.

Network	Sites (Total)	Sites (Used)	Available Time	Depth (cm)	Temporal Resolution	Sensor
COSMOS	109	6	28 April 2008– 29 March 2020	0–10	Hourly	Cosmic-ray Probe
iRON	9	9	21 August 2012– 1 January 2020	5	Hourly	EC5 II, 10HS, EC5 I, HMP155, EC5
PBO_H2O	159	140	27 September 2004– 16 December 2017	0–5	Daily	GPS
SCAN	239	188	1 January 1996– Now	5	Hourly	n.s., 5.0 Volt, 2.5 Volt, linear
SNOTEL	441	130	1 October 1980– Now	5	Hourly	n.s., 5.0 Volt, 2.5 Volt
SOILSCAPE	171	114	3 August 2011– 29 March 2017	5	Hourly	EC5
USCRN	115	113	15 November 2000– Now	5	Hourly	Stevens HydraProbe II Sdi-12

### 2.3. Multi-Source Microwave Remote Sensing Data

National Aeronautics and Space Administration (NASA) launched the SMAP satellite in January 2015 to monitor global SM and landscape freeze-thaw conditions. On 31 March 2015, the L-band (1.41G Hz) radiometer was used to continuously collect scientific data [59]. Its nominal incident angle is 40° and it can achieve global coverage every 2–3 days. SMAP carries a radiometer and a radar to provide active and passive microwave remote sensing data at the same time. However, the radar stopped working after about three months of operation [60]. The ascending and descending overpasses of SMAP satellite are at 6 p.m. and 6 a.m., respectively, which are synchronized with the sun. As the thermal equilibrium of vegetation canopy and near-surface soil increases with temperature, SM retrieval is more stable in the early morning [7]. Therefore, we use SMAP Level-3 radiometer global daily 36-km EASEv2-grid soil moisture (SPL3SMP) descending overpass data, including brightness temperatures at vertical and horizontal polarization (SMAP\_TB<sub>V</sub>, SMAP\_TB<sub>H</sub>), the 4th Stokes' parameters (SMAP\_TB<sub>4</sub>) [38], surface temperature (SMAP\_Ts), vegetation water content (SMAP\_VWC), albedo (SMAP\_albedo), landcover classification (SMAP\_landcover), latitude, and longitude. SMAP are available from <https://nsidc.org/data/SPL3SMP> (accessed on 15 January 2021).

AMSR2 was launched by the Japan Aerospace Exploration Agency (JAXA) on 18 May 2012, and its first scientific observation began on July 3, 2012. As the successor of AMSR-E, AMSR2 continues to provide observations similar to AMSR-E. The orbit and basic settings of AMSR2 are consistent with AMSR-E. The ascending overpass of AMSR2 is 1:30 p.m., and the descending overpass was 1:30 a.m. [61]. We used the descending overpass, because it was closer to the early morning. We used AMSR2 Level-2 product with a spatial resolution of 25 km, including C-band 36GHz brightness temperatures at vertical and horizontal polarization (AMSR2\_TB<sub>V</sub>, AMSR2\_TB<sub>H</sub>), C-band surface temperature (AMSR2\_Ts), C-band optical depth (AMSR2\_optx), and X-band optical depth (AMSR2\_optc). AMSR2 are available from <https://search.earthdata.nasa.gov/> (accessed on 15 January 2021) and <https://suzaku.eorc.jaxa.jp/> (accessed on 15 January 2021).

SMOS was developed by the European Space Agency (ESA), which is the first polar orbit L-band radiometer. It was successfully launched on 2 November 2009, and has become one of the most important satellites for monitoring the global water cycle [62]. SMOS observations cover the globe approximately every 3 days with the ascending overpass at 6 a.m. and the descending overpass at 6 p.m., respectively. We used SMOS L3 ascending overpass observations with a spatial resolution of 25 km, including H and V polarization brightness temperature data (SMOS\_TB<sub>V</sub>, SMOS\_TB<sub>H</sub>), and optical depth (SMOS\_opt). SMOS data can be obtained from <https://smos-diss.eo.esa.int/oads/access/> (accessed on 20 February 2021).

FY-3B was successfully launched by the China National Space Administration (CNSA) on 5 November 2010, and the data service ceased on June 2020. It was equipped with a passive microwave radiometer called Microwave Radiation Imager (MWRI) [39]. MWRI provides observations with a total of five frequencies of 10.65, 18.7, 23.8, 36.5, and 89.0 GHz in both ascending (1:40 p.m.) and descending (1:40 a.m.) overpass. Each frequency has horizontal and vertical polarizations. We used 25km FY-3B level-1 10.65 GHz H and V polarization brightness temperature data with descending overpass (FY-3B\_TB<sub>V</sub>, FY-3B\_TB<sub>H</sub>). FY-3B data can be obtained from <http://satellite.nsmc.org.cn/portalsite/default.aspx> (accessed 20 February 2021).

The specific characteristics of all of the above microwave remote sensing products are reported in Table 2.

**Table 2.** The characteristics of the microwave remote sensing products.

Microwave Remote Sensing Product	Band	Spatial Resolution (km)	Temporal Resolution (days)	Available Time	Orbit
SMAP	L	36	~3	April 2015–Now	6:00 p.m. (A) 6:00 a.m. (D)
SMOS	L	25	~3	January 2010–Now	6:00 a.m. (A) 6:00 p.m. (D)
AMSR2	C/X	25	~2	July 2012–Now	1:30 p.m. (A) 1:30 a.m. (D)
FY-3B	X/Ku/K/Ka/E	25	~2	July 2011–June 2020	1:40 p.m. (A) 1:40 a.m. (D)

Note: A for ascending and D for descending.

#### 2.4. Auxiliary Data

MODIS was developed by NASA to understand global climate changes. We used three 0.05° MODIS products, including MOD13C2, MOD11C3 and MCD12C1. MOD13C2 provides monthly global Normalized Difference Vegetation Index (MODIS\_NDVI). MOD11C3 provides monthly night surface temperature (MODIS\_Ts). MCD12C1 is a global land-cover classification product (MODIS\_landcover). MODIS data are available from <https://modis.gsfc.nasa.gov/> (accessed on 13 December 2020).

ERA-Interim is a global atmospheric reanalysis product provided by the Medium-range Weather Forecasts (ECMWF) [5]. The time coverage of the product ranged from January 1979 to August 2019. We used surface roughness (ERA\_SR), surface temperature (ERA\_Ts), and albedo (ERA\_albedo) with a spatial resolution of 0.125° at 6 a.m. from this product. ERA-Interim can be obtained from <http://apps.ecmwf.int/datasets/> (accessed on 13 December 2020).

The 30-m Global Land Cover Dataset (GlobeLand30) is produced by National Geomatics of China [63]. The data covers the land area of 80° N to 80° S, which consisted of 10 land cover types. GlobeLand30 are available from <http://kmap.ckcest.cn/> (accessed on 13 December 2020). DEM data used in this experiment come from Shuttle Radar Topography Mission (SRTM) with a spatial resolution of 90 m. SRTM started in February 2000, and it covers an area of more than 119 million square kilometers between 60° N and 56° S. SRTM data can be obtained from <http://srtm.csi.cgiar.org/> (accessed on 13 December 2020). The soil texture data come from the 1 km Harmonized World Soil Database version 1.2 (HWSD), which contains 12 soil textures. HWSD can be obtained from <http://www.fao.org/> (accessed on 13 December 2020). The day of year (DOY) should also be considered as a necessary feature to reflect time variations.

### 3. Methodology

In this work, we designed a novel Multi-MDA-RF method to estimate SM values in CONUS. The presented method includes multi-feature generation, feature evaluation, and RF. Firstly, we matched multi-source data and in-situ measurements spatially and temporally to generate multiple features. Secondly, we evaluated these features by using various feature selection methods. According to the feature importance ranking, the

optimal subset was obtained by using SFS. Then, RF was employed to establish a nonlinear relationship between in-situ SM measurements and the optimal feature subset. Finally, the SM retrieval model was evaluated from many aspects in terms of different models, products, in-situ measurements, and U.S. states. A flowchart of this work is exhibited in Figure 2, and the details are described as follows.

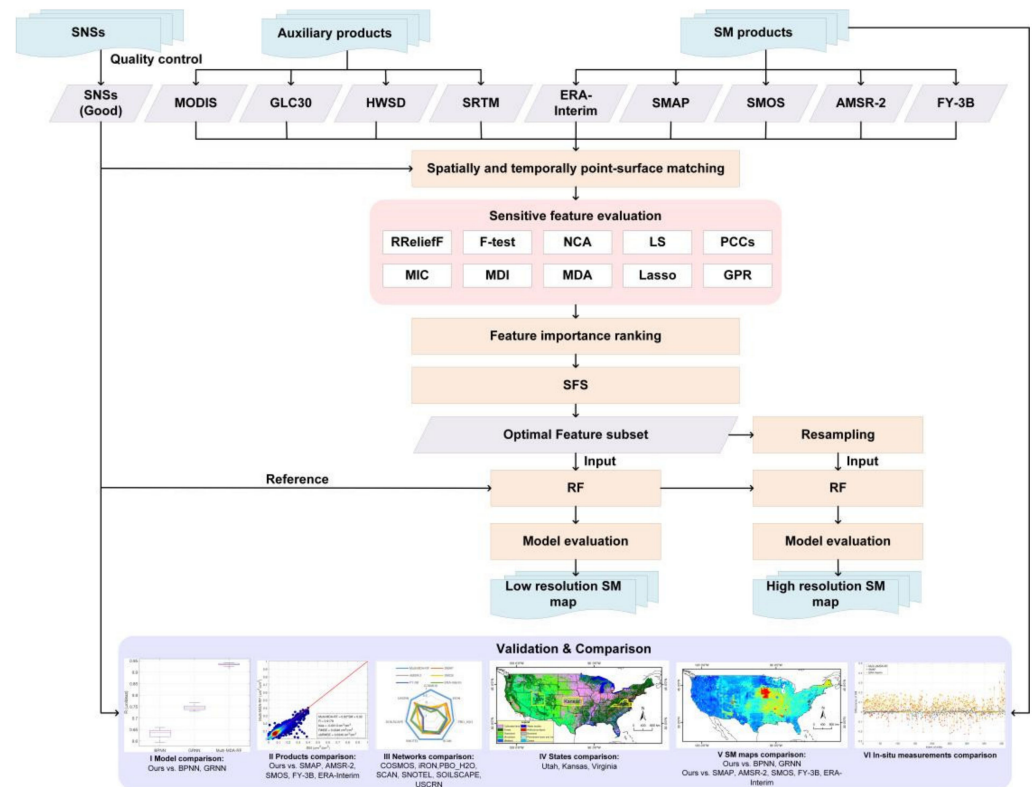


Figure 2. The flowchart of Multi-MDA-RF.

### 3.1. Multi-Feature Generation

For in-situ measurements, PBO\_H2O records SM data once at 12 p.m. every day, and the other networks record data per hour. We used the measurements recorded at 6 a.m. for all networks except for PBO\_H2O, which was in accordance with the time of satellite observations. In order to ensure the authenticity of in-situ SM measurements, we only selected the data with a quality mark “G (Good)”.

In order to comprehensively consider the variables related to SM retrieval, 29 multi-source features from passive microwave remote sensing data (SMAP, AMSR2, SMOS, FY-3B), optical remote sensing data (MODIS), LSM (ERA-Interim), and some other auxiliary data (GlobeLand30, SRTM, HWSD, DOY) were generated, as listed in Table 3. These features included brightness temperature data, surface parameters, vegetation parameters, soil texture, land use classification, geographical location, time, which are commonly used features in SM retrieval. There are some repetitive variables, which may be based on different criteria or methods, such as brightness temperature data from different passive microwave remote sensing data. We used feature selection to remove the redundant features among them, and then selected the features that were more suitable for SM retrieval.

For the multi-source microwave remote sensing data and the auxiliary data, they were converted to the same projection coordinate system of SMAP and resampled to a spatial resolution of 36 km.

**Table 3.** Multi-source data considered for SM retrieval.

Data	Index	Feature	Spatial Resolution	Description
SMAP	1	SMAP_TB <sub>H</sub>	36 km	Brightness temperatures (H)
	2	SMAP_TB <sub>V</sub>		Brightness temperatures (V)
	3	SMAP_TB <sub>4</sub>		4th Stokes' parameters
	4	SMAP_Ts		Daily surface temperature
	5	SMAP_VWC		Daily vegetation water content
	6	SMAP_albedo		Daily single-scattering albedo
	7	SMAP_landcover		Daily landcover classification
	8	Latitude		Center latitude
	9	Longitude		Center longitude
AMSR2	10	AMSR2_TB <sub>H</sub>	25 km	C-band brightness temperatures (H)
	11	AMSR2_TB <sub>V</sub>		C-band brightness temperatures (V)
	12	AMSR2_Ts		C-band daily surface temperature
	13	AMSR2_optc		C-band optical depth
	14	AMSR2_optx		X-band optical depth
FY-3B	15	FY-3B_TB <sub>H</sub>	25 km	X-band brightness temperatures (H)
	16	FY-3B_TB <sub>V</sub>		X-band brightness temperatures (V)
SMOS	17	SMOS_TB <sub>H</sub>	25 km	Brightness temperatures (H)
	18	SMOS_TB <sub>V</sub>		Brightness temperatures (V)
	19	SMOS_opt		optical depth
ERA-Interim	20	ERA_SR	0.125°	Daily surface roughness
	21	ERA_Ts		Daily surface temperature
	22	ERA_albedo		Daily albedo
GlobeLand30	23	GLC30_landcover	30 m	Landcover classification (2010)
MODIS	24	MODIS_NDVI	0.05°	Monthly Normalized Difference Vegetation Index
	25	MODIS_Ts		Monthly night surface temperature
	26	MODIS_landcover		Landcover classification (2015)
SRTM	27	DEM	90 m	Elevation
HWSD	28	Soil texture	1 km	Soil texture (FAO74)
DOY	29	DOY	\	Day of year

Then, the multi-source data should be matched with in-situ measurements, spatially and temporally. In order to relieve the scale difference between in-situ points and satellite pixels, the in-situ measurements within a multi-source data grid were averaged, which was adopted in many previous studies [5,64]. After the point-surface matching, 410 spatially isolated sites were available.

### 3.2. Sensitive Feature Evaluation Methods

The feature selection algorithms can be divided into filter, wrapper, and embedded type methods. The filter methods measure feature importance based on different indicators, which are independent of the adopted predictor. The wrapper methods train the predictor using a subset of features, and then add or remove a feature based on a selected criterion. The embedded methods obtain the feature importance in the model training process [65].

In this experiment, we used 10 filter or embedded type feature selection methods to obtain the importance ranking of each feature. Then, we combined these methods with



SFS to find the optimal feature subset. The filter type methods included regression ReliefF (RReliefF), F-test, neighborhood component analysis (NCA), Laplacian score (LS), Pearson correlation coefficient (PCCs), and maximal information coefficient (MIC). The embedded methods included mean decrease impurity (MDI), mean decrease accuracy (MDA), Lasso, and the feature optimization of Gaussian process regression (GPR) model.

### 3.2.1. The Filter Methods

1. RReliefF: RReliefF is inspired from Relief [66], which is very powerful in estimating the quality of features [67,68]. RReliefF penalizes the input features that give different values to neighbors with the same response values, and rewards the input features that give different values to neighbors with different response values. We used the 29 features as the input data and the in-situ measurements as the response values. The algorithm selects a random observation and finds the  $k$ -nearest observations to it. Then, the weight of SM features can be calculated as follows:

$$w = \frac{w_{dydx}}{w_{dy}} - \frac{w_{dr} - w_{dydr}}{m - w_{dy}} \quad (1)$$

where  $w_{dy}$  is the weight of having different values for the response  $y$ ,  $w_{dr}$  is the weight of having different values for the feature  $r$ ,  $w_{dydr}$  is the weight of having different response values and different values for the feature  $r$ ,  $m$  is the number of iterations. The importance ranking of each feature can be obtained according to this weight.

2. F-test: F-test is a statistical test by calculating the  $f$ -score of each feature [69]. We examined the importance of each feature individually using F-test, which calculates the values of  $f$ -score as follows, and the features were ranked based on  $f$ -scores.

$$f - \text{score} = -\log(p) \quad (2)$$

where  $p$  is the  $p$  values between features and in-situ measurements.

3. NCA: a novel nearest neighbor-based feature selection method was proposed by [70]. This feature selection method performs feature selection with regularization to learn feature weights for minimization of an objective function that measures the average leave-one-out regression loss over the training data. The objective function of minimization is as follows:

$$f(w) = \frac{1}{n} \sum_{i=1}^n l_i + \lambda \sum_{r=1}^p w_r^2 \quad (3)$$

where  $n$  is the number of observations,  $l_i$  is the distance between the in-situ measurements and  $y$ ,  $w_r$  the feature weight,  $\lambda$  is the regularization parameter,  $p$  is the average accuracy.

4. S: Laplacian score is a feature selection algorithm introduced by [71]. The locality preserving power for each feature was reflected by calculating the Laplacian score. Then, we can rank features using the Laplacian scores computed as follows:

$$L_i = \frac{\tilde{r}_i^T L \tilde{r}_i}{\tilde{r}_i^T D_g L \tilde{r}_i} \quad (4)$$

where  $r_i$  is the  $i$ -th feature,  $D_g$  is the degree matrix, and  $L$  is the Laplacian matrix.

5. PCCs: Pearson correlation coefficient is a simple method that can help to understand the relationship between features and response variables. This method measures the linear correlation between variables. The value range of the result is  $(-1, 1)$ , where “ $-1$ ” represents the complete negative correlation, “ $+1$ ” represents the complete positive correlation, and “ $0$ ” represents no linear correlation. The feature with the larger absolute value of the correlation is considered more important.

6. MIC: MIC is a powerful measure for relevance [72]. It is used to measure the degree of correlation between two variables  $r$  and  $y$ , and is often used in feature selection of machine learning. MIC can eliminate the feature with less information, so as to make the variable used in model more representative. MIC between the feature  $r$  and the response values  $y$  can be computed as follows:

$$\text{MIC}(r; y) = \max_{a \times b < B} \frac{I(r; y)}{\log_2 \min(a, b)} \quad (5)$$

where  $I(r; y)$  is the mutual information between  $r$  and  $y$ ,  $a, b$ , is the number of grids in  $r$  and  $y$  directions, and  $B$  is a variable, approximately set to the 0.6th power of the amount of data.

### 3.2.2. The Embedded Methods

1. MDI: RF based feature selection methods can be divided into MDI and MDA [73]. MDI computes feature importance for tree by summing changes in the mean squared error (MSE) due to splits on every feature and dividing the sum by the number of branch nodes. The importance of each feature segmentation is as follows:

$$\text{imp}(r_i) = (R_1 - R_2 - R_3) / N_{\text{branch}} \quad (6)$$

where  $R_i$  is the MSE of each node,  $N_{\text{branch}}$  is the total number of nodes.

2. MDA: MDA quantifies variable importance by measuring the change in prediction accuracy when the values of the variable are randomly permuted [74]. The importance of the feature  $r$  is then calculated using the following equation:

$$\text{imp}(r_i) = \bar{d}_r / \sigma_r \quad (7)$$

where  $\bar{d}_r$  is the average differences of the features,  $\sigma_r$  is the standard deviation of the features.

3. Lasso: this method trains a linear regression model with Lasso regularization. For a given value of  $\lambda$ , a nonnegative parameter, Lasso solves the problem:

$$w = \min_{\beta_0, \beta} \left( \frac{1}{2N} \sum_{i=1}^N (y_i - \beta_0 - r_i^T \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right) \quad (8)$$

where  $N$  is the number of observations,  $y_i$  is the response at observation  $i$ ,  $r_i$  is the  $i$ -th feature, a vector of length  $p$  at observation  $i$ ,  $\lambda$  is a nonnegative regularization parameter corresponding to one value of Lambda, and the parameters  $\beta_0$  and  $\beta$  are a scalar and a vector of length  $p$ , respectively.

4. GPR: This method is a feature selection method of GPR model [75]. It trains a GPR model and finds the predictor weights by taking the exponential of the negative learned length scales. Then, we can normalize the weights and obtain the importance ranking.

### 3.2.3. Sequential Forward Selection (SFS)

SFS is a bottom-up search procedure [76], in which the features are added to an empty set in the specified order until the addition of further features does not decrease the criterion. The criterion used in this experiment was the root mean square error (RMSE), and Pearson correlation coefficient (R) serves as a reference. The order of forward input features was the importance ranking of each feature selection method.

### 3.3. The Random Forest (RF) Method

RF was developed by [77], which is a popular method in *Applied Statistics* field to solve classification and regression problems using multiple decision trees. One advantage

of RF is that it has powerful generalization performance by using multiple regression trees, which is beneficial to reducing the variability of the model. Another advantage of RF is that it does not require complex parameter adjustments since it only has two parameters: the number of trees (*ntree*) and the number of features (*mtry*). RF selects features from the entire set using replacement sampling to establish the decision tree. To model the relationship between SM values and sparse network stations, a set of training input–output pairs should be given. The input variables are the optimal subset selected from the 29 features, and the output variables are the in-situ SM measurements. Once the SM retrieval model is trained, we then can estimate SM values by feeding the new samples into the model.

### 3.4. Evaluation Method

Four commonly used error metrics including RMSE, the mean bias (bias), R, and the unbiased root mean square error (ubRMSE) were used to evaluate the performance between the SM products and the in-situ SM measurements [78]. Those error metrics are defined as follows:

$$R = \frac{E[(x - E[x])(y - E[y])]}{\sigma_x \sigma_y} \quad (9)$$

$$RMSE = \sqrt{E[(x - y)^2]} \quad (10)$$

$$bias = E[x] - E[y] \quad (11)$$

$$ubRMSE = \sqrt{E[((x - E[x]) - (y - E[y]))^2]} \quad (12)$$

where  $x$  is the SM product,  $y$  is the in-situ measurements,  $\sigma_x$  and  $\sigma_y$  indicate the standard deviation of  $x$  and  $y$ , respectively.

## 4. Results

### 4.1. Experimental Settings

For validation set, there were a total of 1999 samples covering all 29 features. We split the validation set into training (60%) and test (40%). We further screened the training set after selecting the optimal feature subset, and there were 5225 training samples left. For the two parameters of RF, *mtry* and *ntree* in the RF model were experimentally set to 4 and 100, respectively, for feature selection. All experimental results were reported by averaging the outputs of 20 independent runs in terms of randomly initializing the training set. Note that our experiments were carried on a personal computer (Intel Core 2.40 GHz processor with 8 GB random access memory). The software implementation was performed using MATLAB (The MathWorks Inc., Natick, MA, USA).

### 4.2. Selection of Sensitive Features

Initially, 29 features from SMAP, AMSR2, SMOS, FY-3B, ERA-interim, MODIS, GlobeLand30, SRTM, HWSD and DOY were used. In this experiment, 10 filter and embedded type feature selection methods with SFS were considered. Different feature selection methods could obtain the importance weights for each feature, and the weights were normalized to 0–1 for a fair comparison. The corresponding feature importance of each feature selection method is shown in Figure 3. The abscissa is 28 features, except DOY, and the ordinate is the stacked importance. Figure 3 shows that the feature importance calculated by different feature selection methods are discrepant. Some features gained high importance in a variety of feature selection methods, such as latitude and MODIS\_NDVI. These features may be more sensitive to SM. Some features are less important after integrating various feature selection methods, such as SMAP\_Tb4. It may reduce the accuracy of SM retrieval and increase the uncertainty.

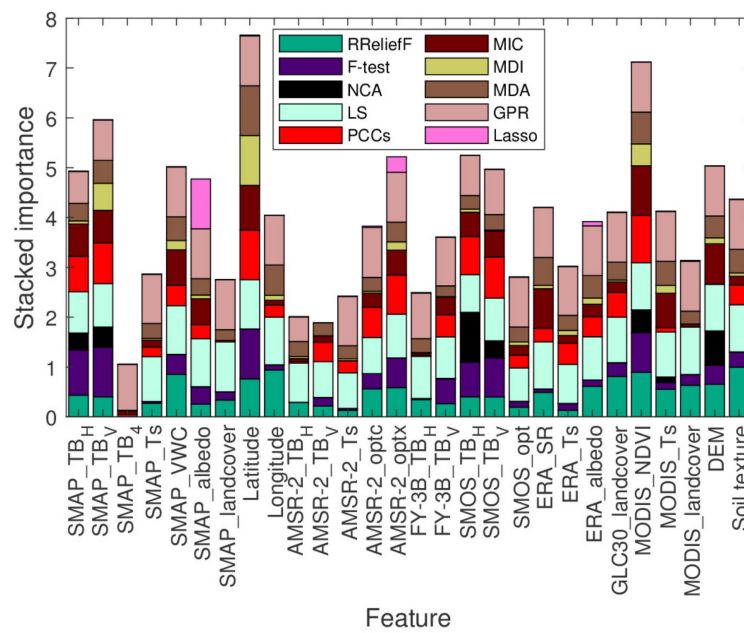


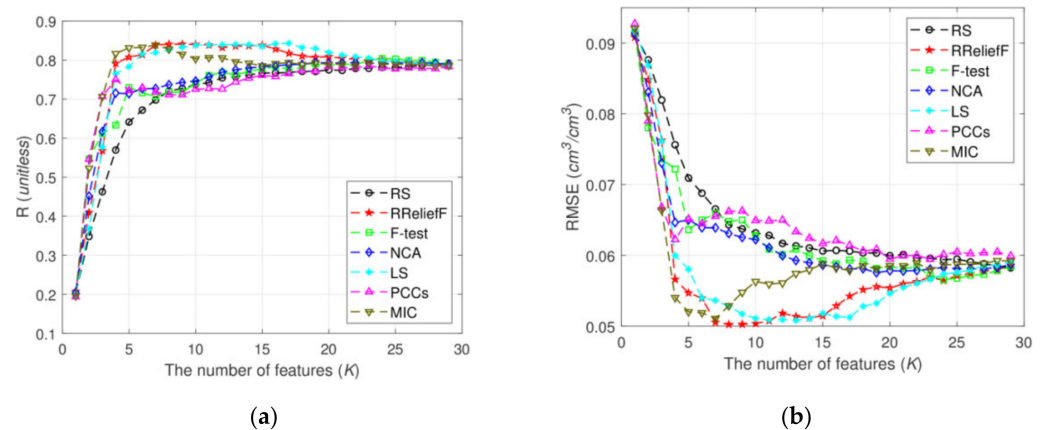
Figure 3. Stacked bars of feature importance obtained by different feature selection methods.

Then, we used the obtained importance to rank these features as listed in Table 4. The data reported in the table represent the importance rankings of different features based on different feature selection methods, and the features with high rankings were considered more sensitive for SM retrieval. It can be found that, even for the same feature, the importance ranking of different feature selection methods is quite different. Even both MDI and MDA methods are based on RF; they have different evaluation results for the same feature. Therefore, it is necessary to find a better feature selection method to get the optimal feature subset.

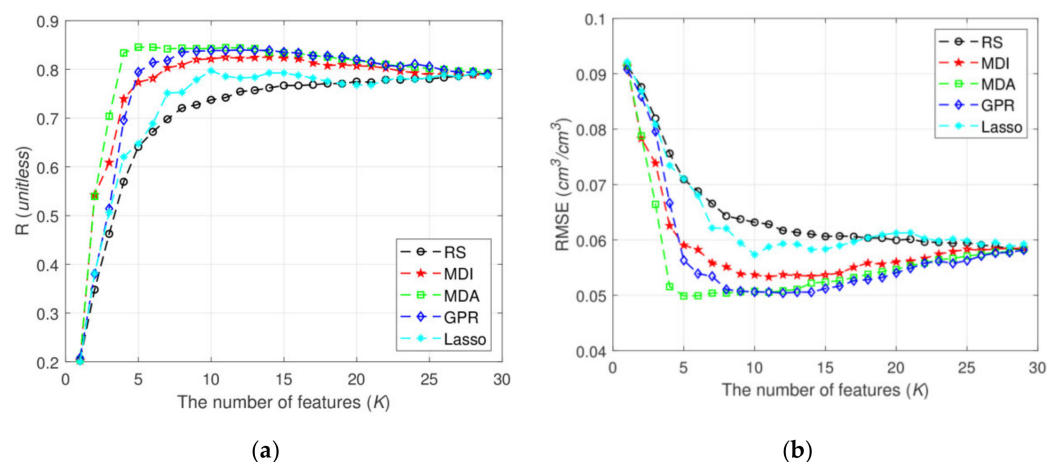
Table 4. The importance ranking of different feature selection methods.

Feature	The Importance Ranking									
	Filter					Embedded				
	RReliefF	F-Test	NCA	LS	PCCs	MIC	MDI	MDA	GPR	Lasso
SMAP_TBH	14	3	6	20	7	8	15	13	26	10
SMAP_TBV	16	2	4	15	4	7	2	8	24	26
SMAP_TB4	28	28	16	28	23	23	22	28	21	11
SMAP_Ts	21	24	18	13	21	20	20	18	18	12
SMAP_VWC	4	9	13	3	12	5	4	6	13	13
SMAP_albedo	23	11	25	4	16	10	11	14	1	1
SMAP_landcover	19	17	9	1	25	28	27	27	2	14
Longitude	6	1	20	2	1	2	1	3	6	6
AMS2_TBH	2	22	15	5	19	22	10	3	4	27
AMS2_TBV	20	27	28	22	24	24	17	17	27	15
AMS2_Ts	24	16	27	26	15	21	28	24	28	23
AMS2_optc	27	25	14	25	20	27	21	23	17	16
AMS2_optx	11	12	21	24	8	14	19	21	15	4
FY-3B_TBH	10	7	22	14	5	11	5	11	14	2
FY-3B_TBV	18	26	17	18	26	26	26	20	22	7
SMOS_TBH	22	8	8	19	10	13	24	26	20	17
SMOS_TBV	15	6	1	23	6	12	16	22	25	28
SMOS_opt	17	5	5	17	3	9	23	15	23	18
ERA_SR	25	21	23	27	18	17	12	16	16	19
ERA_Ts	13	23	10	9	17	4	14	4	5	25
ERA_albedo	26	19	26	21	11	19	9	19	19	20
GLC30_landcover	9	20	24	16	13	15	7	9	12	3
MODIS_NDVI	5	14	12	11	9	16	18	12	6	24
MODIS_Ts	3	4	3	8	2	1	3	2	7	9
MODIS_landcover	12	18	7	12	22	6	6	5	8	21
DEM	8	15	19	6	27	25	25	25	9	5
Soil texture	7	10	2	10	28	3	8	10	10	8
Soil texture	1	13	11	7	14	18	13	7	11	22

With the importance ranking at hand, we then used SFS to stack features one-by-one, and used the RF model to test the stacked features. It is worth noting that DOY is a necessary feature to reflect the temporal variation of SM. The SM retrieval results after feature selection are shown in Figure 4 (filter type) and Figure 5 (embedded type), where we plot the retrieval accuracies (using R and RMSE) obtained by different feature selection methods as a function of the number of features ( $K$ ). We also included a random feature selection (RS) method to validate the effects of feature selection. The ranking of features was randomly shuffled in the RS method.



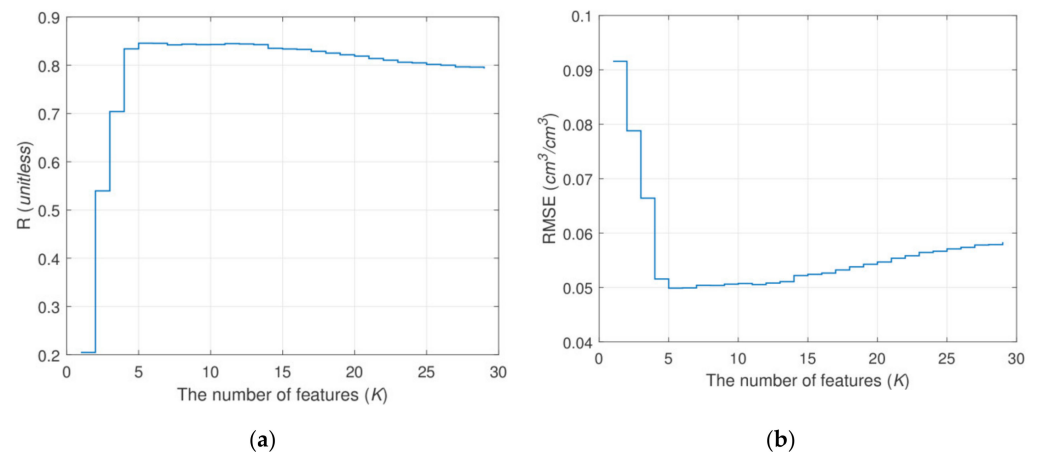
**Figure 4.** (a) R and (b) RMSE based on different filter type feature selection methods as a function of the number of features.



**Figure 5.** (a) R and (b) RMSE based on different embedded type feature selection methods as a function of the number of features.

Generally, the curves obtained by embedded type methods are smoother than that of filter type methods, which is due to the fact that the filter type methods are independent of the adopted predictor. The overall trends of different feature selection methods are consistent, i.e., the accuracies gradually increase to the maximum value and then decrease, which verifies that too many features will lead to accuracy reduction in the SM retrieval models. Specifically, among the filter type methods, RReliefF achieves a minimum RMSE value of  $0.0502 \text{ cm}^3/\text{cm}^3$ , and LS achieves a maximum R value of 0.8431. Whereas, among the embedded type methods, all feature selection methods are better than RS, which shows the effectiveness of feature selection. MDA obtained a minimum RMSE value of  $0.0499 \text{ cm}^3/\text{cm}^3$  and a maximum R value of 0.8457 with the top 5 features. Compared with all 29 features as input, R value increased by 0.0527 and RMSE value decreased by  $0.0085 \text{ cm}^3/\text{cm}^3$ , indicating that feature selection is worthy of attention in SM retrieval problems.

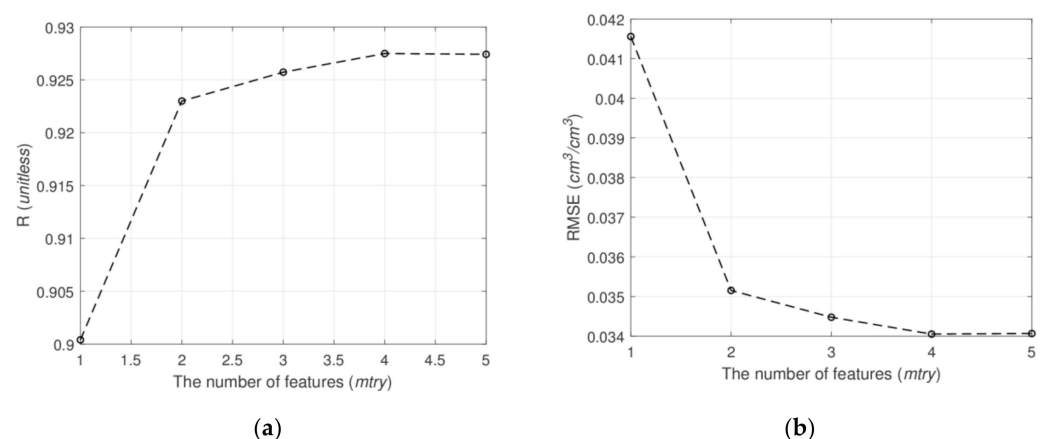
As we can see, MDA is the best feature selection method since it achieves the smallest RMSE and highest R values. Therefore, we chose to use the optimal features selected by MDA. In order to observe the results more clearly and to find the optimal  $K$  value, we drew the step diagrams of RMSE and R values obtained by MDA in Figure 6. Our retrieval model achieved the best accuracy when  $K = 5$ . Therefore, the optimal feature subset was determined as DOY, latitude, MODIS\_NDVI, longitude, and ERA\_SR. The five selected features include geographic location (latitude, longitude), time (DOY), surface parameter (ERA\_SR), and vegetation parameter (MODIS\_NDVI). These factors are closely related to SM retrieval, which also shows that the feature selection method is scientific in this work.



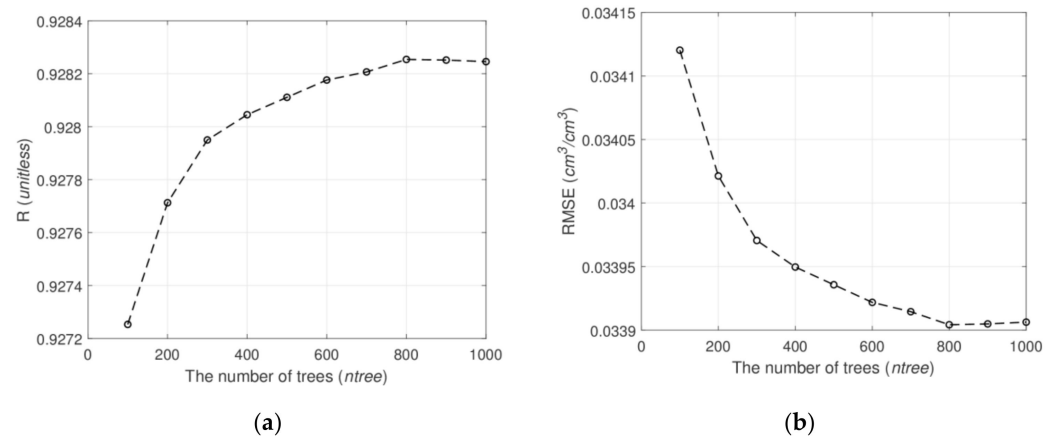
**Figure 6.** Step diagrams of (a) R and (b) RMSE based on MDA as a function of the number of features.

#### 4.3. Parameters Selection for Multi-MDA-RF

In order to analyze the impacts of parameters on SM retrieval accuracy, we tested two parameters, including  $mtry$  and  $n tree$  and determined their optimum values. In this experiment, we experientially set  $mtry$  to (1, 2, ..., 5), and set  $n tree$  to (100, 200, ..., 1000). The variations of R and RMSE values against  $mtry$  and  $n tree$  parameters are depicted in Figures 7 and 8, respectively. It can be seen that the best results appear when  $mtry = 4$  and  $n tree = 800$ . To sum up, the  $mtry$  parameter is set to 4 and the  $n tree$  parameter is set to 800 in the following experiments.



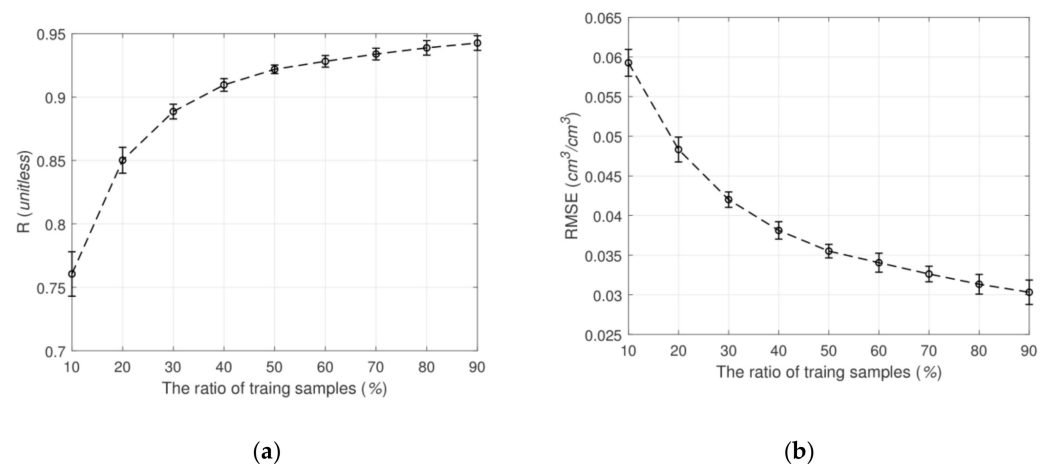
**Figure 7.** Effects of  $mtry$  parameter on the SM retrieval performance. (a) R and (b) RMSE.



**Figure 8.** Effects of ntree parameter on the SM retrieval performance. (a) R and (b) RMSE.

#### 4.4. Generalization Performance Analysis

In general, the performance could be better when we use more training samples since the model will be well-trained with sufficient prior knowledge. In this experiment, we tested the generalization performance in our model under different numbers of training samples. To this end, the ratio of training samples was set to (0.1, 0.2, . . . , 0.9). As shown in Figure 9, the accuracy increases as the ratio of training samples also increases, which confirms our assumption. In the following experiments, the ratio of training samples is set to 0.7, to obtain better results and retain sufficient validation data.



**Figure 9.** Effects of the ratio of training samples (%) on the SM retrieval performance. (a) R and (b) RMSE.

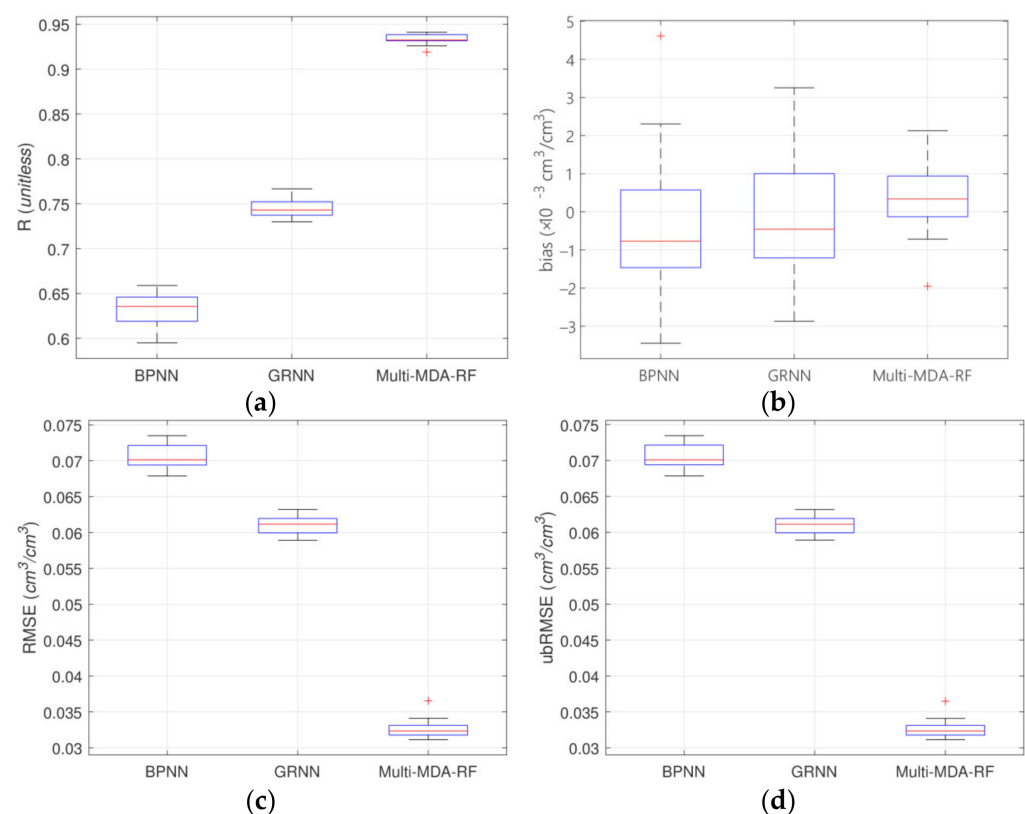
#### 4.5. Evaluation of Different Retrieval Models

Based on the above experiments, the proposed Multi-MDA-RF model is determined. In order to give a comprehensive and reliable analysis of our model, BPNN and GRNN are chosen for comparison with the same training samples. The results of different models are summarized in Table 5. It can be observed that Multi-MDA-RF obtains much higher accuracy than BPNN and GRNN, with R (ubRMSE) values improved by 0.30 (0.039 cm³/cm³) and 0.19 (0.029 cm³/cm³) for BPNN and GRNN, respectively. For the operation time, our model is reasonable with a time cost of 37 s. However, GRNN takes the longest time among the three models. Although BPNN has a shorter calculation time, its accuracy is lower than the other models.

**Table 5.** The results of Multi-MDA-RF model compared to BPNN and GRNN. The bold numbers indicate better evaluation results.

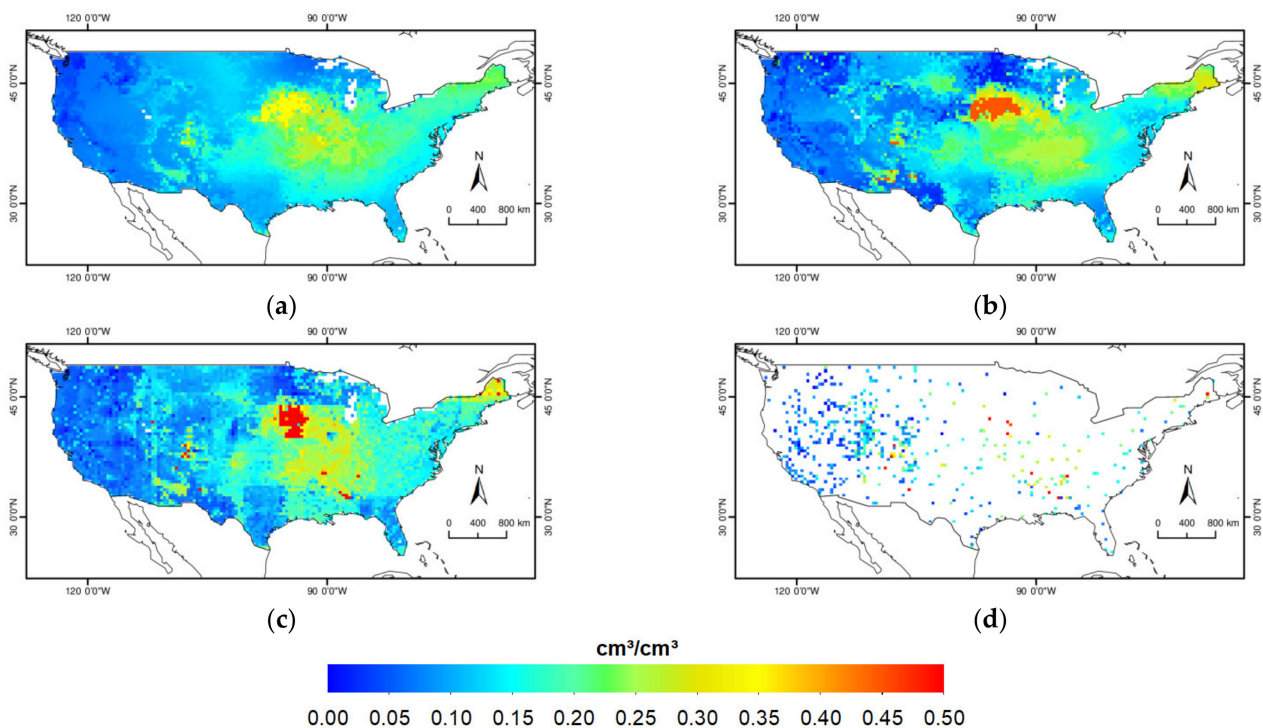
Model	Training Samples	Testing Samples	Time (s)	R	Bias ( $\text{cm}^3/\text{cm}^3$ )	RMSE ( $\text{cm}^3/\text{cm}^3$ )	ubRMSE ( $\text{cm}^3/\text{cm}^3$ )
BPNN	5225	2239	2	0.63	0.000	0.071	0.071
GRNN	5225	2239	114	0.74	0.000	0.061	0.061
Multi-MDA-RF	5225	2239	37	<b>0.93</b>	0.000	<b>0.033</b>	<b>0.032</b>

We then drew box plots in Figure 10 to analyze the stabilities of different models. More intuitively, it can be seen that the results of our model are more compact, indicating that it is more stable than the other two models. Among the three models, BPNN is the least accurate and the most unstable one. SM retrieval maps based on the three models are generated in Figure 11, which displays the mean SM maps during August 2015 for different models and the in-situ measurements. From a spatial perspective, all three models show low SM values in the west, with an increase toward the east, which agrees with climate of CONUS. However, both BPNN and GRNN underestimate SM values, especially in the center of CONUS. The result of our model illustrated in Figure 11c is wetter than BPNN and GRNN models, which is well matched with the spatial pattern of the in-situ measurements.



**Figure 10.** Box plots for (a) R, (b) bias, (c) RMSE, and (d) ubRMSE obtained by BPNN, GRNN, and Multi-MDA-RF. The central mark indicates the median, while the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers plotted individually using the “+” symbol.





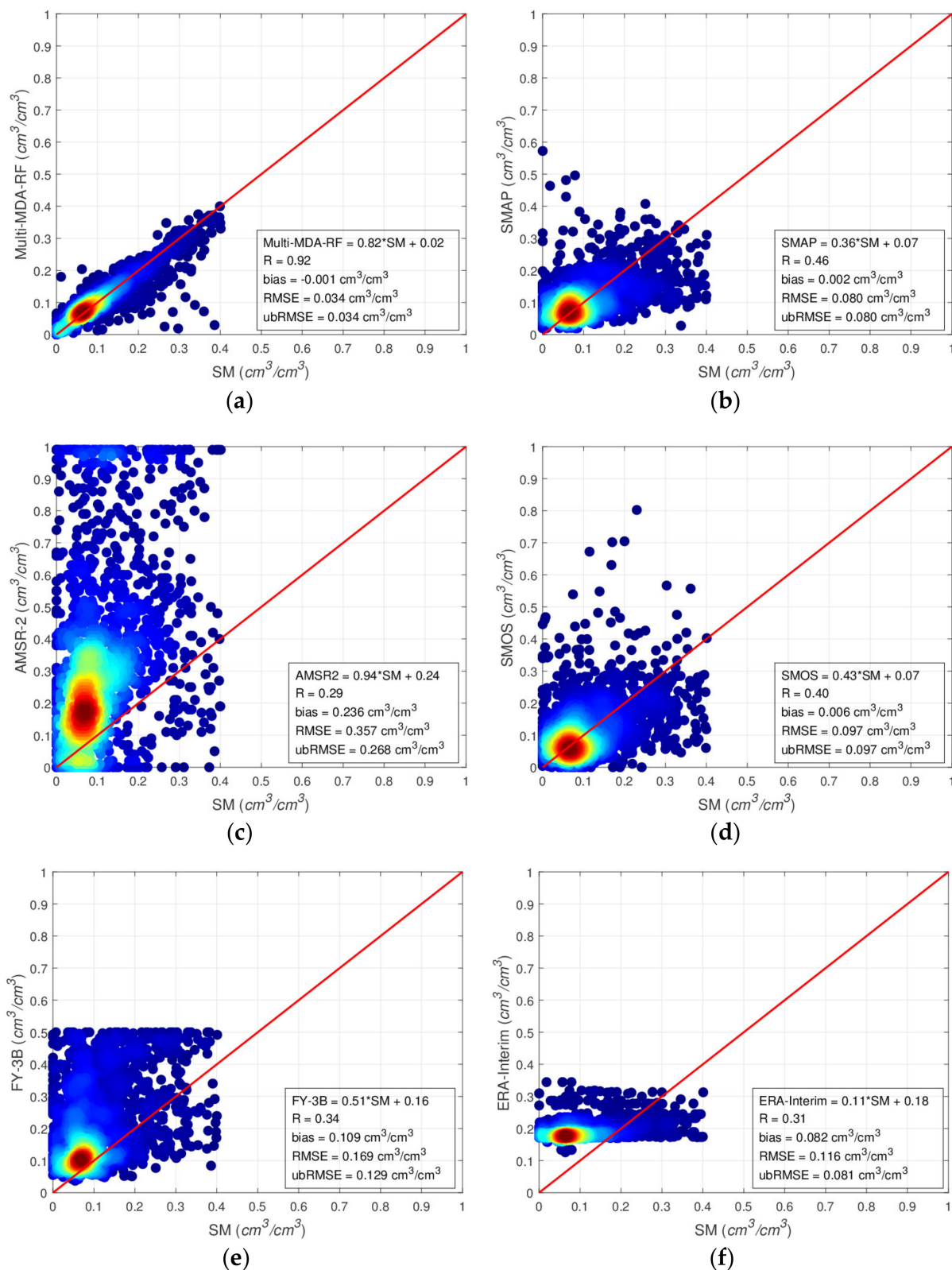
**Figure 11.** Mean SM maps of August 2015 obtained by different models. (a) BPNN, (b) GRNN, (c) Multi-MDA-RF, and (d) in-situ measurements.

#### 4.6. Evaluation of Different SM Products

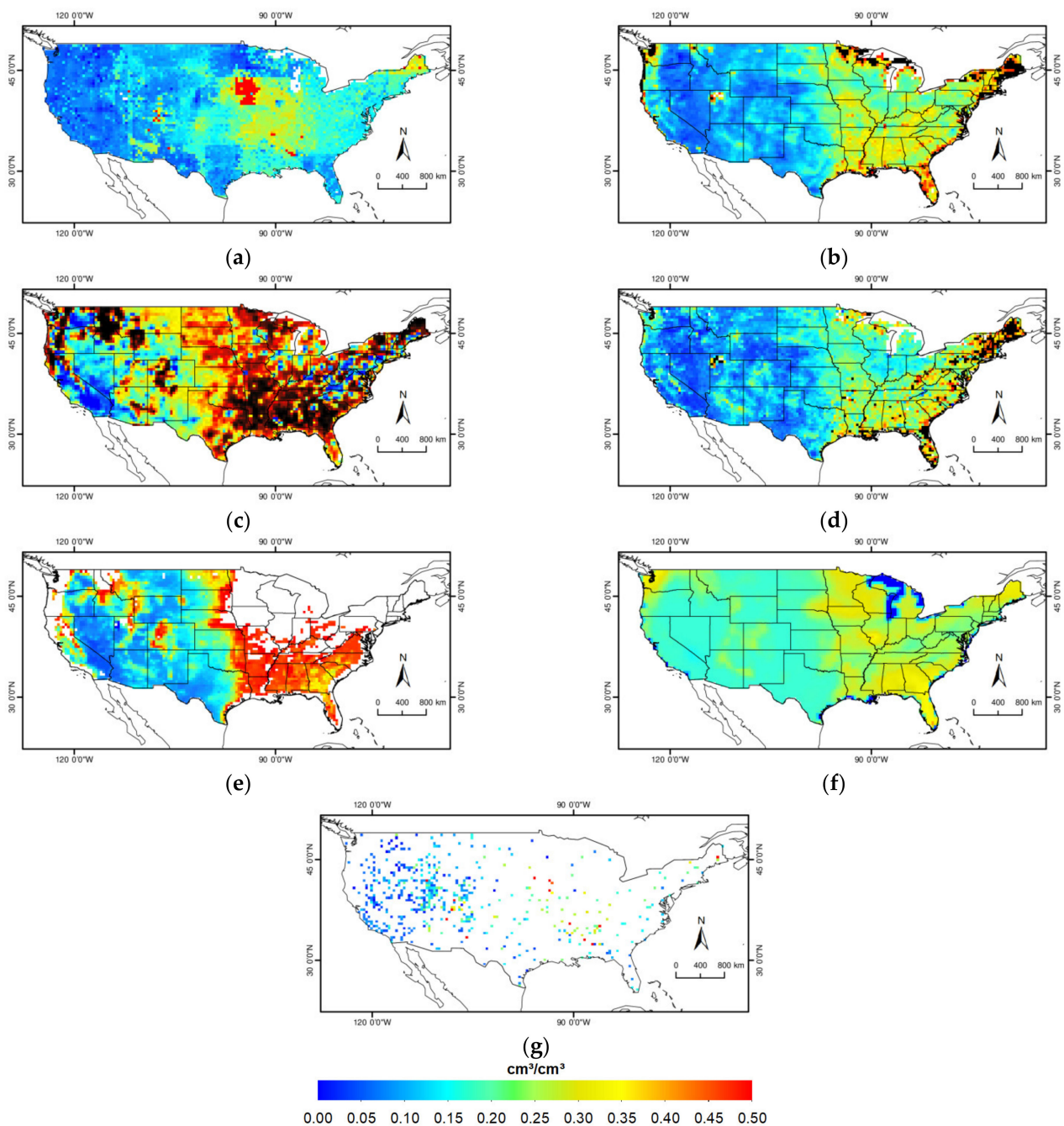
In this experiment, Multi-MDA-RF is compared with other five SM products including SMAP, AMSR2, SMOS, FY-3B, and ERA-Interim. We used 1577 points of spatiotemporal matching to draw scatter plots as shown in Figure 12. According to the scatter plots, Multi-MDA-RF matches the 1:1 line better than the other products, indicating that our product has the highest accuracy. Most SM values of SMAP, SMOS, and FY-3B agree with the in-situ measurements, but there are still some deviation values that make the accuracy relatively low. AMSR2 is less accurate with the R value of 0.29, and its SM values deviate from the 1:1 line. Most SM values of ERA-Interim are concentrated between 0.1 and 0.3  $\text{cm}^3/\text{cm}^3$ , which are inconsistent with the in-situ measurements. Note that the satellite data used in our model are not screened, and we did not conduct quality control for all of the considered products for a fair comparison, which results in the deviations of the other products.

According to the mean SM maps (Figure 13), our product can well capture the spatial dynamics and outperform the other products. The results of SMAP and SMOS are not “good”, and their SM estimates involve too many black pixels near the continental margins and water systems, where the water bodies or wet soil may lead to the poor performance. AMSR2 seems to be the worst with a lot of black pixels and it overestimates the SM values, especially in densely vegetated areas. This may be because C-band satellites do not have capacity to penetrate vegetation as well as L-band satellites [79]. Note that, those black pixels in Figure 13b–d represent that the corresponding SM values are greater than 0.5  $\text{cm}^3/\text{cm}^3$ , which are considered as outliers [80]. Therefore, SMAP, SMOS, and AMSR2 need conducting carefully quality control to avoid errors in practical applications. FY-3B is also very poor, and there are a lot of gaps in the map. The gaps in satellite-based SM products are intrinsic due to satellite orbits and retrieval algorithms [9]. The data gaps in CONUS accounts for about 25%, which makes the application value of FY-3B extremely limited. In addition, microwave remote sensing products are easily affected by Radio-Frequency Interference (RFI), which may be another reason for the low accuracies of these products. ERA-Interim overestimates the SM values in the east and underestimates the SM values in the west of CONUS. Its estimations are too smooth to reflect variations

since the distribution of SM values cannot be so uniform in practice. This may be because ERA-interim is a reanalysis product from the data assimilation system, the product itself has certain deviations.



**Figure 12.** Scatter plots of (a) Multi-MDA-RF, (b) SMAP (c) AMSR2, (d) SMOS, (e) FY-3B, and (f) ERA-Interim. The red line is the 1:1 line.

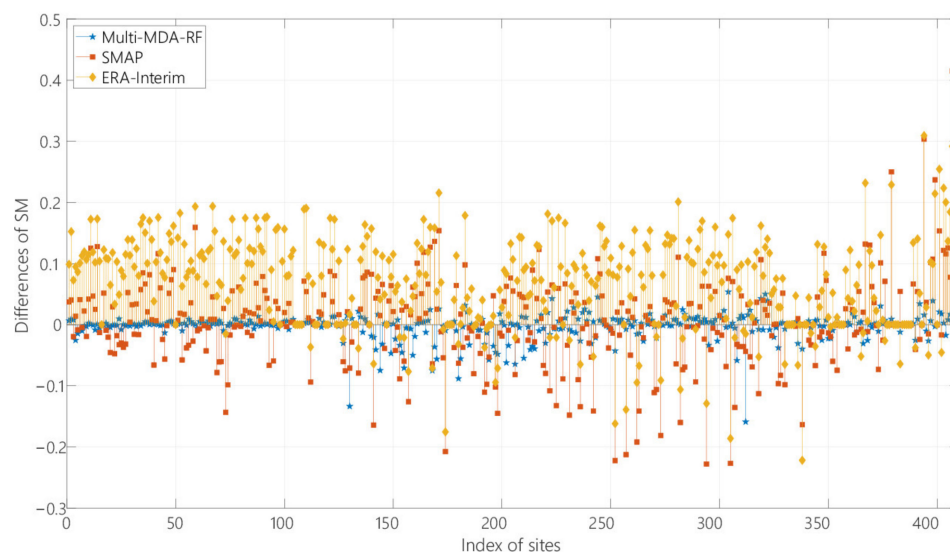


**Figure 13.** Mean SM maps of August 2015 for (a) Multi-MDA-RF, (b) SMAP, (c) AMSR2, (d) SMOS, (e) FY-3B, (f) ERA-Interim, and (g) in-situ measurements. Black pixels are considered as outliers.

In this context, our product is more consistent with the in-situ measurements, which means it outperforms the other compared products, including official satellites and LSM. In addition, the obtained scatter plot and the map of our product are fairly good, demonstrating it is able to accurately describe the relationship between the selected features and in-situ measurements. Therefore, Multi-MDA-RF showed the best predictive ability and can well capture spatial variations in SM even without quality control.

To further evaluate the performance of the proposed method, we selected two typical products, including SMAP and ERA-Interim using 410 spatially isolated sites. Figure 14 shows the monthly mean difference between Multi-MDA-RF, SMAP, and ERA-Interim with in-situ measurements. It can be seen that our product has the smallest difference, i.e.,

most of the points are closer to the 0 line, indicating the good accuracy and potential of Multi-MDA-RF in SM retrieval.



**Figure 14.** Difference diagram by comparing Multi-MDA-RF, SMAP, EAR-Interim with in-situ measurements.

4.7. Evaluation of Different SM Networks

In order to evaluate the predictive power of the Multi-MDA-RF model over each network, Table 6 and Figure 15 show the specific evaluations at different networks. It can be observed that our model performs slightly better at SCAN and USCRN than the other five networks. This is because the two networks have sufficient, widespread and uniform in-situ measurements, which are ideal SM observation networks. Table 6 shows that the R values of PBO\_H2O and SOILSCAPE are lower than the others. This may be due to the fact that most of the sites of these two networks are located in rugged areas of the west CONUS. In addition, PBO\_H2O only records SM data at 12 p.m. once a day. However, the satellite data used in the experiment are close to 6 a.m., which may be another reason for the low correlation of this network. Box plots in Figure 15 show that the results of SCAN, SNOTEL and USCRN are compact and stable. However, the results of COSMOS, iRON, and SOILSCAPE are unstable. This is mainly due to the fact that only a few sites of these three networks participate in model training, resulting in unstable prediction results.

**Table 6.** Average statistics of the evaluation of SM retrieval against in-situ measurements over each network.

Network	Training Samples	R	Bias (cm <sup>3</sup> /cm <sup>3</sup> )	RMSE (cm <sup>3</sup> /cm <sup>3</sup> )	ubRMSE (cm <sup>3</sup> /cm <sup>3</sup> )
COSMOS	48	0.95	−0.013	0.050	0.0489
iRON	36	0.94	−0.002	0.029	0.029
PBO_H2O	1640	0.78	0.000	0.028	0.028
SCAN	2197	0.96	0.002	0.028	0.028
SNOTEL	2569	0.95	−0.001	0.030	0.030
SOILSCAPE	49	0.88	−0.008	0.026	0.024
USCRN	1481	0.96	0.003	0.029	0.029

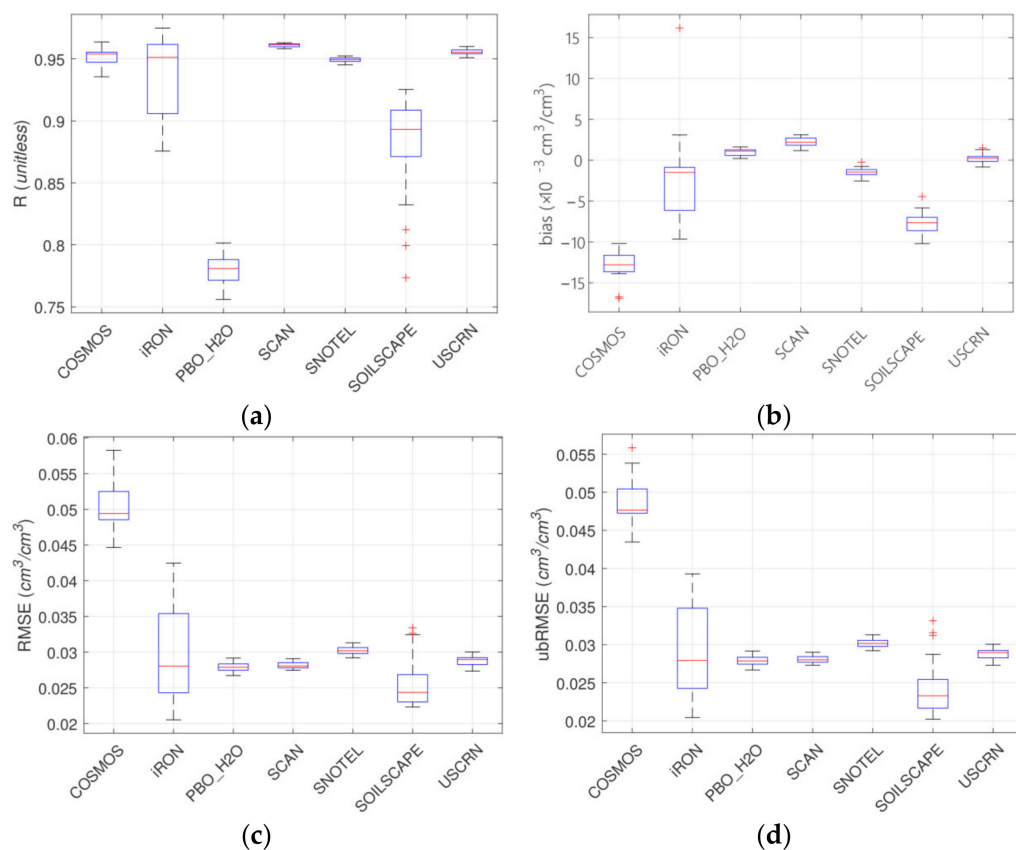


Figure 15. Box plots for (a) R, (b) bias, (c) RMSE, and (d) ubRMSE of each network.

Radar plot is usually used for comprehensive analysis of multiple indicators, which has the advantages of integrity, clarity and intuition. Figure 16 shows the performance of different models at seven networks. The results show that the performance of Multi-MDA-RF is better than the other models, which means that our model is more adaptable at different networks. In addition, BPNN and GRNN have low accuracy in predicting SM values at iRON, which may be because the two models are not dominant when there are few training samples. Figure 17 shows the performance of different SM products at seven networks. Multi-MDA-RF still performs best at all networks. The other products show low R values at COSMOS except for our product, which further indicates that Multi-MDA-RF has a wider range of applications. The RMSE values of AMSR2 are much higher than the other products at all networks, because there are many abnormal values in this product.

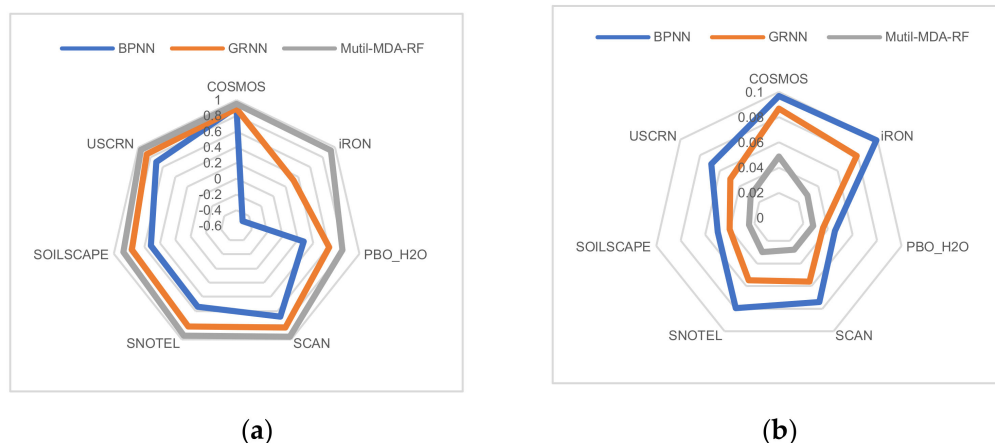
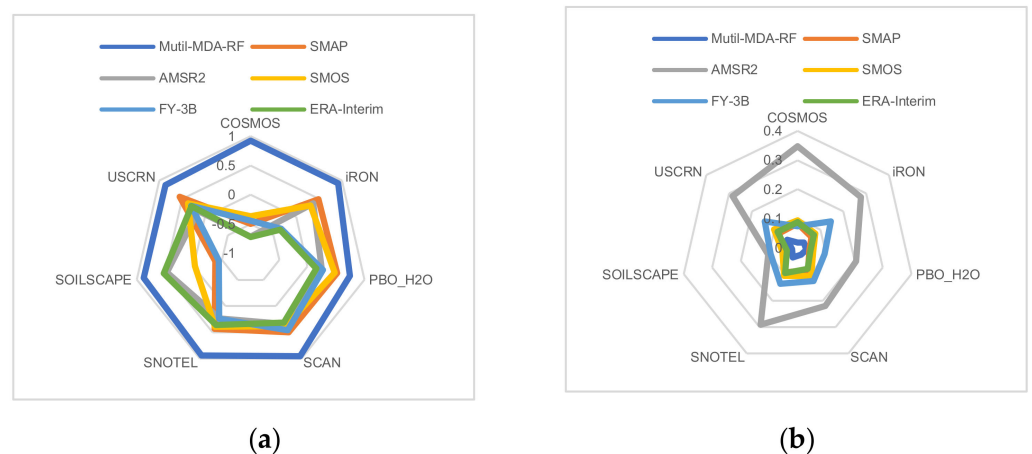


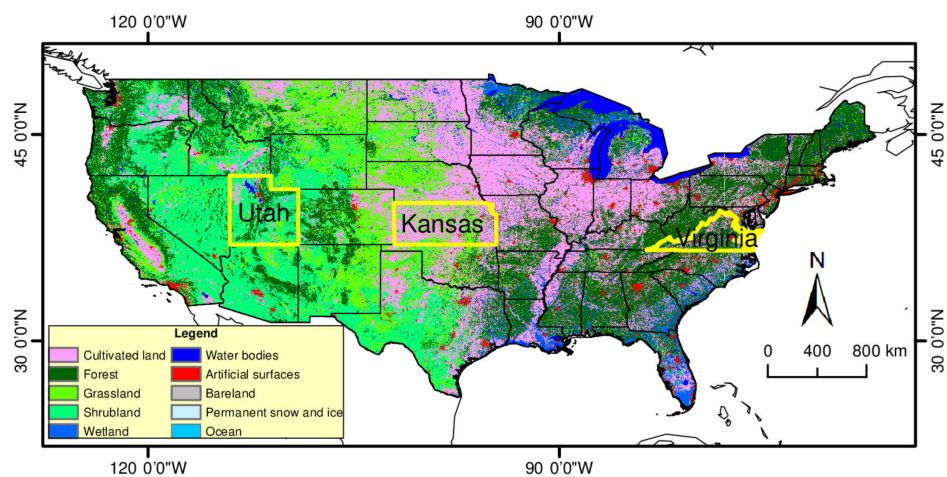
Figure 16. Radar plots for (a) R and (b) ubRMSE of three models at different networks.



**Figure 17.** Radar plots for (a) R and (b) ubRMSE of six SM products at different networks.

#### 4.8. Evaluation of Different U.S. States

To evaluate the spatial variations between different states, we compared the SM values from three U.S. states with similar latitudes, i.e., Utah in the west, Kansas in the central, and Virginia in the east of CONUS. As shown in Figure 18, the topographies of the three areas are different. Utah is mainly covered with forest and grassland with high altitude. Kansas is mainly cultivated land and grassland, and the terrain is gentle. Virginia is mainly covered with forests, hills and low mountains. Table 7 shows the evaluation results of three selected U.S. states. The R value of Utah is the highest, which may be due to the richer in-situ measurements in the western region. Kansas has the lowest RMSE value, this is because the terrain of the central region is flat, which is conducive to the construction of SM retrieval model. The performance of Virginia is not very good, because the eastern region has less in-situ measurements and the dense vegetation cover is unfavorable for SM retrieval.



**Figure 18.** The distribution and land cover of U.S. states, where Utah, Kansas, and Virginia are highlighted. The source of land cover is GlobeLand30.

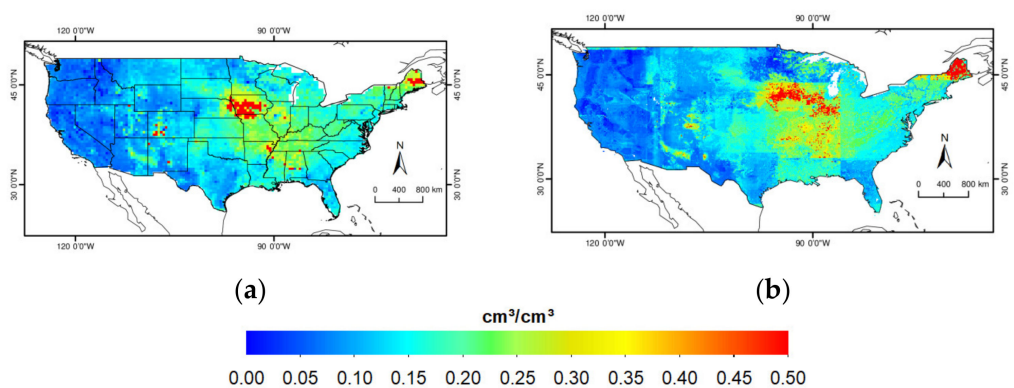
**Table 7.** Evaluation of different U.S. states.

Scheme 3.	Training Samples	R	Bias (cm <sup>3</sup> /cm <sup>3</sup> )	RMSE (cm <sup>3</sup> /cm <sup>3</sup> )	ubRMSE (cm <sup>3</sup> /cm <sup>3</sup> )
Utah	1032	0.94	0.000	0.027	0.027
Kansas	112	0.88	0.000	0.026	0.026
Virginia	64	0.87	0.002	0.031	0.030

#### 4.9. Producing High Resolution SM Map

Although our product has potential for SM retrieval, a SM product with relatively low spatial resolution is inadequate to be applied to practical use in some fields. Therefore, in this experiment, we used higher spatial resolution features as the input data to retrain the proposed Multi-MDA-RF model, and obtain a higher spatial resolution SM product.

In previous experiments, we resampled all of the features to the same spatial resolution of 36 km. After feature selection, the selected five features could be resampled according to the lowest resolution among the features, i.e., 0.125° of ERA-Interim product. The evaluation results of the retraining model are measured as  $R = 0.94$ , bias =  $0.000 \text{ cm}^3/\text{cm}^3$ , RMSE =  $0.033 \text{ cm}^3/\text{cm}^3$ , ubRMSE =  $0.033 \text{ cm}^3/\text{cm}^3$ , which means that the SM map with high spatial resolution still maintain high accuracy. Figure 19 shows a comparison of the lower and the higher resolution SM maps. It can be found that the SM map with higher spatial resolution can well represent the global and local variations with more clear details.



**Figure 19.** Comparison of SM maps with different resolutions; (a) 36 km and (b) 0.125°.

## 5. Discussion

The comparison of SM products shows that satellite products and LSM products have great uncertainty. Existing studies used these products as reference to train SM retrieval models, which will inevitably bring great uncertainty to the results [5,38,81]. The proposed model has greatly reduced the uncertainty from several aspects. Firstly, we used the in-situ measurements as the reference, which was beneficial to yielding more accurate results. Secondly, we conducted sensitive analysis of 29 features generated from multi-source data. Thirdly, we equipped the RF model with feature selection for SM retrieval, which had higher generalization performance with limited training samples. Previous comprehensive experiments have validated the ability of our model in terms of weakening uncertainty. For example, our model is more stable compared to BPNN and GRNN according to the box plots. Our product is closer to the in-situ measurements compared to other products according to the scatter plots, SM maps, and difference diagram. Our method equally performs best at different networks according to the radar plots.

To demonstrate the advantages of Multi-MDA-RF in terms of generalization performance, we compared several other studies in the same study area in Table 8. Firstly, it can be found that our method uses the most abundant features from multi-source products. However, the input data of other studies are all from a single source except auxiliary products. Most of the experiments have not conducted feature selection, while our experiment optimizes sensitive features. Although Senyurek et al. [45] also selected features, they considered fewer features, which may be not comprehensive enough. Secondly, we used fewer training samples (5225), which shows that our product also can be used in the areas with insufficient in-situ measurements. Thirdly, we achieve higher accuracies with the  $R$  value of 0.93 and the ubRMSE value of  $0.032 \text{ cm}^3/\text{cm}^3$ , indicating that our product has more potential to predict SM values. Fourthly, we produced two SM maps with different spatial resolutions and provide the highest spatial resolution in the analogous studies. The 0.125° map also has high accuracy and more clear details.





shows the highest accuracy when  $K = 5$ . Then, RF was employed to establish a nonlinear relationship between the optimal feature subset and the in-situ measurements with fewer training samples (i.e., a total of 5225). Compared with BPNN and GRNN, our model is more stable and achieves higher accuracy with the R value of 0.93 and ubRMSE value of  $0.032 \text{ cm}^3/\text{cm}^3$ . Compared with other SM products, our product is more consistent with the in-situ measurements, and it shows the best predictive ability and can well capture SM spatial dynamics. The evaluation of different SM networks indicates that SCAN and USCRN are ideal SM observation networks, and our method is more adaptable at different networks compared with other models and products. The evaluation of different U.S. states shows that the flat area with rich in-situ measurements is more suitable for the implementation of SM retrieval work. We also produced a SM product with higher spatial resolution of  $0.125^\circ$ , which can well represent the global and local variations with more clear details.

To conclude, the Multi-MDA-RF method used in this study shows great potential in estimating reliable regional SM values. The ideology of our work may be extended to SM retrievals from other meaningful SM features and applied in other geographical regions in the world. Future work will focus on more comprehensive SM feature selection, quality control of the in-situ data, and more intensive monthly data. We can also consider the incorporation of some state-of-the-art deep learning techniques to replace RF.

**Author Contributions:** L.Z.: visualization, writing—original draft. Z.Z.: data curation, investigation, formal analysis. Z.X.: conceptualization, methodology. H.L.: supervision, funding acquisition, resources. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China (41971279) and the Fundamental Research Funds for the Central Universities (B200202012).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data sharing not applicable.

**Acknowledgments:** The authors would like to thank the editor and the two anonymous reviewers who greatly helped us to improve the quality and presentation of the paper.

**Conflicts of Interest:** The authors declare that there is no conflict of interest regarding the publication of this article.

## References

1. Peng, J.; Loew, A.; Merlin, O.; Verhoest, N.E.C. A review of spatial downscaling of satellite remotely sensed soil moisture. *Rev. Geophys.* **2017**, *55*, 341–366. [[CrossRef](#)]
2. Kovačević, J.; Cvijetinović, Ž.; Stančić, N.; Brodić, N.; Mihajlović, D. New Downscaling Approach Using ESA CCI SM Products for Obtaining High Resolution Surface Soil Moisture. *Remote Sens.* **2020**, *12*, 1119. [[CrossRef](#)]
3. Karthikeyan, L.; Pan, M.; Wanders, N.; Kumar, D.N.; Wood, E.F. Four decades of microwave satellite soil moisture observations: Part 1. A review of retrieval algorithms. *Adv. Water Resour.* **2017**, *109*, 106–120. [[CrossRef](#)]
4. Alemohammad, S.H.; Kolassa, J.; Prigent, C.; Aires, F.; Gentile, P. Global downscaling of remotely sensed soil moisture using neural networks. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 5341–5356. [[CrossRef](#)]
5. Ge, L.; Hang, R.; Liu, Y.; Liu, Q. Comparing the Performance of Neural Network and Deep Convolutional Neural Network in Estimating Soil Moisture from Satellite Observations. *Remote Sens.* **2018**, *10*, 1327. [[CrossRef](#)]
6. Mohanty, B.P.; Cosh, M.H.; Lakshmi, V.; Montzka, C. Soil Moisture Remote Sensing: State-of-the-Science. *Vadose Zone J.* **2017**, *16*, 1–9. [[CrossRef](#)]
7. Zeng, J.; Chen, K.-S.; Bi, H.; Chen, Q. A preliminary evaluation of the SMAP radiometer soil moisture product over United States and Europe using ground-based measurements. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4929–4940. [[CrossRef](#)]
8. Kim, S.; Liu, Y.Y.; Johnson, F.M.; Parinussa, R.M.; Sharma, A. A global comparison of alternate AMSR2 soil moisture products: Why do they differ? *Remote Sens. Environ.* **2015**, *161*, 43–62. [[CrossRef](#)]
9. Cui, Y.; Long, D.; Hong, Y.; Zeng, C.; Zhou, J.; Han, Z.; Liu, R.; Wan, W. Validation and reconstruction of FY-3B/MWRI soil moisture using an artificial neural network based on reconstructed MODIS optical products over the Tibetan Plateau. *J. Hydrol.* **2016**, *543*, 242–254. [[CrossRef](#)]

10. Feng, X.; Li, J.; Cheng, W.; Fu, B.; Wang, Y.; Lü, Y.; Shao, M. Evaluation of AMSR-E retrieval by detecting soil moisture decrease following massive dryland re-vegetation in the Loess Plateau, China. *Remote Sens. Environ.* **2017**, *196*, 253–264. [[CrossRef](#)]
11. Zhuo, L.; Han, D. Multi-source hydrological soil moisture state estimation using data fusion optimisation. *Hydrol. Earth Syst. Sci.* **2017**, *21*, 3267–3285. [[CrossRef](#)]
12. Lee, C.S.; Sohn, E.; Park, J.D.; Jang, J.-D. Estimation of soil moisture using deep learning based on satellite data: A case study of South Korea. *GIScience Remote Sens.* **2019**, *56*, 43–67. [[CrossRef](#)]
13. Toride, K.; Sawada, Y.; Aida, K.; Koike, T. Toward high-resolution soil moisture monitoring by combining active-passive microwave and optical vegetation remote sensing products with land surface model. *Sensors* **2019**, *19*, 3924. [[CrossRef](#)] [[PubMed](#)]
14. Wang, Q.; van der Velde, R.; Ferrazzoli, P.; Chen, X.; Bai, X.; Su, Z. Mapping soil moisture across the Tibetan Plateau plains using Aquarius active and passive L-band microwave observations. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *77*, 108–118. [[CrossRef](#)]
15. Sathyanadh, A.; Karipot, A.; Ranalkar, M.; Prabhakaran, T. Evaluation of soil moisture data products over Indian region and analysis of spatio-temporal characteristics with respect to monsoon rainfall. *J. Hydrol.* **2016**, *542*, 47–62. [[CrossRef](#)]
16. Chen, F.; Crow, W.T.; Bindlish, R.; Colliander, A.; Burgin, M.S.; Asanuma, J.; Aida, K. Global-scale evaluation of SMAP, SMOS and ASCAT soil moisture products using triple collocation. *Remote Sens. Environ.* **2018**, *214*, 1–13. [[CrossRef](#)] [[PubMed](#)]
17. Jing, W.; Song, J.; Zhao, X. Validation of ECMWF Multi-Layer Reanalysis Soil Moisture Based on the OzNet Hydrology Network. *Water* **2018**, *10*, 1123. [[CrossRef](#)]
18. Kolassa, J.; Gentile, P.; Prigent, C.; Aires, F.; Alemohammad, S. Soil moisture retrieval from AMSR-E and ASCAT microwave observation synergy. Part 2: Product evaluation. *Remote Sens. Environ.* **2017**, *195*, 202–217. [[CrossRef](#)]
19. Kim, H.; Parinussa, R.; Konings, A.G.; Wagner, W.; Cosh, M.H.; Lakshmi, V.; Zohaib, M.; Choi, M. Global-scale assessment and combination of SMAP with ASCAT (active) and AMSR2 (passive) soil moisture products. *Remote Sens. Environ.* **2018**, *204*, 260–275. [[CrossRef](#)]
20. Bulut, B.; Yilmaz, M.T.; Afshar, M.H.; Şorman, A.Ü.; Yücel, I.; Cosh, M.H.; Şimşek, O. Evaluation of Remotely-Sensed and Model-Based Soil Moisture Products According to Different Soil Type, Vegetation Cover and Climate Regime Using Station-Based Observations over Turkey. *Remote Sens.* **2019**, *11*, 1875. [[CrossRef](#)]
21. Duygu, M.B.; Akyürek, Z. Using cosmic-ray neutron probes in validating satellite soil moisture products and land surface models. *Water* **2019**, *11*, 1362. [[CrossRef](#)]
22. Qu, Y.; Zhu, Z.; Chai, L.; Liu, S.; Montzka, C.; Liu, J.; Yang, X.; Lu, Z.; Jin, R.; Li, X.; et al. Rebuilding a Microwave Soil Moisture Product Using Random Forest Adopting AMSR-E/AMSR2 Brightness Temperature and SMAP over the Qinghai–Tibet Plateau, China. *Remote Sens.* **2019**, *11*, 683. [[CrossRef](#)]
23. Kerr, Y.H.; Al-Yaari, A.; Rodriguez-Fernandez, N.; Parrens, M.; Molero, B.; Leroux, D.; Bircher, S.; Mahmoodi, A.; Mialon, A.; Richaume, P.; et al. Overview of SMOS performance in terms of global soil moisture monitoring after six years in operation. *Remote Sens. Environ.* **2016**, *180*, 40–63. [[CrossRef](#)]
24. Agutu, N.O.; Awange, J.L.; Ndehedehe, C.; Mwaniki, M. Consistency of agricultural drought characterization over Upper Greater Horn of Africa (1982–2013): Topographical, gauge density, and model forcing influence. *Sci. Total Environ.* **2020**, *709*, 135149. [[CrossRef](#)]
25. Hagan, D.F.T.; Parinussa, R.M.; Wang, G.; Draper, C.S. An evaluation of soil moisture anomalies from global model-based datasets over the people’s republic of China. *Water* **2020**, *12*, 117. [[CrossRef](#)]
26. Sahebi, M.R.; Angles, J. An inversion method based on multi-angular approaches for estimating bare soil surface parameters from RADARSAT-1. *Hydrol. Earth Syst. Sci.* **2010**, *14*, 2355–2366. [[CrossRef](#)]
27. Baghdadi, N.; Chaaya, J.A.; Zribi, M. Semiempirical Calibration of the Integral Equation Model for SAR Data in C-Band and Cross Polarization Using Radar Images and Field Measurements. *IEEE Geosci. Remote Sens. Lett.* **2010**, *8*, 14–18. [[CrossRef](#)]
28. Kseneman, M.; Gleich, D.; Cucej, Z. Soil moisture estimation using high-resolution spotlight TerraSAR-X data. *IEEE Geosci. Remote Sens. Lett.* **2011**, *8*, 686–690. [[CrossRef](#)]
29. Lievens, H.; Verhoest, N.E.C. On the retrieval of soil moisture in wheat fields from L-Band SAR based on water cloud modeling, the IEM, and effective roughness parameters. *IEEE Geosci. Remote Sens. Lett.* **2011**, *8*, 740–744. [[CrossRef](#)]
30. Merzouki, A.; McNairn, H.; Pacheco, A. Mapping soil moisture using RADARSAT-2 data and local autocorrelation statistics. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2011**, *4*, 128–137. [[CrossRef](#)]
31. Wang, S.G.; Li, X.; Han, X.J.; Jin, R. Estimation of surface soil moisture and roughness from multi-angular ASAR imagery in the Watershed Allied Telemetry Experimental Research (WATER). *Hydrol. Earth Syst. Sci.* **2011**, *15*, 1415–1426. [[CrossRef](#)]
32. Shen, X.; Mao, K.; Qin, Q.; Hong, Y.; Zhang, G. Bare surface soil moisture estimation using Double-Angle and dual-polarization L-Band radar data. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 3931–3942. [[CrossRef](#)]
33. Chakravorty, A.; Chahar, B.R.; Sharma, O.P.; Dhanya, C.T. A regional scale performance evaluation of SMOS and ESA-CCI soil moisture products over India with simulated soil moisture from MERRA-Land. *Remote Sens. Environ.* **2016**, *186*, 514–527. [[CrossRef](#)]
34. Wigneron, J.P.; Jackson, T.J.; O’Neill, P.; De Lannoy, G.; de Rosnay, P.; Walker, J.P.; Ferrazzoli, P.; Mironov, V.; Bircher, S.; Grant, J.P.; et al. Modelling the passive microwave signature from land surfaces: A review of recent results and application to the L-band SMOS & SMAP soil moisture retrieval algorithms. *Remote Sens. Environ.* **2017**, *192*, 238–262.

35. Das, N.; Entekhabi, D.; Dunbar, R.; Chaubell, M.; Colliander, A.; Yueh, S.; Jagdhuber, T.; Chen, F.; Crow, W.; O'Neill, P.; 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](#)]
36. Karthikeyan, L.; Pan, M.; Konings, A.G.; Piles, M.; Fernandez-Moran, R.; Kumar, D.N.; Wood, E.F. Simultaneous retrieval of global scale Vegetation Optical Depth, surface roughness, and soil moisture using X-band AMSR-E observations. *Remote Sens. Environ.* **2019**, *234*, 111473. [[CrossRef](#)]
37. Azimi, S.; Dariane, A.B.; Modanesi, S.; Bauer-Marschallinger, B.; Bindlish, R.; Wagner, W.; Massari, C. Assimilation of Sentinel 1 and SMAP—Based satellite soil moisture retrievals into SWAT hydrological model: The impact of satellite revisit time and product spatial resolution on flood simulations in small basins. *J. Hydrol.* **2020**, *581*, 124367. [[CrossRef](#)] [[PubMed](#)]
38. Kolassa, J.; Reichle, R.H.; Liu, Q.; Alemohammad, S.H.; Gentine, P.; Aida, K.; Asanuma, J.; Bircher, S.; Caldwell, T.; Colliander, A.; et al. Estimating surface soil moisture from SMAP observations using a Neural Network technique. *Remote Sens. Environ.* **2018**, *204*, 43–59. [[CrossRef](#)]
39. Cui, Y.; Chen, X.; Xiong, W.; He, L.; Lv, F.; Fan, W.; Luo, Z.; Hong, Y. A Soil Moisture Spatial and Temporal Resolution Improving Algorithm Based on Multi-Source Remote Sensing Data and GRNN Model. *Remote Sens.* **2020**, *12*, 455. [[CrossRef](#)]
40. Xu, H.; Yuan, Q.; Li, T.; Shen, H.; Zhang, L.; Jiang, H. Quality Improvement of Satellite Soil Moisture Products by Fusing with In-Situ Measurements and GNSS-R Estimates in the Western Continental U.S. *Remote Sens.* **2018**, *10*, 1351. [[CrossRef](#)]
41. Eroglu, O.; Kurum, M.; Boyd, D.; Gurbuz, A.C. High Spatio-Temporal Resolution CYGNSS Soil Moisture Estimates Using Artificial Neural Networks. *Remote Sens.* **2019**, *11*, 2272. [[CrossRef](#)]
42. Ma, C.; Li, X.; Wei, L.; Wang, W. Multi-scale validation of SMAP soil moisture products over cold and arid regions in northwestern China using distributed ground observation data. *Remote Sens.* **2017**, *9*, 327. [[CrossRef](#)]
43. Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J.; et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sens. Environ.* **2020**, *241*, 111716. [[CrossRef](#)]
44. Yuan, Q.; Xu, H.; Li, T.; Shen, H.; Zhang, L. Estimating surface soil moisture from satellite observations using a generalized regression neural network trained on sparse ground-based measurements in the continental U.S. *J. Hydrol.* **2020**, *580*, 125843. [[CrossRef](#)]
45. Senyurek, V.; Lei, F.; Boyd, D.; Kurum, M.; Gurbuz, A.C.; Moorhead, R. Machine Learning-Based CYGNSS Soil Moisture Estimates over ISMN sites in CONUS. *Remote Sens.* **2020**, *12*, 1168. [[CrossRef](#)]
46. Fang, K.; Shen, C.; Kifer, D.; Yang, X. Prolongation of SMAP to Spatiotemporally Seamless Coverage of Continental U.S. Using a Deep Learning Neural Network. *Geophys. Res. Lett.* **2017**, *44*, 11030–11039. [[CrossRef](#)]
47. Zreda, M.; Desilets, D.; Ferre, T.P.A.; Scott, R.L. Measuring soil moisture content non-invasively at intermediate spatial scale using cosmic-ray neutrons. *Geophys. Res. Lett.* **2008**, *35*. [[CrossRef](#)]
48. Zreda, M.; Shuttleworth, W.J.; Zeng, X.; Zweck, C.; Desilets, D.; Franz, T.E.; Rosolem, R. COSMOS: The COsmic-ray Soil Moisture Observing System. *Hydrol. Earth Syst. Sci.* **2012**, *16*, 4079–4099. [[CrossRef](#)]
49. Montzka, C.; Bogena, H.R.; Zreda, M.; Monerris, A.; Morrison, R.; Muddu, S.; Vereecken, H. Validation of Spaceborne and Modelled Surface Soil Moisture Products with Cosmic-Ray Neutron Probes. *Remote Sens.* **2017**, *9*, 103. [[CrossRef](#)]
50. Osenga, E.C.; Arnott, J.C.; Endsley, K.A.; Katzenberger, J.W. Bioclimatic and Soil Moisture Monitoring Across Elevation in a Mountain Watershed: Opportunities for Research and Resource Management. *Water Resour. Res.* **2019**, *55*, 2493–2503. [[CrossRef](#)]
51. Larson, K.M.; Small, E.; Gutmann, E.; Bilich, A.L.; Braun, J.J.; Zavorotny, V.U. Use of GPS receivers as a soil moisture network for water cycle studies. *Geophys. Res. Lett.* **2008**, *35*. [[CrossRef](#)]
52. Schaefer, G.L.; Cosh, M.H.; Jackson, T.J. The USDA Natural Resources Conservation Service Soil Climate Analysis Network (SCAN). *J. Atmos. Ocean. Technol.* **2007**, *24*, 2073–2077. [[CrossRef](#)]
53. Lu, Y.; Wei, C. Evaluation of microwave soil moisture data for monitoring live fuel moisture content (LFMC) over the coterminous United States. *Sci. Total. Environ.* **2021**, *771*, 145410. [[CrossRef](#)]
54. Leavesley, G.; David, O.; Garen, D.; Lea, J.; Marron, J.; Pagano, T.; Perkins, T.; Strobel, M. A Modeling Framework for Improved Agricultural Water Supply Forecasting. AGU Fall Meeting, San Francisco, CA, USA, 15–19 December 2008; Volume 2008, p. C21A-0497.
55. Moghaddam, M.; Entekhabi, D.; Goykhman, Y.; Li, K.; Liu, M.Y.; Mahajan, A.; Nayyar, A.; Shuman, D.; Teneketzis, D. A Wireless Soil Moisture Smart Sensor Web Using Physics-Based Optimal Control: Concept and Initial Demonstrations. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2010**, *3*, 522–535. [[CrossRef](#)]
56. Bell, J.E.; Palecki, M.A.; Baker, C.B.; Collins, W.G.; Lawrimore, J.H.; Leeper, R.; Hall, M.E.; Kochendorfer, J.; Meyers, T.P.; Wilson, T.; et al. U.S. Climate Reference Network Soil Moisture and Temperature Observations. *J. Hydrometeorol.* **2013**, *14*, 977–988. [[CrossRef](#)]
57. Dorigo, W.; Wagner, W.; Hohensinn, R.; Hahn, S.; Paulik, C.; Xaver, A.; Gruber, A.; Drusch, M.; Mecklenburg, S.; Van Oevelen, P.; et al. The International Soil Moisture Network: A data hosting facility for global in situ soil moisture measurements. *Hydrol. Earth Syst. Sci.* **2011**, *15*, 1675–1698. [[CrossRef](#)]
58. Dorigo, W.A.; Xaver, A.; Vreugdenhil, M.; Gruber, A.; Hegyiová, A.; Sanchis-Dufau, A.D.; Zamojski, D.; Cordes, C.; Wagner, W.; Drusch, M. Global Automated Quality Control of In Situ Soil Moisture Data from the International Soil Moisture Network. *Vadose Zone J.* **2013**, *12*. [[CrossRef](#)]

59. Chen, Q.; Zeng, J.; Cui, C.; Li, Z.; Chen, K.-S.; Bai, X.; Xu, J. Soil Moisture Retrieval From SMAP: A Validation and Error Analysis Study Using Ground-Based Observations Over the Little Washita Watershed. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 1394–1408. [[CrossRef](#)]
60. Singh, G.; Das, N.N.; Panda, R.K.; Colliander, A.; Jackson, T.J.; Mohanty, B.P.; Entekhabi, D.; Yueh, S.H. Validation of SMAP Soil Moisture Products Using Ground-Based Observations for the Paddy Dominated Tropical Region of India. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8479–8491. [[CrossRef](#)]
61. Bindlish, R.; Caldwell, T.; Collins, C.H.; McNairn, H.; Martinez-Fernandez, J.; Prueger, J.; Rowlandson, T.; Seyfried, M.; Starks, P.; Thibeault, M.; et al. GCOM-W AMSR2 soil moisture product validation using core validation sites. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 209–219. [[CrossRef](#)]
62. Peng, J.; Misra, S.; Piepmeier, J.R.; Dinnat, E.P.; Hudson, D.; Le Vine, D.M.; De Amici, G.; Mohammed, P.N.; Bindlish, R.; Yueh, S.H.; et al. Soil moisture active/passive L-Band microwave radiometer postlaunch calibration. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5339–5354. [[CrossRef](#)]
63. Liu, J.; Chai, L.; Lu, Z.; Liu, S.; Qu, Y.; Geng, D.; Song, Y.; Guan, Y.; Guo, Z.; Wang, J.; et al. Evaluation of SMAP, SMOS-IC, FY3B, JAXA, and LPRM soil moisture products over the Qinghai-Tibet Plateau and its surrounding areas. *Remote Sens.* **2019**, *11*, 792. [[CrossRef](#)]
64. Zeng, J.; Li, Z.; Chen, Q.; Bi, H.; Qiu, J.; Zou, P. Evaluation of remotely sensed and reanalysis soil moisture products over the Tibetan Plateau using in-situ observations. *Remote Sens. Environ.* **2015**, *163*, 91–110. [[CrossRef](#)]
65. Guyon, I.; Elisseeff, A. An Introduction of Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
66. Kira, K.; Rendell, L. The Feature Selection Problem: Traditional Methods and a New Algorithm. In Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI'92), San Jose, CA, USA, 12–16 July 1992.
67. Robnik-Sikonja, M.; Kononenko, I. An adaptation of Relief for attribute estimation in regression. In Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), Nashville, TN, USA, 8–12 July 2000.
68. Robnik-Šikonja, M.; Kononenko, I. Theoretical and Empirical Analysis of Relief and RRelief. *Mach. Learn.* **2003**, *53*, 23–69. [[CrossRef](#)]
69. Dhanya, R.; Paul, I.R.; Akula, S.S.; Sivakumar, M.; Nair, J.J. F-test feature selection in Stacking ensemble model for breast cancer prediction. *Procedia Comput. Sci.* **2020**, *171*, 1561–1570. [[CrossRef](#)]
70. Yang, W.; Wang, K.; Zuo, W. Neighborhood Component Feature Selection for High-Dimensional Data. *JCP* **2012**, *7*, 161–168. [[CrossRef](#)]
71. He, X.; Cai, D.; Niyogi, P. Laplacian score for feature selection. In Proceedings of the 18th International Conference on Neural Information Processing Systems (NIPS'05), Vancouver, BC, Canada, 5–8 December 2005.
72. Petković, M.; Kocev, D.; Džeroski, S. Feature ranking for multi-target regression. *Mach. Learn.* **2019**, *109*, 1179–1204. [[CrossRef](#)]
73. Gwetu, M.V.; Tapamo, J.-R.; Viriri, S. Exploring the impact of purity gap gain on the efficiency and effectiveness of random forest feature selection. In Proceedings of the International Conference on Computational Collective Intelligence (ICCCI'19), Hendaye, France, 4–6 September 2019; pp. 340–352.
74. Behnamian, A.; Millard, K.; Banks, S.N.; White, L.; Richardson, M.; Pasher, J. A Systematic Approach for Variable Selection With Random Forests: Achieving Stable Variable Importance Values. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1988–1992. [[CrossRef](#)]
75. Rasmussen, C.; Williams, C. *Gaussian Process for Machine Learning*; The MIT Press: Cambridge, MA, USA, 2006.
76. Marcano-Cedeño, A.; Quintanilla, J.; Cortina-Januchs, G.; Andina, D. Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network. In Proceedings of the IECON 2010-36th Annual Conference on IEEE Industrial Electronics Society, Glendale, AZ, USA, 7–10 November 2010.
77. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
78. Entekhabi, D.; Reichle, R.H.; Koster, R.D.; Crow, W.T. Performance Metrics for Soil Moisture Retrievals and Application Requirements. *J. Hydrometeorol.* **2010**, *11*, 832–840. [[CrossRef](#)]
79. Ma, H.; Zeng, J.; Chen, N.; Zhang, X.; Cosh, M.H.; Wang, W. Satellite surface soil moisture from SMAP, SMOS, AMSR2 and ESA CCI: A comprehensive assessment using global ground-based observations. *Remote Sens. Environ.* **2019**, *231*. [[CrossRef](#)]
80. Ge, X.; Wang, J.; Ding, J.; Cao, X.; Zhang, Z.; Liu, J.; Li, X. Combining UAV-based hyperspectral imagery and machine learning algorithms for soil moisture content monitoring. *PeerJ* **2019**, *7*, e6926. [[CrossRef](#)] [[PubMed](#)]
81. Yang, X.; Zhang, C.; Cui, Z.; Yu, F.; Wang, J.; Han, Y. Filling method for soil moisture based on BP neural network. *J. Appl. Remote Sens.* **2018**, *12*, 042806.
82. Chatterjee, S.; Huang, J.; Hartemink, A.E. Establishing an Empirical Model for Surface Soil Moisture Retrieval at the U.S. Climate Reference Network Using Sentinel-1 Backscatter and Ancillary Data. *Remote Sens.* **2020**, *12*, 1242. [[CrossRef](#)]