Computers and Electronics in Agriculture 192 (2022) 106623

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Developing machine learning models with multisource inputs for improved land surface soil moisture in China

Lei Wang [a,b], Shibo Fang [a,*], Zhifang Pei [c], Dong Wu [d], Yongchao Zhu [e], Wen Zhuo [a]

[a] State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing 100081, PR China
[b] National Climate Center, China Meteorological Administration, Beijing 100081, PR China
[c] School of Architecture, Nanyang Institute of Technology, Nanyang 473004, PR China
[d] College of Resources and Environment, Anhui Agricultural University, Hefei 230036, PR China
[e] Meteorological Observation Center, China Meteorological Administration, Beijing 100081, PR China

ARTICLE INFO

Keywords:
Land surface soil moisture estimation
Machine learning
Satellite data
In situ data
Remote sensing

ABSTRACT

Accurate and spatially continuous land surface soil moisture (SSM) data will greatly benefit analyses of heat transfer, energy exchange and agricultural dryness. To obtain spatiotemporally consistent SSM information, five machine learning (ML) models, i.e., polynomial regression (PR), ridge regression (RR), lasso regression (LR), elastic net regression (EnR) and random forest regression (RfR) models, were generated to map the regional SSM in the 0–10 cm soil layer across the study area. Multiple features, including the geographical location, elevation, vegetation coverage, soil texture, seasonal patterns and satellite-retrieved SSM product from Fengyun-3C (FY-3C), were selected as the input variables for the proposed ML models. In situ SSM measurements from the Chinese Automatic Soil Moisture Observation Stations (CASMOS) were used as the reference dataset. The error metrics, including the coefficient of correlation (R), mean relative error (MRE), unbiased RMSE (ubRMSE) and mean absolute error (MAE), between the measured SSM values and those estimated using the different models were calculated. Among those ML models, the RfR model showed the best performance in the training (R = 0.981, MRE = 7.3%, ubRMSE = 0.021 cm^3/cm^3, and MAE = 0.015 cm^3/cm^3) and testing (R = 0.789, MRE = 22.2%, ubRMSE = 0.065 cm^3/cm^3, and MAE = 0.047 cm^3/cm^3) processes and was applied to map the regional SSM values and measure the importance of each input feature. The results indicated that geographical location, i.e., latitude (35.84%) and longitude (16.96%), contributed the most to the SSM estimation model, followed by elevation (14.88%), vegetation coverage (9.75%), the FY-3C SSM product (8.30%), the soil texture (8.04%) and seasonal patterns (6.23%). In addition, the SSM estimations across mainland China matched the spatiotemporal patterns of historical precipitation well, which indicated the feasibility of achieving accurate and consistent land surface (0–10 cm) soil moisture monitoring results using the established RfR model with appropriately selected input features.

1. Introduction

The occurrence of drought events, which are usually driven by factors such as low surface soil moisture (SSM) and high evapotranspiration (ET), has rapidly increased in the past several decades in China, causing substantial environmental and societal losses (Wu et al., 2019). In recent years, SSM has become widely considered a key parameter for monitoring agricultural, hydrological and meteorological droughts (Huang et al., 2020). In addition, SSM plays an important role in the development of weather and precipitation patterns and has been extensively adopted to analyze global energy, heat and water exchange processes (Crow et al., 2008). Generally, there are two methods to calculate the SSM content, i.e., the direct and indirect methods. Among them, the gravimetric method, which get the mass of water in the soil (known as the oven-dry weight) by measuring the difference between the moist soil and the soil dried at 105 °C, is frequently utilized to estimate soil moisture content as the direct method, as it provides more accurate soil moisture data compared to other device-based indirect

* Corresponding author at: State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Zhongguancun South Street No. 46, Haidian, Beijing 100081, PR China.
E-mail addresses: wangleic137955@cma.cn (L. Wang), fangshibo@cma.gov.cn (S. Fang), zhifangpei@stu.cdut.edu.cn (Z. Pei), dongwu@ahau.edu.cn (D. Wu), zhuyc10@cma.gov.cn (Y. Zhu), zhuowen1992@cau.edu.cn (W. Zhuo).

https://doi.org/10.1016/j.compag.2021.106623
Received 9 March 2021; Received in revised form 6 December 2021; Accepted 7 December 2021
Available online 17 December 2021

measurement (Sekertekin et al., 2020). Besides, there are several indirect methods, including the station-based measurement sensors, satellite-based remote sensing technology and models that assimilate multisource data (Jackson et al., 2008; Singh et al., 2018), for detecting the land SSM.

Continuous station-based in situ SSM measurements are crucial to analyze agricultural drought patterns and quantitatively validate the reliability of remotely sensed SSM products (Zhu et al., 2019). A large number of techniques, including gamma attenuation, frequency domain reflectometry (FDR), time domain reflectometry (TDR) and soil heat flux, have been employed to measure soil moisture content and profiles (Jackson et al., 2008). However, most of the current methods are expensive, cumbersome to perform and likely to be affected by environmental noise (Vreugdenhil et al., 2013). In the past two decades, substantial efforts have been devoted to designing and establishing regional and continental soil moisture networks to quantify the accuracy of various satellite retrieved and modeled SSM products (Bircher et al., 2012). These heavily instrumented soil moisture networks, such as the international soil moisture network (ISMN), soil climate analysis network (SCAN), FLUXNET and the national automatic soil moisture network of China, i.e., the Chinese automatic soil moisture observation stations (CASMOS), can acquire continuous and widespread moisture metrics and providing insight into the validation processes of satellite SSM retrievals. Among those networks, CASMOS, with more than 2000 observation stations densely distributed across all of China, was constructed to measure and record hourly soil moisture-related parameters at different soil layers (Zhu et al., 2019). Thus, CASMOS is suitable for drought monitoring and satellite data validation.

In recent decades, with the rapid development of remote sensing technology, a large number of spatiotemporally continuous SSM products from various satellite platforms have been employed to explore the role of land SSM in land-atmosphere interactions and to detect agricultural droughts (Srivastava, 2017). Microwave remote sensing, in both its active and passive forms, has been widely recognized as one of the most promising SSM monitoring approaches due to its high sensitivity to SSM content and ability to monitor SSM dynamics under all weather conditions (Sabaghy et al., 2018). Active microwave remote sensing, such as synthetic aperture radar (SAR), has the capability of observing land SSM over a wide range and at high spatial resolutions (Torres et al., 2012). Additionally, its revisit time has been significantly reduced from 35 days or longer to 6–12 days through the Global Monitoring for Environment and Security (GMES) Sentinel-1 constellation operated by the European Space Agency (ESA). This reduction greatly improves the suitability of SAR for hydrological applications (Kornelsen and Coulibaly, 2013). In general, passive microwave remote sensing has a larger number of developed algorithms and a higher temporal resolution but a lower spatial resolution than active microwave remote sensing (Wang et al., 2020). Currently, a wide variety of SSM products are readily available from multiple satellites, such as the soil moisture active passive (SMAP) mission, the advanced microwave scanning radiometer-2 (AMSR-2), the soil moisture and ocean salinity (SMOS) mission and the FengYun-3 (FY-3) series meteorological satellites. Those missions and satellites have been extensively applied to monitor regional SSM conditions.

Despite the advantages of microwave remote sensing in achieving regional SSM coverage, its accuracy is likely to be affected by factors such as the land surface roughness, vegetation biomass and vegetation water content (VWC) (Sabaghy et al., 2018). Many researchers have developed various techniques to enhance satellite SSM products to better meet the requirements of different applications in recent years (Choi and Hur, 2012; Shi et al., 2006). Among those techniques, machine learning (ML) models, such as back-propagation neural networks (BPNNs), the support vector machine (SVM) approach, the random forest method, the XGBoost method and general regression neural networks (GRNNs), which have high potential for feature selection and parameter optimization, have been extensively employed to comprehensively estimate and downscale the regional SSM (Wang et al., 2020; Zhang et al., 2020). Frequently selected features for these ML models include the precipitation, vegetation index (VI), land surface temperature (LST), ET, digital elevation model (DEM), land cover (LC), brightness temperature (BT), albedo, latitude, longitude, soil texture, seasonal difference and remotely sensed SSM data (Sabaghy et al., 2018; Wang et al., 2020; Zhang et al., 2020). Also, the soil moisture networks and in situ measurements have been played an important role as the reference datasets during the training and validation of the ML models for estimating land surface and root-zone soil moisture at the continental scale (Rodriguez-Fernandez et al., 2017). Hence, as one of the most promising techniques, ML models with input variables from multiple sources should be established and compared to obtain SSM estimations with high accuracy.

This paper aims to achieve accurate and spatiotemporally continuous land SSM coverage within the 0–10 cm soil layer by building and comparing models with five frequently used ML techniques: polynomial regression (PR), ridge regression (RR), lasso regression (LR), elastic net regression (EnR) and random forest regression (RfR). For this purpose, the CASMOS measurements from 2017 to 2019 were employed as the reference dataset. Twenty SSM-related features, including the geographical location, vegetation information, remotely sensed SSM images, soil texture information and seasonal pattern, were generated from multiple data sources to train and validate those ML models. The best model (estimator) was selected by calculating the error parameters between the in situ SSM measurements and the SSM estimations obtained by different models. Then, the best model was applied to measure the relative importance of each input feature and to map the regional SSM across the study area.

2. Study area and datasets

2.1. Study area

China is located in the eastern Asia on the west coast of the Pacific Ocean with an approximately 9.6 million km^2 land area, and its continental coastline is more than 18,000 km. This region is under the impact of the monsoon climate, and the precipitation patterns varies significantly from the northwestern to southeastern parts during different seasons. In the past several decades, China experiences a rapid increase in water use as for the fast development of industry and agricultural irrigation. The uneven precipitation and increase of the water use have led to a higher frequency of droughts and floods occurrence, especially in the northern China (Zhao et al., 2017). Therefore, it is urgent and of great importance to achieve accurate and continuous soil moisture dataset at the regional scale for China's drought monitoring and flood detection.

2.2. Datasets

2.2.1. In situ SSM measurements from CASMOS

CASMOS was constructed and developed by the Chinese Meteorological Administration (CMA) and has been widely used as an extensive national soil moisture network to monitor agricultural droughts, provide early warnings of meteorological disasters, and validate the accuracy of remotely sensed SSM products (Zhu et al., 2019). The number of SSM data collection stations has gradually increased to 2075 in the past decade. These stations are densely and evenly distributed across most parts of China, especially in the main agricultural regions, such as the North China Plain at elevations lower than 100 m (Fig. 1). Four soil moisture parameters, including the soil volumetric water content (SVWC), soil weight water content (SWWC), relative soil humidity (RSH), and soil available water storage (SAWS), at eight soil depth levels, i.e., 0–10 cm, 10–20 cm, 20–30 cm, 30–40 cm, 40–50 cm, 50–60 cm, 70–80 cm, and 90–100 cm, are recorded per hour. In this study, monthly in situ SVWC measurements at the 0–10 cm depth were

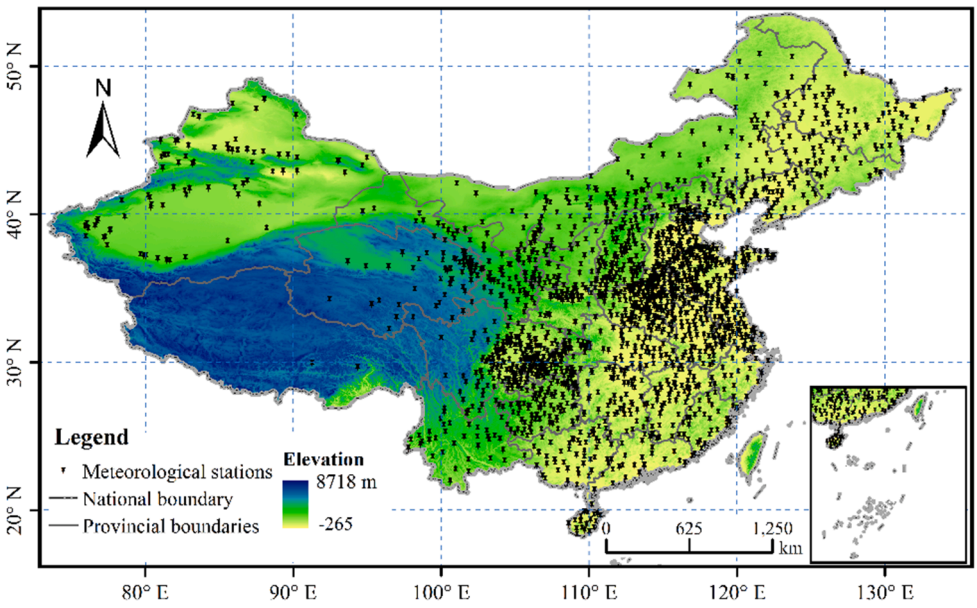

Fig. 1. The spatial distribution of CASMOS stations in the study area.

produced from the hourly SSM values using the averaging method. Then, the mean value of those in situ soil moisture measurements in each FY-3C footprint was calculated and adopted as the target datasets for training and validating the SSM estimation models.

2.2.2. Remotely sensed SSM data from FY-3C satellite

The Fengyun-3C (FY-3C) satellite operated by the CMA has been used extensively to obtain global meteorological time series data. Its capability for quantitatively detecting the atmosphere, land surface and sea surface under all weather conditions has been greatly improved compared with those of the FY-1 series satellites, which are China's first generation of polar-orbiting meteorological satellites. The FY-3C satellite is equipped with various instruments, including the microwave radiometer imager (MWRI), for collecting BT data and generating products, such as global SSM monitoring images at different time intervals and a spatial resolution of 25 km (Wang et al., 2020). The FY-3C SSM products are generated and optimized using the Q_p model, which was built using advanced integral equation (AIE) model simulations of microwave emissions (Shi et al., 2006). Compared with the single channel SSM retrieval algorithm, both the vertical and horizontal polarizations of the X-band (10.65 GHz) were applied in the Q_p model to decrease the effects of roughness and vegetation coverage. The FY-3C SSM data since May 2014 are freely available on the website of China's National Satellite Meteorological Center (NSMC) (satellite.nsmc.org.cn/). Here, the monthly FY-3C/MWRI SSM products from January 2017 to December 2019 were resized and reprojected to reflect the spatiotemporal patterns of the land surface (0–5 cm) soil moisture across the study area.

2.2.3. NDVI from Aqua-MODIS

The remotely sensed normalized difference vegetation index (NDVI) is the most well-developed and widely used approach for monitoring vegetation conditions and enhancing the spatial consistency of SSM retrievals using microwave remote sensing at the regional scale (Choi and Hur, 2012). In this study, the NDVI was selected as the input variable to correct the SSM overestimation issue caused by the vegetation coverage when the single FY-3C satellite data was used. For this purpose, monthly NDVI products (MYD13C2, collection v006) with 0.1° resolution derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) during 2017–2019 were downloaded from the website of the National Aeronautics and Space Administration (NASA) (https://ladsweb.modaps.eosdis.nasa.gov/). The spatial resolution of MODIS NDVI products from the Aqua (afternoon) orbit were resampled from 0.1° to 25 km using the nearest neighbor method to be consistent with the FY-3C SSM data. Also, the Aqua-MODIS NDVI products were reprojected from the sinusoidal to WGS-84 and the images covering China were extracted to reflect the vegetation information and improve the SSM estimation accuracy in this study.

2.2.4. Elevation data from the GMTED2010

The global multiresolution terrain elevation data for 2010 (GMTED2010) dataset, with spatial resolutions of 225 m, 450 m, and 1 km, was jointly released by the U.S. National Geospatial-Intelligence Agency (NGA) and the U.S. Geological Survey (USGS). The GMTED2010 dataset has been extensively applied to analyze the regional terrain characteristics and has proven to be more consistent and vertically accurate than the previous GTOPO30 elevation model (Danielson and Gesch, 2011). Additionally, the GMTED2010 dataset is freely available and is convenient to load onto geographic information processing platforms, such as the Environment for Visualizing Images (ENVI). In this study, the GMTED2010 dataset with a spatial resolution of 1 km was resampled to a 25 km spatial resolution using the nearest neighbor method. Then, the resampled elevation data were applied to map the spatial characteristics of the topography across the study area.

2.2.5. Soil texture data from the HWSD

Harmonized World Soil Database (HWSD) was jointly conducted by multiple institutions, including the Food and Agriculture Organization of the United Nations (FAO), the International Institute for Applied Systems Analysis (IIASA), the ISRIC-World Soil Information, the Institute of Soil Science-Chinese Academy of Sciences (ISSCAS) and the Joint Research Centre of the European Commission (JRC) (http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/). The HWSD includes the global soil map at the 1:5,000,000 scale released by the FAO/Unesco, as well as the recent updates of soil information at the regional and national scales. The HWSD datasets have been implemented to analyze the runoff and the agricultural modeling applications (Jones and Thornton, 2015). In this study, the soil map of China at a scale of 1:1 million distributed by the Institute of Soil Science in Nanjing was resampled to 25 km and generated as the input variables for soil moisture estimation.

2.2.6. Precipitation data from GPCC

The spatiotemporal patterns of the SSM are strongly determined by precipitation events. Thus, precipitation products are frequently applied to validate SSM consistency using satellite data (Wagner et al., 2003; Zhu et al., 2019). The Global Precipitation Climatology Centre (GPCC) provides long-term precipitation datasets with 1.0°, 2.5°, 0.25°, and 0.5° spatial resolutions on its website (https://opendata.dwd.de/climate_environment/GPCC/html/download_gate.html). In this study, the average monthly precipitation data during 1891–2016 at 0.25° were generated using the full GPCC dataset (version 2018), which was created by Schneider et al. in 2018 (see reference (Schneider et al., 2018)). Then, the historical precipitation was mapped to analyze the consistency between the estimated SSM and the precipitation events and to explain the input features (mainly the latitude and longitude) that made great contributions to the establishment of the SSM estimation models.

3. Methodology

3.1. Machine learning models

In this study, five ML models, namely the polynomial regression (PR), ridge regression (RR), lasso regression (LR), elastic net regression (EnR) and random forest regression (RfR), were regarded for SSM estimation. Those ML models present information about the relationship between variables such as remotely sensed SSM from FY-3C satellite, vegetation coverage, soil texture, geographical location and seasonal variation pattern, and the ground SSM measurements. In those models, the aforementioned variables were regarded as a function of measured SSM as shown in Eq. (1).

$$SSM_{sta} = f(SSM_{FY}, VI, Lat, Lon, Elev, ST, SP) \tag{1}$$

where SSM_{sta} is the dependent variable and represents for the SSM measurements from observation stations. SSM_{FY} is the SSM retrievals from the FY-3C satellite. VI, Lat, Lon, Elev, ST, and SP represent for vegetation information, Latitude, Longitude, Elevation, soil texture and seasonal pattern respectively.

In general, those established models with multiple input variables are likely to be highly complex and have classic overfitting issues. Therefore, the regularization technique is usually adopted to constrain the parameters, reduce the overfitting of the model by adding a regularization term to the loss function. A regularization term is a monotonic increasing function of the model's complexity, which indicates a larger regularization value with the increasing of the complexity.

The cross-validation is frequently used to select the model with the best performance. There are three steps: 1) randomly divide the given data into k subsets of the same size, and then 2) the model is trained using the k-1 subsets and tested using the remaining subset, 3) repeat the former step for k times and calculate the mean of the k tests. Therefore, the regularization technique and the cross-validation method were applied in this study to reduce the complexity, optimize the parameters and comprehensively assess the performance of PR models and the other ML models in the training and testing datasets.

3.1.1. Polynomial regression model

Standard and modified polynomial regression (PR) models are commonly used in modeling and predicting nonlinear relationships and functions between a dependent numeric variable and the values of one or several independent variables (Li et al., 2019). Examples of PR equations with one (x) and two (x_1, x_2) independent variables are given as follows:

$$\hat{y} = w_0 + w_1 x + w_2 x^2 + \ldots + w_m x^m \tag{2}$$

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 + w_5 x_1 x_2 + \ldots \tag{3}$$

where w_j ($j = 0, 1, \ldots, m$) are regression coefficients and m is the degree of the proposed PR model based on Eq. (2).

The performance of a PR model is commonly evaluated using the mean squared error (MSE), which is defined as:

$$E(x_n, w) = \sum_{n=1}^{N} |\hat{y}(x_n, w) - t_n|^2 \tag{4}$$

where n is the number of input variables and $\hat{y}(x_n, w)$ and t_n are the predicted and target (actual) values of the dependent variable, respectively. Generally, MSE values become increasingly close to zero with increasing model degree (m). However, a PR model with high m values is likely to be affected by the overfitting issue, and the regularization and cross-validation technique should be applied to optimize the PR model.

3.1.2. Ridge regression model

Ridge regression (RR) is an essential concept in data science and has been widely used to address multicollinearity and instability problems (Zou, 2020). A ridge penalty on the regression coefficients is applied to penalize the least squares loss in the RR model as follows:

$$E(\beta) = \sum (y - X\beta)^2 + \lambda \|\beta\|_2^2 \tag{5}$$

where β is the regression coefficient, λ ($\lambda \geq 0$) is a constant, and y and X are the dependent and independent variables of the RR model, respectively.

Generally, the key to conducting ridge regression is the selection of an appropriate λ value to balance the variance and bias of the RR model. Two commonly used methods of optimizing the λ parameter are the ridge trace method and the cross-validation method. Here, the dataset was divided into 10 groups for scoring and selecting the best λ value and the best estimator using the 'RidgeCV' import from the Python library 'Scikit-learn' (Pedregosa et al., 2011). Similarly, the input dataset was split into 10 subsets during the cross-validation processes using the other ML models in this study.

3.1.3. Lasso regression model

The least absolute shrinkage and selection operator (Lasso) was introduced by Robert Tibshirani in 1998 (Tibshirani, 1998) and has been extensively used for simulating parameters and selecting variables in survival analysis. Similar to ridge regression (Eq. (5)), a Lasso penalty, which applies the L1 norm instead of the L2 norm, is considered in the loss equation of the Lasso regression (LR) model as follows:

$$E(\beta) = \sum (y - X\beta)^2 + \lambda \|\beta\|_1 \tag{6}$$

Generally, the Lasso regularization used in the LR model makes it more appropriate for feature selection than the RR model and provides regression results that are easier to interpret. In this study, the values of the parameter λ were compared and selected using the 'LassoCV' module from 'Scikit-learn' (Pedregosa et al., 2011).

3.1.4. Elastic net regression model

The elastic net regression (EnR) model has attracted wide attention in fields such as statistics and machine learning due to its ability to achieve good performance under weak regularization (Wang et al., 2019). Similar to the LR model, the EnR model enables the assignment of unimportant parameters to values of zero and thus is suitable for extracting useful information from large datasets. The coefficient β is estimated by minimizing the loss function of the EnR model, which combines the penalty terms of the ridge and Lasso regression models. The coefficient β is defined as:

$$E(\beta) = \sum (y - X\beta)^2 + \lambda \sum (1 - \alpha) \|\beta\|_2^2 + \alpha \|\beta\|_1 \tag{7}$$

where α ($0 < \alpha < 1$) is the tuning parameter; the method applies the ridge or Lasso regression model when the α value equals 0 or 1, respectively. Here, the λ values were determined using the cross-

validation method with 10 subsets in Python, i.e., 'ElasticNetCV', for each varied value of α from 0.01 to 0.99 with 0.01 intervals.

3.1.5. Random forest regression model

The random forest concept introduced by Breiman in 2001 (Breiman, 2001), is an extension of classification and regression trees (CART), and it has been well developed and extensively adopted as an ensemble learning method to address categorical, continuous and time-to-event outcomes. Compared with other ML algorithms, the advantages of the random forest method are its high computational speed, the availability of feature importance information and its convenient procedures for feature selection (Ziegler and Koenig, 2014). In this study, a random forest regression (RfR) model was developed to predict continuous outputs.

In general, the steps for building and training an RfR model are 1) randomly select a certain number of samples as a subset from the training dataset; 2) construct a regression tree (RT) for each subset, define the important parameters and determine the best estimator by minimizing the MSE as follows:

$$min[\sum_{x_i \in D_1(A,s)} (y_i - \bar{y}_1)^2 + \sum_{x_i \in D_2(A,s)} (y_i - \bar{y}_2)^2] \quad (8)$$

where A is a randomly selected feature, s is the division point, and D_1 and D_2 are the datasets partitioned by s. x_i and y_i are the input and predicted values for the RfR model, respectively, and $\bar{y}_1$ and $\bar{y}_2$ are the averages of the target values in D_1 and D_2, respectively; 3) calculate the predicted value using each RT; and 4) obtain the final predicted value by averaging the predictions in step 3) using the following equation:

$$P(x) = \frac{\sum_{j=1}^{N} P_j(x)}{N} \quad (9)$$

where P is the final prediction and N is the number of RTs.

3.2. Performance criteria metrics

To quantify the accuracies of FY-3C SSM products and the estimated SSM using multivariate machine learning models, four statistical indicators, including the correlation coefficient (R), mean absolute error (MAE), unbiased root mean square error (ubRMSE) and mean relative error (MRE), were adopted in this study. The R, MAE and MRE were calculated by Eqs. (10), (11) and (12) respectively as follows.

$$R = \frac{\sum_{i=1}^{N}(SSM_i - \overline{SSM})(MSM_i - \overline{MSM})}{\sqrt{\sum_{i-1}^{N}(SSM_i - \overline{SSM})^2}\sqrt{\sum_{i-1}^{N}(MSM_i - \overline{MSM})^2}} \quad (10)$$

$$MAE = \frac{1}{N}\sum_{i-1}^{N}(SSM_i - MSM_i) \quad (11)$$

$$MRE = \frac{1}{N}\sum_{i-1}^{N}\frac{|SSM_i - MSM_i|}{SSM_i} \times 100\% \quad (12)$$

where SSM_i represents the SSM (cm^3/cm^3) values retrieved from FY-3C or the estimated SSM values, MSM_i represents the measured soil moisture (cm^3/cm^3), $\overline{SSM}$ represents the average of the FY-3C SSM retrievals or the SSM estimates using those ML models of all pixels in the study area, $\overline{MSM}$ represents the average of in situ soil moisture measurements, and N represents the total number of valid samples in each month during 2017–2019. The R values range between [−1, 1], and a large absolute value of R indicates a strong correlation between the SSM product and the soil moisture measurements. A positive MAE indicates an overestimation, and a negative MAE indicates an underestimation in those SSM product.

The ubRMSE is widely used to evaluate and compare the performance of different remotely sensed SSM products. In this study, the ubRMSE was adopted to reflect the absolute difference between the station-based measurements and the SSM product from FY-3C or SSM values estimated by those ML models. The ubRMSE were calculated with Eqs. (13)–(15) as follows.

$$ubRMSE = \sqrt{RMSE^2 - MD^2} \quad (13)$$

where RMSE can be calculated by Eq. (14), and MD is the abbreviation of the mean deviation, which can be calculated by Eq. (15).

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

$$MD = \frac{1}{N}\sum_{i=1}^{N}(SSM_i - MSM_i) \quad (15)$$

4. Results and discussion

4.1. Consistency assessment between FY-3C retrieved and ground measured SSM

To assess the consistency between the FY-3C SSM retrievals and the in situ SSM measurements in different months, the ubRMSE and MAE in the months from January to December 2017 were calculated (Fig. 2). Fig. 2 indicates that the ubRMSEs between the satellite-retrieved and in situ measured SSM in different months ranged between [0.10, 0.13] (cm^3/cm^3), and the errors in June, July, August, September and December were higher than those in the other months. The ubRMSEs tended to gradually decrease in spring, increase to above 0.12 cm^3/cm^3 from June to September, and then slightly decrease in October and November. The MAEs in most of the months were negative, especially in months from January to March, which were lower than −0.10 cm^3/cm^3. These results demonstrated a significant underestimation of SSM in the FY-3C satellite data in those months. However, the MAEs in July, August and September were positive, which indicated that the retrieved SSM values were generally higher than the in situ measured SSM values, especially in August.

According to Fig. 2, the SSM values from the FY-3C satellite data were obviously overestimated in August and underestimated in February, respectively. Here, the difference values between the FY-3C and the in situ SSM values were mapped as shown in Fig. 3. Fig. 3(a) indicated that the FY-3C retrieved SSM values were much lower than the soil moisture measurements at most of the meteorological stations across the study area in February 2017, especially in the southern central part of the study region, where the differences between the satellite retrieved and ground measured SSMs were lower than −0.20 cm^3/cm^3. The stations with overestimated SSM values were distributed sporadically in the southwestern study area. However, the SSM values were most overestimated using the FY-3C satellite data in August. This was especially true in areas with high vegetation coverage, such as the North China Plain and Northeast China, where the intensively planted corn was at its heading-milk stage in August (Fig. 3(b)). Therefore, the SSM estimation accuracy using the FY-3C satellite data was closely related to the vegetation coverage, the seasonal characteristics, as well as the geographical location.

4.2. Establishment of machine learning models for land SSM estimation

In this study, twenty features, including the FY-3C SSM, MODIS NDVI, months from January to December, latitude, longitude, elevation and soil separates including the clay, sand and silt, were selected as the input variables for five ML models (PR, RR, LR, EnR and RfR), to achieve accurate land surface soil moisture at the regional scale. The dataset

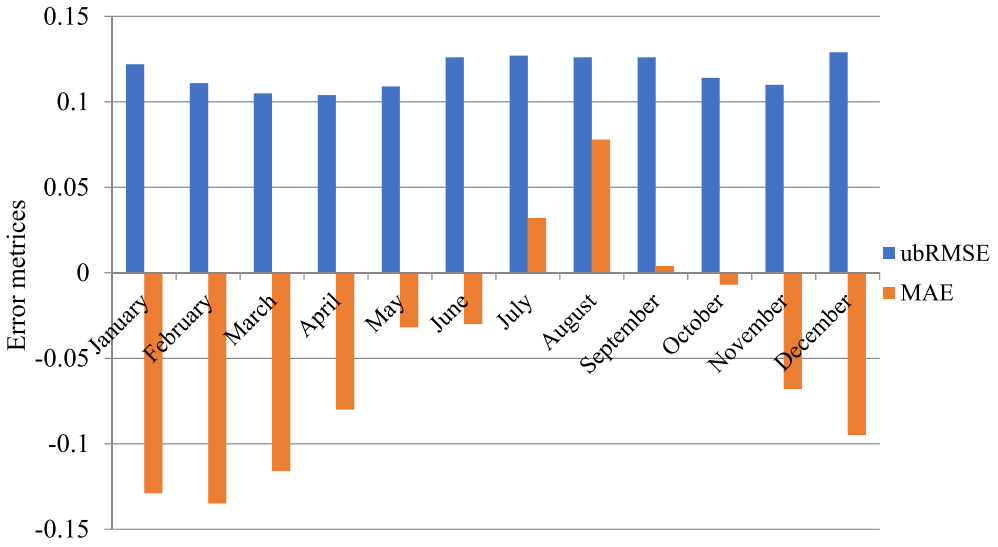

Fig. 2. The ubRMSE and MAE between the measured SSM and FY-3C SSM in different months.

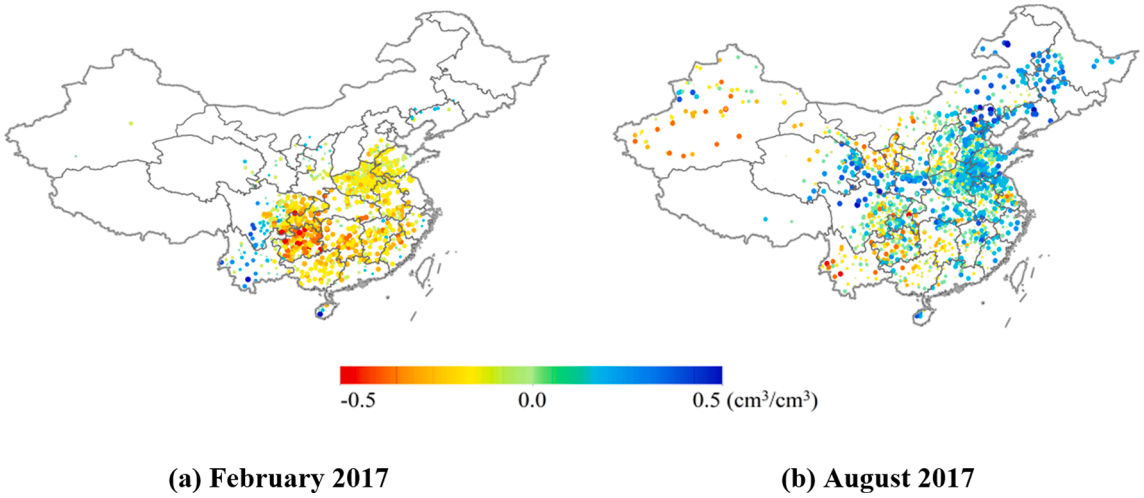

Fig. 3. Maps of the difference between the measured and the FY-3C retrieved SSM in February and August 2017.

during the years 2017–2019 was divided into a training dataset (2017–2018, 68.3%) and a testing dataset (2019, 31.7%). The training and testing error metrics (R, ubRMSE, MRE and MAE) of those models were calculated pixel-by-pixel and shown as Table 1. According to Table 1, among all of the models, the RfR model showed the best performance in the training (R = 0.981, MRE = 7.3%, ubRMSE = 0.021 cm^3/cm^3, and MAE = 0.015 cm^3/cm^3) and testing (R = 0.789, MRE = 22.2%, ubRMSE = 0.065 cm^3/cm^3, and MAE = 0.047 cm^3/cm^3) processes, followed by the RR and PR models, whose R values were approximately 0.61 and 0.57 in the training and testing processes, respectively. The performance of the EnR and LR models was relatively poor in this study, with training and testing R values of approximately or lower than 0.50.

To further evaluate the performance of the RfR model at the regional scale, the SSM data from the ground observations, the FY-3C retrievals, as well as the SSM estimations using the RfR and RR models with a spatial resolution of 25 km in January 2017 were mapped and shown in Fig. 4. It is obvious that the measured SSM values around the purple circle were above 0.25 cm^3/cm^3 (Fig. 4(a)). However, those values were greatly underestimated by the FY-3C satellite data in this region (Fig. 4 (b)). According to Fig. 4(c) and Fig. 4(d), the estimations using RR and RfR models were more consistent with the in situ SSM measurements than the pure satellite data. Moreover, the RfR model was able to present more spatial details in its monitoring results compared with the RR model. According to Fig. 4(a) and Fig. 4(b), the pixel values by the FY-3C satellite were apparently higher than their measurements in the orange

Table 1
Performance of five machine learning models.

Model Parameter	PR		RR		LR		EnR		RfR	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
R	0.608	0.572	0.606	0.571	0.485	0.478	0.520	0.509	0.981	0.789
MRE (%)	27.7	29.9	27.7	29.9	30.8	32.5	29.8	31.4	7.3	22.2
ubRMSE (cm^3/cm^3)	0.080	0.087	0.081	0.086	0.090	0.094	0.087	0.092	0.021	0.065
MAE (cm^3/cm^3)	0.063	0.068	0.063	0.068	0.072	0.075	0.069	0.073	0.014	0.047

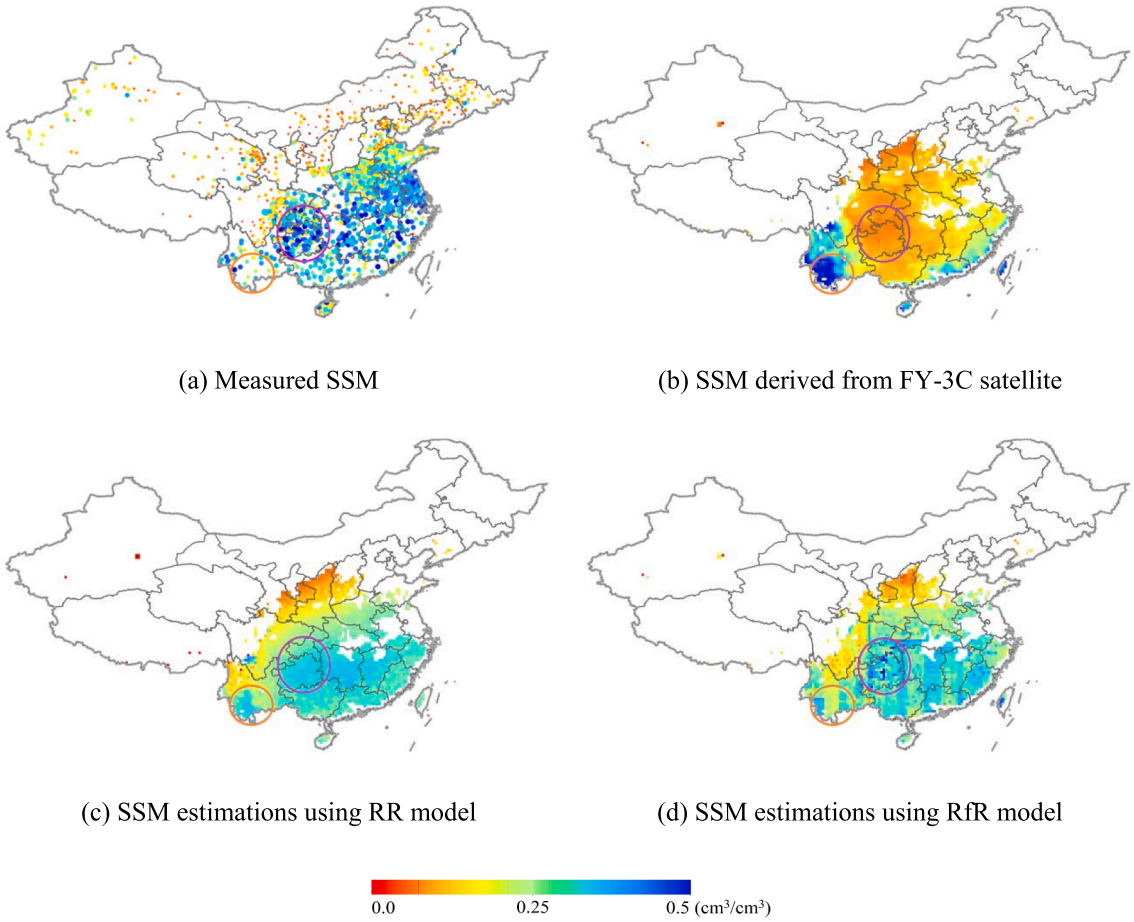

Fig. 4. Comparison between the measured, FY-3C retrieved and estimated SSM using the RR and RfR models in January 2017.

circles, and the overestimation issue was significantly improved in the SSM images estimated using the RR and RfR models (Fig. 4(c), Fig. 4(d)). To conclude, the SSM products estimated by the RR and RfR models were more consistent with the CASMOS measurements than the FY-3C satellite retrievals. These results indicated the feasibility of applying the proposed ML models with multiple and appropriate input features. In this study, the RfR model, which achieved the highest accuracy among the five models (Table 1), was used to estimate the regional land

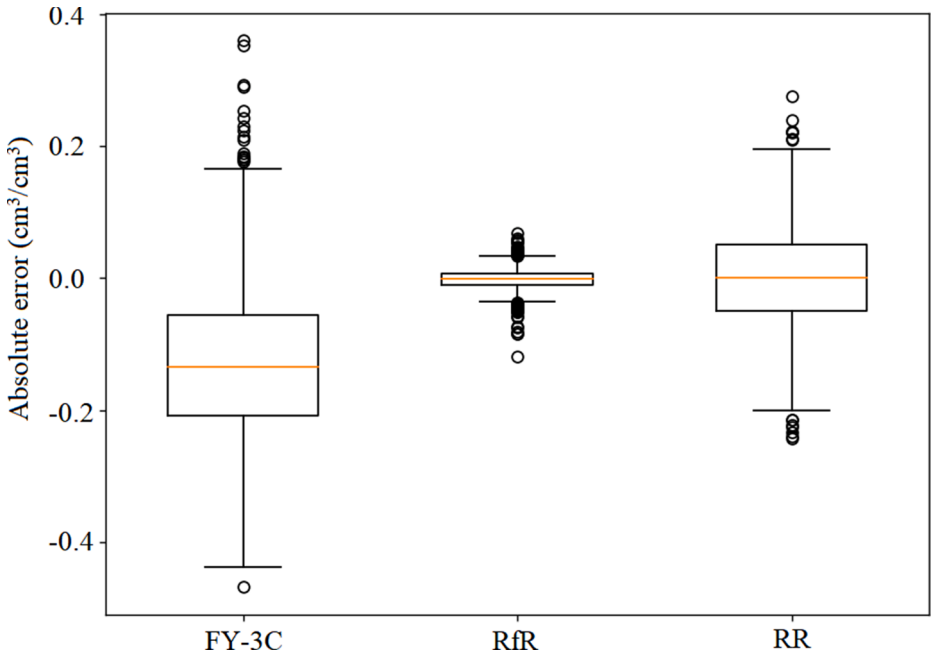

Fig. 5. The absolute errors of the satellite-based, in situ measured and estimated SSM values.

surface soil moisture in the study area. Besides, the importance of those features was measured by RfR model to enhance the interpretability of the established model and its regression results.

To quantitatively assess the performance of the RR and RfR models, the absolute errors between the in situ measurements, the FY-3C retrievals, as well as the SSM estimations using the RR and RfR models, were calculated site-by-site and depicted as Fig. 5. The results showed that both RfR and RR models demonstrated a good performance for estimating the SSM values, with their median values of the absolute errors approximating to 0.0 cm^3/cm^3. Additionally, the distribution ranges of the RR errors between the 25% (Q1) and 75% (Q3) quantiles ([-0.05, 0.05]) were wider than those for the RfR model, whose absolute errors between Q1 and Q3 were distributed quite close to 0. The results indicated that the ML models with multiple inputs, especially the RfR model, were suitable and promising approaches for achieving SSM estimations with high accuracy.

4.3. The importance of input features in the RfR model

The importance of each input feature was calculated using the RfR model and shown in Fig. 6. Generally, geographical location played the most important role in the comprehensive SSM estimation, and the proportions of the total importance made up by latitude and longitude were 35.84% and 16.96%, respectively. In addition to the geographical location, the elevation was also highly relevant to the SSM, occupying a proportion of 14.88%, followed by the vegetation coverage as reflected by the MODIS NDVI, with a proportion of 9.75%. The FY-3C satellite data, soil texture information and the seasonal variation characteristics contributed 8.30%, 8.04% and 6.23% to the RfR model for estimating the regional SSM in the study area, respectively. To be specific, the soil texture data included the proportions of the clay, sand and silt, which contributed 2.07%, 2.97% and 3.00% to the RfR model respectively. Regarding to the seasonal variation characteristics, i.e., the importance of twelve months from January to December, the importance of the months from May to November were 0.60%, 0.58%, 0.56%, 0.72%, 0.64%, 1.07% and 0.50% respectively, which was slightly higher than that of the months from December to April, whose proportions were 0.34%, 0.20%, 0.24%, 0.34% and 0.43% respectively.

4.4. Results of monitoring the regional SSM using the RfR model

In this study, the RfR model was adopted to monitor the monthly land SSM pixel-by-pixel in the study area. The multisource input features of RfR model included latitude, longitude, elevation, MODIS NDVI, FY-3C SSM product, soil separates and seasonal differences, i.e., the months from January to December. The estimated SSM images in China from January 2017 to December 2017 were mapped as shown in Fig. 7. According to Fig. 7(a)–(l), the spatial distribution of the SSM images in different months showed a similar pattern, with the SSM values gradually increasing from northwestern to southeastern China. The northwestern part of the study area with the SSM values lower than 0.20 cm^3/cm^3 was much drier than the southern and northeastern parts, where the SSM values were greater than 0.30 cm^3/cm^3. Regarding the temporal variation characteristics, the proportion of dry areas in the north, such as the North China Plain, slightly increased from April to June and then decreased from July to September 2017. In contrast, the areas with SSM values lower than 0.25 cm^3/cm^3 in the southern region gradually increased from July 2017 and then became wetter since September 2017. These results matched the spatial and seasonal precipitation patterns of the study area well (Gao et al., 2020), indicating that the established RfR model with the selected input features provided accurate and consistent land SSM monitoring results.

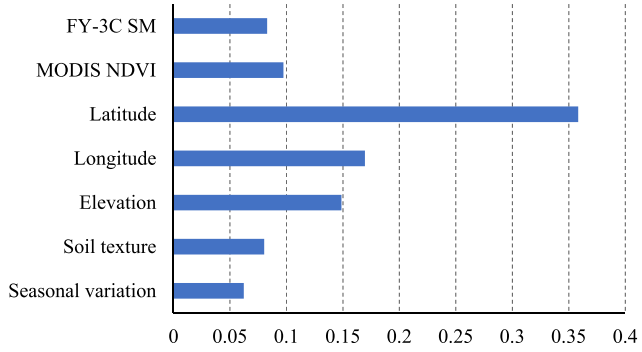

Fig. 6. Features importance generated for the RfR model.

4.5. Discussion

4.5.1. The importance of the soil moisture ground measurements

The integration of satellite images, in situ soil moisture data and ML models is a promising method to achieve accurate and consistent land surface soil moisture data at the regional scale. In this study, the volumetric soil moisture values at the 0–10 cm level measured by the CASMOS stations were employed as the reference dataset for the establishment and validation of the ML models. The large number of in situ soil moisture measurements from more than 2000 meteorological stations across the study area provided massive training and testing samples for the SSM estimation models and therefore played an important role in this study. Although the calibrated in situ measurements were crucial and useful, their spatial resolution (point-scale) was quite different from the satellite observations. Field investigations could be conducted in the future to improve the reference datasets and optimize the current method.

4.5.2. The high correlation between the precipitation patterns and the SSM distribution monitored by the established RfR model

The present study applied five frequently used machine learning methods to estimate the land surface (0–10 cm) soil moisture in the study area. More input features were considered and more reliable and interpretable results were achieved compared with the authors' former research (Wang et al., 2020). For example, the established RfR model, which achieved the best performance here, was employed to measure the importance of each input feature. The results demonstrated a high correlation between the geographical location and the distribution of the in situ measurements. This is likely to be caused by the gradually increasing rainfall with decreasing latitude and increasing longitude (Fig. 8) under the influence of the monsoonal climate. The importance of the remotely sensed SSM retrieved from the FY-3C satellite was 8.30%, which was lower than the latitude, longitude, elevation and vegetation information. The possible reason is that the soil levels detected by the satellite (0–5 cm) and the in situ sensors (0–10 cm) were mismatched, thus resulting in significant differences between the satellite-retrieved and in situ-measured SSM values.

4.5.3. Additional features and models to be considered in the future research

Although the current results indicated the feasibility of the ML models with multiple inputs from various data sources for achieving reliable and consistent SSM estimates, the models can be further trained to achieve wider and more accurate estimating results with additional related features, such as the precipitation and land cover classifications, as inputs in future research. The RfR model, as a type of ensemble learning method, has proven its suitability and superiority for regional SSM estimation compared with the other methods. To extend the current research, more ensemble ML models, as well as the various deep learning approaches, such as convolutional neural networks, should be further developed. Accordingly, the number of the samples needs to be further increased to obtain the full potential of those approaches. Moreover, the models established in this study are only applicable in a

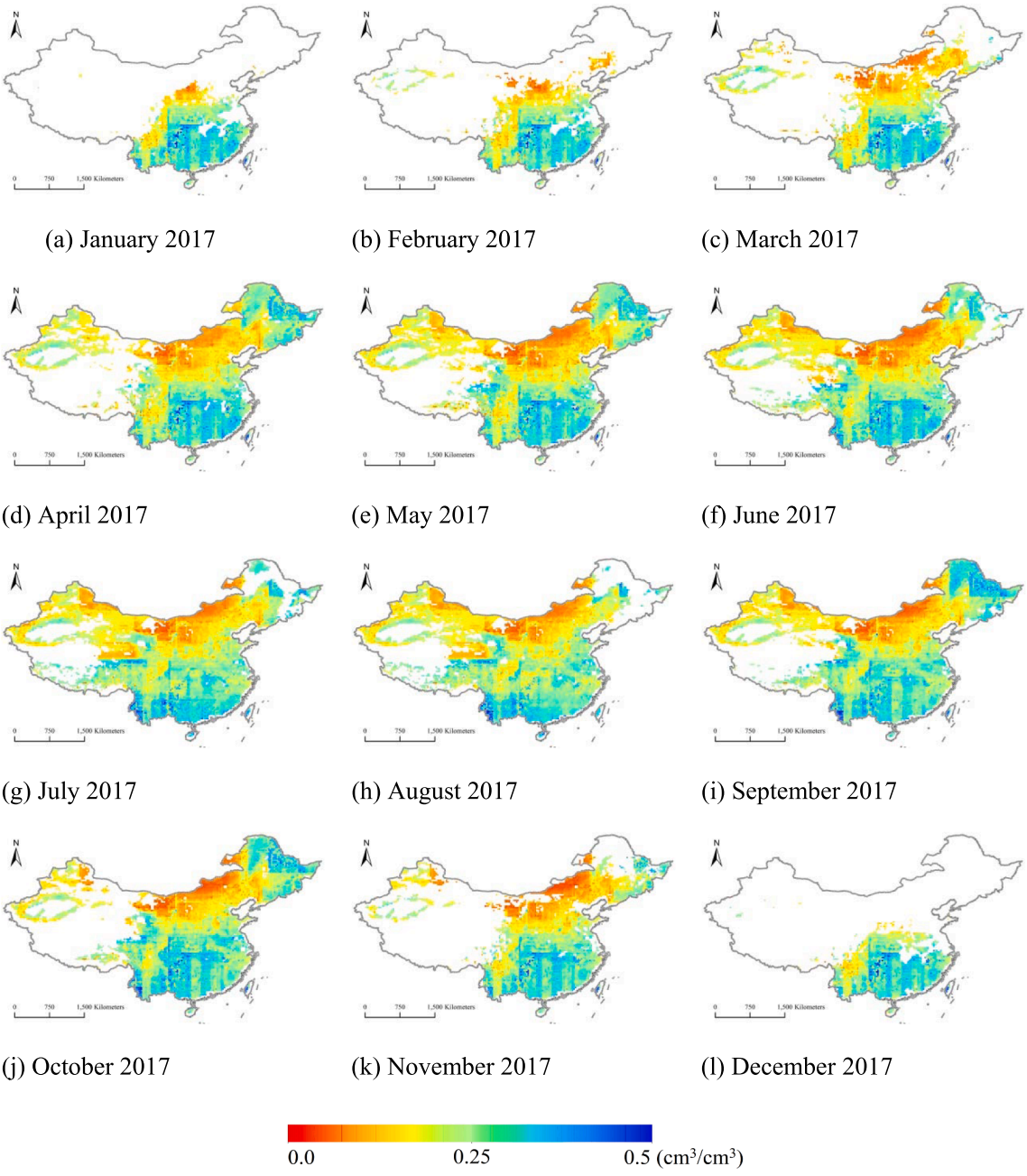

Fig. 7. Monthly estimated SSM (cm^3/cm^3) values obtained using the RfR model from January 2017 to December 2017.

certain region, and their input features and main parameters would need to be adjusted to employ them for SSM monitoring in other regions.

5. Conclusions

The consistency assessment between the in situ measured and the remotely sensed SSM indicated a relatively high SSM monitoring error using FY-3C satellite data during different months (ubRMSE greater than 0.10 cm^3/cm^3), which indicated that it is insufficient to achieve more accurate SSM monitoring results based on the single remotely sensed data source. Multiple features, including the remotely sensed SSM from the FY-3C satellite, the vegetation information represented with MODIS NDVI, the seasonal characteristics, the soil information from HWSD and the in situ measurements from CASMOS, were selected as the inputs for five ML models to obtain accurate and consistent SSM estimations at the regional scale. Among those proposed ML models, the ensemble learning method RfR achieved the best performance during both the training (R = 0.981, MRE = 7.3% and ubRMSE = 0.021 cm^3/cm^3) and testing (R = 0.789, MRE = 22.2% and ubRMSE = 0.065 cm^3/cm^3) processes, followed by the RR and PR models. Additionally, the SSM monitoring images obtained using the RfR model were more consistent with the ground soil moisture than the pure FY-3C SSM product and the SSM estimations from the RR model. These results indicate the superiority of the RfR method for accurately monitoring the land SSM across the study area.

Geographical location was the most crucial input feature according to the importance values generated by the RfR model, followed by the elevation, MODIS NDVI, FY-3C SSM and soil texture. The months from January to December were the least important among all the input features. The SSM images from January to December 2017 estimated by the best ML model showed a gradually decreasing trend from the northwestern to the southeastern part of the study area. The results

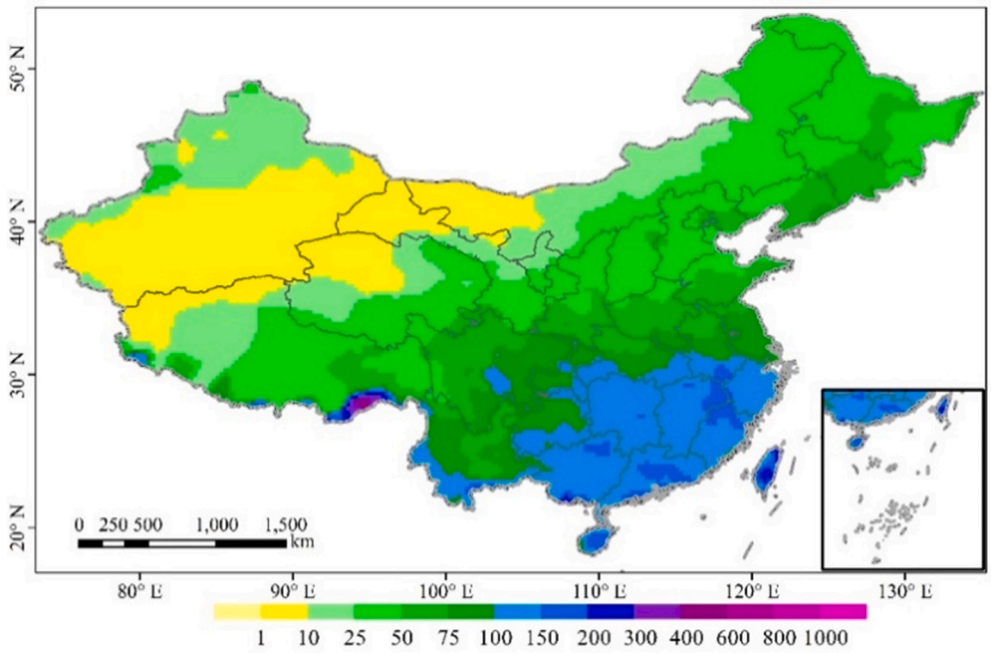

Fig. 8. The spatial distribution of the monthly rainfall (mm) from 1891 to 2016 in the study area.

matched the spatial and seasonal distribution of the historical monthly rainfall in the study area well. Therefore, the comprehensive application of data from multiple data sources combined with the appropriate ML model is a promising strategy for improving SSM estimation accuracy at the national scale.

Declaration of Competing Interest

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

Acknowledgments

This manuscript was supported by the National Key Research and Development Programs of China under Grant 2019YFC1510200, National Natural Science Foundation of China under Grant 42075193 and the Fundamental Research Funds under Grant 2019Z010.

References

Bircher, S., Skou, N., Jensen, K.H., Walker, J.P., Rasmussen, L., 2012. A soil moisture and temperature network for SMOS validation in Western Denmark. Hydrol. Earth Syst. Sci. 16 (5), 1445–1463.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Choi, M., Hur, Y., 2012. A microwave-optical/infrared disaggregation for improving spatial representation of soil moisture using AMSR-E and MODIS products. Remote Sens. Environ. 124, 259–269.

Crow, W.T., Kustas, W.P., Prueger, J.H., 2008. Monitoring root-zone soil moisture through the assimilation of a thermal remote sensing-based soil moisture proxy into a water balance model. Remote Sens. Environ. 112 (4), 1268–1281.

Danielson, J., Gesch, D., 2011. Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010). U.S. Geological Survey Open-File Report 2011-1073 26p.

Gao, X., Guo, M., Yang, Z., Zhu, Q., Xu, Z., Gao, K., 2020. Temperature dependence of extreme precipitation over mainland China. J. Hydrol. 583, 12.

Huang, J., Zhuo, W., Li, Y., Huang, R., Sedano, R., Su, W., Dong, J., Tian, L., Huang, Y., Zhu, D., Zhang, X., 2020. Comparison of three remotely sensed drought indices for assessing the impact of drought on winter wheat yield. Int. J. Digit. Earth 13 (4), 504–526.

Jackson, T., Mansfield, K., Saafi, M., Colman, T., Romine, P., 2008. Measuring soil temperature and moisture using wireless MEMS sensors. Meas. 41 (4), 381–390.

Jones, P.G., Thornton, P.K., 2015. Representative soil profiles for the Harmonized World Soil Database at different spatial resolutions for agricultural modelling applications. Agr. Syst. 139, 93–99.

Kornelsen, K.C., Coulibaly, P., 2013. Advances in soil moisture retrieval from synthetic aperture radar and hydrological applications. J. Hydrol. 476, 460–489.

Li, Z., Sun, G., He, C., Liu, X., Zhang, R., Li, Y., Zhao, D., Liu, H., Zhang, F., 2019. Multivariable regression methods using modified Chebyshev polynomials of class 2. J. Comput. Appl. Math. 346, 609–619.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Rodriguez-Fernandez, N.J., de Souza, V., Kerr, Y.H., Richaume, P., Al Bitar, A., 2017. Soil moisture retrieval using SMOS brightness temperatures and a neural network trained on in situ measurements. IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 1574–1577.

Sabaghy, S., Walker, J.P., Renzullo, L.J., Jackson, T.J., 2018. Spatially enhanced passive microwave derived soil moisture: Capabilities and opportunities. Remote Sens. Environ. 209, 551–580.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Ziese, M., 2018. GPCC Full Data Monthly Product Version 2018 at 0.25°: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historical Data.

Sekertekin, A., Marangoz, A.M., Abdikan, S., 2020. ALOS-2 and Sentinel-1 SAR data sensitivity analysis to surface soil moisture over bare and vegetated agricultural fields. Comput. Electron. Agr. 171 (105303).

Shi, J., Jiang, L., Zhang, L., Chen, K., Wigneron, J., Chanzy, A., Jackson, T.J., 2006. Physically based estimation of bare-surface soil moisture with the passive radiometers. IEEE Trans. Geosci. Remote Sens. 44 (11), 3145–3153.

Singh, V.K., Singh, B.P., Kisi, O., Kushwaha, D.P., 2018. Spatial and multi-depth temporal soil temperature assessment by assimilating satellite imagery, artificial intelligence and regression based models in arid area. Comput. Electron. Agr. 150, 205–219.

Srivastava, P.K., 2017. Satellite Soil Moisture: Review of Theory and Applications in Water Resources. Water Resour. Manag. 31 (10), 3161–3176.

Tibshirani, R., 1998. The lasso method for variable selection in the cox model. Stat. Med. 16 (4), 385–395.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., Traver, I.N., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., L'Abbate, M., Croci, R., Pietropaolo, A., Huchler, M., Rostan, F., 2012. GMES Sentinel-1 mission. Remote Sens. Environ. 120, 9–24.

Vreugdenhil, M., et al., 2013. Towards a high-density soil moisture network for the validation of SMAP in Petzenkirchen, Austria. IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 1865–1868.

Wagner, W., Scipal, K., Pathe, C., Gerten, D., Lucht, W., Rudolf, B., 2003. Evaluation of the agreement between the first global remotely sensed soil moisture data with model and precipitation data. J. Geophys. Res. Atmos. 108 (D19).

Wang, L., Fang, S., Pei, Z., Zhu, Y., Khoi, D.N., Han, W., 2020. Using FengYun-3C VSM Data and Multivariate Models to Estimate Land Surface Soil Moisture. Remote Sens. 12 (6), 1038.

Wang, Y., Yang, X., Lu, Y., 2019. Informative gene selection for microarray classification via adaptive elastic net with conditional mutual information. Appl. Math. Model. 71, 286–297.

Wu, D., Fang, S., Li, X., He, D., Zhu, Y., Yang, Z., Xu, J., Wu, Y., 2019. Spatial-temporal variation in irrigation water requirement for the winter wheat-summer maize

rotation system since the 1980s on the North China Plain. Agric. Water Manage. 214, 78–86.
Zhang, S., Weng, F., Yao, W., 2020. A Multivariable Approach for Estimating Soil Moisture from Microwave Radiation Imager (MWRI). J. Meteorol. Res. 34 (4), 732–747.
Zhao, S., Cong, D., He, K., Yang, H., Qin, Z., 2017. Spatial-temporal variation of drought in China from 1982 to2010 based on a modified temporature vegetation drought index (mTVDI). Sci. Rep. 7 (17473).
Zhu, Y., Li, X., Pearson, S., Wu, D., Sun, R., Johnson, S., Wheeler, J., Fang, S., 2019. Evaluation of Fengyun-3C Soil Moisture Products Using In-Situ Data from the Chinese Automatic Soil Moisture Observation Stations: A Case Study in Henan Province, China. Water 11 (2), 23.
Ziegler, A., Koenig, I.R., 2014. Mining data with random forests: current options for real-world applications. Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery 4 (1), 55–63.
Zou, H., 2020. Comment: Ridge Regression-Still Inspiring After 50 Years. Technometrics 62 (4), 456–458.