Potential of Remote Sensing Images for Soil Moisture Retrieving Using Ensemble Learning Methods in Vegetation-Covered Area

Ya Gao, Liguo Wang, Geji Zhong, Yitong Wang, and Jinghui Yang

***Abstract*—Soil moisture (SM) plays a critical role in various fields such as agriculture, hydrology, and land-atmosphere interactions. This study aims to evaluate the performance of the categorical boosting algorithm (CatBoost) in comparison to other multiple-boosting algorithms for SM prediction. Appropriate feature selection is vital for achieving accurate predictions, and this study focuses on identifying relevant features and assessing CatBoost's suitability for the task. The study incorporates several boosting algorithms including Gradient Boosting Decision Tree (GBDT), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and CatBoost to estimate SM. Results indicate that radar backscatter coefficient, soil roughness, and digital elevation model (DEM) are crucial features for SM retrieval. Comparatively, CatBoost outperforms GBDT, XGBoost, and LightGBM in various feature combinations. The most favorable results are obtained when utilizing all features as inputs for the algorithm. These optimal results yield a mean absolute error (MAE) of 2.40 vol.%, mean relative error (MRE) of 0.16 vol.%, root mean square error (RMSE) of 3.26 vol.%, and Pearson correlation coefficient of 0.73. Additionally, the study analyzes the inversion results for different ranges of SM and Normalized Difference Vegetation Index (NDVI). Within the range of SM from 0 to 25 vol.% and NDVI from 0 to 0.7, utilizing all features yields the most accurate results. Using CatBoost, this approach achieves an MAE of 1.52 vol.%, MRE of 0.12 vol.%, RMSE of 2.11 vol.%, and R of 0.81. The study suggests that applying boosting algorithms, especially CatBoost, holds promise in accurately estimating surface SM.**

***Index Terms*—Boosting, categorical boosting algorithm (CatBoost), extreme gradient boosting (XGBoost), gradient boosting decision tree (GBDT), light gradient boosting machine (lightGBM), sentinel-1, soil moisture retrieval.**

Manuscript received 30 June 2023; revised 4 August 2023; accepted 21 August 2023. Date of publication 1 September 2023; date of current version 13 September 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62071084 and Grant 62001434 and in part by the Leading Talents Project of the State Ethnic Affairs Commission. *(Corresponding author: Liguo Wang.)*

Ya Gao and Yitong Wang are with the College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China (e-mail: gaoya0001@hrbeu.edu.cn; wangyitong@hrbeu.edu.cn).

Liguo Wang is with the College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China, and also with the College of Information and Communication Engineering, Dalian Minzu University, Dalian 116600, China (e-mail: wangliguo@hrbeu.edu.cn).

Geji Zhong is with the School of Ethnology and Sociology, Minzu University of China, Beijing 100081, China (e-mail: 21410014@muc.edu.cn).

Jinghui Yang is with the School of Information Engineering, China University of Geosciences, Beijing 100083, China (e-mail: yangjh@cugb.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3311096

I. Introduction

SOIL moisture (SM) is a pivotal parameter in the realm of agricultural development, hydrological science, and climate change studies [1], [2], [3]. Microwave remote sensing technology has emerged as a powerful means to retrieve SM due to its ability to capture the sensitivity of radar signals to variations in SM [4], [5], [6]. It also works in all times and weather conditions.

Traditional methods of monitoring SM, such as ground-based sampling, are mainly point-based and have a limited number of samples. Establishing ground observation stations to monitor SM in real time is too costly. Remote sensing technology has gradually become an important tool for monitoring SM in recent years. Especially with the development of microwave remote sensing, more and more scholars are using microwave remote sensing to estimate SM [7].

The presence of vegetation complicates the SM inversion. Vegetation produces complex scattering that reduces the radar signal's sensitivity to SM and contributes to the total backscattering. Therefore, current SM inversion models for vegetation-covered areas account for both the scattering contribution of vegetation and the attenuation of surface backscattering [8], [9]. The water cloud model (WCM), is widely used to remove the effect of vegetation during the quantification of SM content. The WCM, which was proposed by Attema and Ulaby [10], was first used to explain the mechanism of crop microwave backscattering. Recently, both physical model-based techniques and machine learning methods have advanced rapidly for estimating SM using microwave remote sensing satellite data in vegetation-covered area [11], [12], [13], [14], [15]. Yadav et al. [16] estimated and evaluated SM by using modified WCM (MWCM) and the evaluation of potential of multitarget random forest regression (MTFER) in the vegetated areas. The results showed that MTRFR combined with MWCM technology could accurately retrieve SM [16]. Based on the artificial neural network model (ANN), Zhang et al. [13] established an ability to estimate SM at different stages of corn growth. The results suggest that the method had operational potential in estimating SSM from Terra-SAR and Landsat-7 data at different stages of early corn growth. Using Sentinel-1 SAR and Landsat-8 optical data, Yang et al. (2020) inverted SM by combining the WCM and deep belief network (DBN). They used tenfold cross-validation to show that DBN performed more stably with different data when combined with the WCM model [17]. Using ALOS-2 microwave data,

Gao et al. [18] developed a genetic algorithm-based method for optimizing BP neural networks (GA-BP) to estimate SM in vegetated areas. Cui et al. [19] established an ANN SMC retrieval algorithm combined with the water cloud model, the advanced integral equation model, and the Oh model to perform SM inversion in high spatial resolution over agroforestry areas. Jarray et al. [20] developed an ANN, extreme gradient boosting (XGBoost), random forest regressor, combined with WCM to retrieval SM with Sentinel-1 and Sentinel-2. Wang et al. [21] used Sentinel-1 and Sentinel-2 data to estimate SM in vegetated areas by combining the WCM and integrated learning models (RF and AdaBoost). They found that RF performed better than AdaBoost when combined with the WCM model [21].

The boosting method begins with a weak learning algorithm and iteratively learns a series of weak classifiers (also called basic classifiers), and then combines these weak classifiers to form a strong classifier in some form. Most of the boosting methods learn a series of weak classifiers by changing the probability distribution of the training data set (weights of different samples of the training data) and calling weak classification algorithms for data with different probability distributions. Boosting not only solves the classification problem, but also the prediction problem. Boosting algorithms such as GBDT, XGBoost, LightGBM, and CatBoost can be used for SM inversion. Zhang et al. proposed a novel approach for high-resolution soil water retrieval using ensemble learning. They combined landSAT8 optical data with various other data sources to improve the accuracy of their retrieval method. The ensemble learning technique integrated two algorithms, namely XGBoost and RF. The experimental results demonstrated that the XGBoost model exhibited a slightly superior performance compared to the RF model [22]. Mehdi Jamei et al. employed the XGBoost and CatBoost algorithms to investigate the long-term, multi-step advance day Root Zone Soil Moisture (RZSM) in the severely cold and warm semi-arid regions of Iran [23]. By utilizing multiple source datasets, Yang et al. conducted an evaluation of four ensemble learning models, namely RF, Extreme Random Trees (ERT), XGBoost, and LightGBM, for the purpose of soil water retrieval [24]. Liu et al. conducted a comprehensive comparative analysis of soil moisture active passive (SMAP) downscaling techniques in the southwestern region of France. The study employed Classification and Regression Trees (CART), RF, GBDT, and Extreme Gradient Boosting (XGB) algorithms for the assessment [25]. However, there is no clear evidence to show which algorithm is more suitable for this task.

The purpose of this study is to establish a SM inversion algorithm based on Sentinel-1 and multisource datasets using the boosting algorithms of ensemble learning algorithm. In this study, we compared the accuracy of SM inversion using four boosting algorithms (GBDT, XGBoost, LightGBM, and CatBoost) in combination with the WCM model. Our aim was to determine, which method was more suitable for inversion in the study area. At the same time, we will evaluate the applicability of the inversion method in different humidity regions.

The rest of this article is organized as follows. Section II describes the materials and methods used in this study. Section III presents the results of the SM inversion using different boosting algorithms. Section IV discusses the implications and limitations of the findings. Finally, Section V concludes this article.

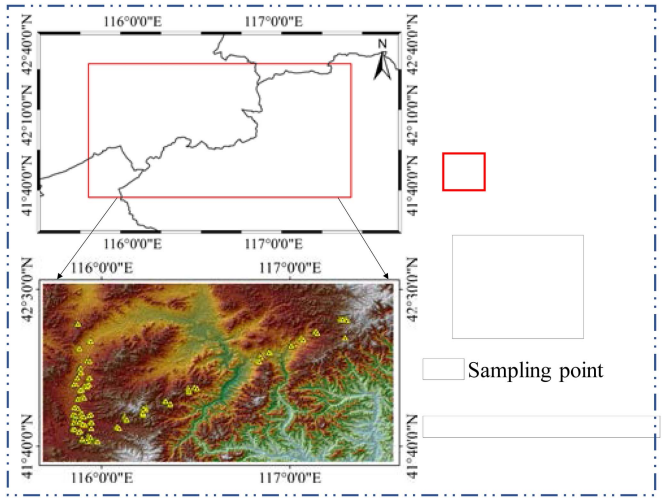

Fig. 1. Geographical location map of the study area.

II. Materials and Methods

A. Study Area and Materials

The upper Luan River basin was selected as the study area for this article. Fig. 1 shows the geographical location of the study area. In this study area, a ground synchronized observation experiment was conducted for the remote sensing experiment of SM in the Luan River Basin. The study area has vegetation cover, including grassland, weedy field, corn field, cauliflower field, potato field, and carrot field. Data were obtained from http://data.tpdc.ac.cn. The simultaneous observation dataset includes soil temperature and moisture as well as soil roughness. In this article, we used SM and soil roughness data (SRD) in September 19, 2018. They selected data on SM at a depth of 0–5 cm. The range of SM is 2 to 40 vol.%. After conducting the screening process, we obtained 917 points of valid SM data. We divided the SM data into two sets: 1) a training dataset (60%) and 2) a validation dataset (40%). The experimental area is located in the upper reaches of the Luanhe River, where they divided it into two routes to measure the soil roughness. In the north and south direction, the terrain is complex and there are many types of typical surface features, such as grassland, agricultural land, wasteland, bare land, and forest land. The surface fluctuation of different surface features in the same large square varies greatly. The northeast southwest direction features a simple type, which is mostly grassland. The root mean square height and correlation length were 0.4 to 3.4 cm and 6 to 34 cm, respectively, in this article.

To obtain the remote sensing data, we downloaded the Sentinel-1 and Sentinel-2 data from the Alaska Satellite Facility[1] and the Earth Explorer[2] websites, respectively. We selected the images that matched the dates of the measured data. Sentinel-1,

[1][Online]. Available: https://asf.alaska.edu/
[2][Online]. Available: https://earthexplorer.usgs.gov/

an Earth observation satellite launched by the European Space Agency (ESA) as part of the Copernicus program, serves as a crucial asset for obtaining high-quality, continuous radar imaging data. Designed to facilitate the monitoring of Earth's surface changes, Sentinel-1 offers all-weather and all-day remote sensing capabilities. Sentinel-2, a crucial facet of the Copernicus program developed and launched by the ESA, serves as a cornerstone for gathering high-resolution and multispectral image data. The satellite's main objectives revolve around the monitoring of surface alterations, meticulous examination of terrestrial and coastal ecosystems, and fostering effective land use and environmental management. In this study, we used data resolution of 30 m.

The preprocessing steps for Sentinel-1 data encompass several vital stages: 1) Data acquisition; 2) Radiometric correction; 3) Atmospheric correction; 4) Geometric correction and registration; 5) Multitemporal data processing; 6) Image denoising and filtering; 7) Data clipping and subset extraction; 8) Data format conversion and storage. The preprocessing steps for Sentinel-2 data typically encompass the following aspects: 1) Geometric correction, radiometric correction, and atmospheric correction; 2) Image stitching and Mosaic; 3) Data tailoring and subset extraction; 4) Data format conversion and storage.

The Sentinel-1 synthetic aperture radar (SAR) images were acquired on September 12, September 16, and September 19, 2018. The Sentinel-2 images were selected to have the same or similar acquisition time as the Sentinel-1 images. We performed preprocessing on the radar signal and calculated the backscattering coefficients of VV and VH polarization. We also calculated the normalized difference vegetation index (NDVI) from the Sentinel-2 images using band arithmetic: NDVI $=$ (band8 $-$ band4)/(band8 $+$ band4), where band8 and band4 are the near-infrared and red bands, respectively.

Digital elevation model (DEM) is a vital data source for studying and analyzing terrain, watershed, and feature identification. We used the ASTER global DEM (GDEM) data, which were derived from observations of NASA's Terra satellite. This data set provides the most comprehensive and accurate global elevation data available. We extracted DEM, slope, and aspect values from the GDEM as the terrain parameters for this study. The spatial resolution of the GDEM was 30 m.

B. Ensemble Learning Algorithm

Ensemble learning is a technique that combines multiple models (weak learners) to achieve better prediction performance. A proper combination of weak learners can produce a more accurate and robust model. Boosting is a powerful ensemble learning method, which is also a supervised classification method. It constructs a strong classifier by iteratively adding weak classifiers and adjusting their weights according to the prediction errors. The boosting algorithm follows these steps.

Step 1: Assign equal weights to each observation and train the first base learner on the data.

Step 2: Increase the weights of the misclassified observations and train the next base learner on the weighted data.

Step 3: Repeat step 2 until a predefined number of learners or a predefined prediction accuracy is reached.

Algorithm 1: GBDT Pseudocode.

```
Input: training set {(x1,y1),…,(xn,yn)}, where xi is a
feature vector and yi is a label
Output: a strong classifier H(x)
Initialize the strong model as a constant: F0(x) =
argmin_c sum(L(yi,c)) for i = 1,…,n
For t = 1,…,T:
  Compute the negative gradient (pseudo response) for
   each instance: ri = -[dL(yi,Ft-1(xi))/dFt-1(xi)]
Train a regression tree ht(x) using the feature vector x
 and pseudo response r as labels
  Compute the optimal multiplier gamma_t for ht(x) by
   solving: gamma_t = argmin_gamma
   sum(L(yi,Ft-1(xi)+gamma∗ht(xi))) for i = 1,…,n
  Update the strong model as Ft(x) = Ft-1(x) + gamma_t
   ∗ ht(x)
Return H(x) = FT(x)
```

There are many specific algorithms for boosting, and in this article we choose four of them: GBDT, XGBoost, LightGBM, and CatBoost.

1) Gradient Boosting Decision Tree: GBDT is a gradient boosting decision tree, where the output of a GBDT model is the sum of several decision trees, each of which is a fit to the residuals of the previous combination of decision trees, a "correction" to the previous model. Gradient boosting trees can be used for both regression problems (in this case known as CART regression trees) and classification problems (in this case known as classification trees. GBDT is an addition model based on Boosting's integration thought. During training, the front-to-distribution algorithm is used for greedy learning. At each iteration, a CART tree is learned to fit the residual between the prediction result of the previous t-1 tree and the real value of the training sample. Trees in GBDT are regression trees, and GBDT is used to make regression predictions. GBDT is to sum up the conclusions of all the trees to make the final conclusion. The core of GBDT is that each tree learns the residual (negative gradient) of the conclusion and all the previous trees. This residual is the cumulative amount of the real value after adding the predicted value [26].

Where, n is the number of instances in the training set. T is the number of decision trees. L is the loss function, such as squared error loss or logistic regression loss. Ft is the strong model obtained after the *t*th iteration. ht is the *t*th round of iterative training to get the decision tree (decision tree). ri is the pseudoresponse value (pseudoresponse value) obtained by the *i*th instance in the t-round iteration, i.e., the loss function obtains a negative gradient value for the output value of the strong model. gamma_t is the optimal multiplier (optimal multiplier) obtained from the output value of the decision tree in the *t*th iteration, i.e., the loss function is minimized to obtain the step size value.

2) Categorical Boosting Algorithm: CatBoost is a decision tree model based on gradient boosting. The main problem addressed is to handle categorical features efficiently and rationally, which can be seen from its name. CatBoost is composed of Categorical and Boosting. In addition, CatBoost solves the problem of gradient bias and prediction bias, thus reducing

Algorithm 2: CatBoost Pseudocode.

Input: training set {(x1,y1), …,(xn,yn)}, where xi is a feature vector and yi is a label
Output: a strong classifier H(x)
Initialize the strong model as a constant: F0(x) = argmin_c sum(L(yi,c)) for i = 1, …,n
For t = 1, …,T:
  Randomly permute the training set
  For i = 1, …,n:
    Compute the gradient statistics for xi using only the instances before it in the permutation: gi = -[dL(yi,Ft-1(xi))/dFt-1(xi)]
    Train a regression tree ht(x) using only categorical features of x and gradient statistics g as labels
  Compute the optimal multiplier gamma_t for ht(x) by solving: gamma_t = argmin_gamma sum(L(yi,Ft-1(xi)+gamma∗ht(xi))) for i = 1, …,n
  Update the strong model as Ft(x) = Ft-1(x) + gamma_t ∗ ht(x)
Return H(x) = FT(x)

Algorithm 3: XGBoost Pseudocode.

Input: training set {(x1,y1), …,(xn,yn)}, where xi is a feature vector and yi is a label
Output: a strong classifier H(x)
Initialize the strong model as a constant: F0(x) = argmin_c sum(L(yi,c)) for i = 1, …,n
For t = 1, …,T:
  Compute the gradient g_t and hessian h_t for each instance using the current model: g_ti = [dL(yi,Ft-1(xi))/dFt-1(xi)], h_ti = [d2L(yi,Ft-1(xi))/dFt-1(xi)2] for i = 1, …,n
  Train a regression tree ht(x) using the feature vector x and (g_t,h_t) as labels
  Compute the optimal multiplier gamma_tj for each leaf j of ht(x) by solving: gamma_tj = -sum(g_ti)/[sum(h_ti)+lambda] for i in leaf j
  Update the strong model as Ft(x) = Ft-1(x) + eta ∗ sum(gamma_tj ∗ I(x in leaf j)) for j = 1, …,J
Return H(x) = FT(x)

the occurrence of overfitting and, thus improving the accuracy and generalization of the algorithm [27], [28]. CatBoost is a GBDT framework based on symmetric decision tree. It mainly deals with Categorical features rationally by taking examples. CatBoost is composed of categorical and Boosting. CatBoost also solves the problem of gradient bias and prediction offset, thereby reducing the occurrence of overfitting and improving the accuracy and generalization ability of the algorithm [27], [29].

Where, n is the number of instances in the training set. T is the number of decision trees. L is the loss function, such as squared error loss or logistic regression loss. Ft is the strong model obtained after the *t*th iteration. ht is the *t*th round of iterative training to get the decision tree (decision tree). gi is the gradient statistic obtained by the *i*th instance in the t-round iteration, i.e., the loss function obtains a negative gradient value for the output value of the strong model, but it is only calculated using the instance before it in the arrangement to avoid information leakage. gamma_t is the optimal multiplier (optimal multiplier) obtained from the output value of the decision tree in the *t*th iteration, i.e., the loss function is minimized to obtain the step size value.

3) Extreme Gradient Boosting: XGBoost is one of the boosting algorithms. the idea of the Boosting algorithm is to integrate many weak classifiers together to form a strong classifier. Since XGBoost is a boosting tree model, it integrates many tree models together to form a very strong classifier. And the tree model used is the CART regression tree model.

The process can also be divided into two types from the algorithm pseudocode. The global approximation is to calculate the quantile points and divide the samples for each feature before a new tree is generated, after which the approximate division is used in each splitting process, while the local approximation is to use the approximation algorithm in the process of splitting the nodes at a specific time [30]. XGBoost is a modification of GBDT or a gradient lifting tree. The basic idea of XGBoost is the same as GBDT, but with some optimizations, such as default missing value handling, adding second derivative information, regularization terms, column sampling, and parallel computation [31].

Where n is the number of instances in the training set. T is the number of decision trees. L is the loss function, such as squared error loss or logistic regression loss. Ft is the strong model obtained after the *t*th iteration. ht is the *t*th round of iterative training to get the decision tree (decision tree). g_ti and h_ti are the gradient value and hessian value obtained by the *i*th instance in the t-round iteration, i.e., the loss function obtains the first-order derivative and the second-order derivative for the output value of the strong model. gamma_tj is the decision tree for each decision tree in the *t*th iteration.

4) Light Gradient Boosting Machine: LightGBM is a framework for implementing the GBDT algorithm, supporting efficient parallel training. lightGBM uses as much data as possible on a single machine without sacrificing speed. When multiple machines are in parallel, the communication cost is as low as possible and the computation can be linearly accelerated [32]. Gradient-based one side sample (GOSS) is a major innovation proposed by LightGBM to reduce training samples and, thus, improve efficiency. LightGBM is a gradient boosting framework that uses decision trees based on learning algorithms. It is distributed and efficient. It is based on the decision tree algorithm of Histogram and adopts GOSS [24].

Where n is the number of instances in the training set. T is the number of decision trees. L is the loss function, such as squared error loss or logistic regression loss. Ft is the strong model obtained after the *t*th iteration. ht is the *t*th round of iterative training to get the decision tree (decision tree). g_ti and h_ti are the gradient value and hessian value obtained by the *i*th instance in the t-round iteration, i.e., the loss function obtains the first-order derivative and the second-order derivative

Algorithm 4: LightGBM Pseudocode.

Input: training set {(x1,y1),…,(xn,yn)}, where xi is a feature vector and yi is a label
Output: a strong classifier H(x)
Initialize the strong model as a constant: F0(x) = argmin_c sum(L(yi,c)) for i = 1,…,n
For t = 1,…,T:
  Compute the gradient g_t and hessian h_t for each instance using the current model: g_ti = [dL(yi,Ft-1(xi))/dFt-1(xi)], h_ti = [d2L(yi,Ft-1(xi))/dFt-1(xi)2] for i = 1,…,n
  Train a regression tree ht(x) using the feature vector x and (g_t,h_t) as labels
  Compute the optimal multiplier gamma_tj for each leaf j of ht(x) by solving: gamma_tj = -sum(g_ti)/[sum(h_ti)+lambda] for i in leaf j
  Update the strong model as Ft(x) = Ft-1(x) + eta * sum(gamma_tj * I(x in leaf j)) for j = 1,…,J
Return H(x) = FT(x)

for the output value of the strong model. gamma_tj is the optimal multiplier (optimal multiplier) obtained from the output value of each leaf node (leaf node) of the decision tree in the tth iteration, i.e., the loss function is minimized to obtain the step size value.

C. Retrieval Soil Moisture Method

1) Radar Backscattering Model: WCM is a simple function model that is widely used to model soil and vegetation parameters from SAR data and to estimate SM [33], [34], [35]. According to WCM, the total backscattered signal consists of the backscattered signals of soil and vegetation. To estimate soil backscattered signal accurately, we need to remove the backscattered signal of vegetation from the total backscattered signal [10]. WCM can be calculated as follows:

$$\sigma^o_{veg} = AV_1 \cos\theta \left(1 - \tau^2\right) \tag{1}$$

$$\tau^2 = \exp\left(-2B \times V_2/\cos\theta\right) \tag{2}$$

$$\sigma^o_{\text{soil}} = \frac{\sigma^o - AV_1 \times \cos\theta\left[1 - \exp\left(-2B \times V_2 \times \sec\theta\right)\right]}{\exp\left(-2B \times V_2 \times \sec\theta\right)} \tag{3}$$

$$\sigma^0_{\text{soil}} = F(SM, \text{RSD}, \text{sensor} - \text{configuration}) \tag{4}$$

V1 and V2 are vegetation descriptions. In this study, both V1 and V2 are represented by NDVI. Many studies have shown that there is no specific parameter for vegetation description, and NDVI, as a vegetation parameter in WCM, has achieved a good effect and accuracy [8], [34], [36], [37], [38]. θ is angle of incidence, and A and B are parameters, determined by satellite parameters. τ^2 is two-way attenuation. σ^o is the total backscattered signal. σ^o_{soil} and $\sigma^o_{\text{soil}}$are backscattered signal of soil and vegetation, respectively. SM is soil moisture, RSD is soil roughness data (correlation length and root mean square height), and sensor-configuration is radar sensor configuration parameters.

We used WCM calculation to obtain the backscatter signal of soil from SAR data and calculated NDVI from Sentinel-2.

TABLE I
DIFFERENT COMBINATIONS OF PREDICTOR VARIABLES FOR SM CONTENT

Model No .	Variables fusion
M Ⅰ	S-1_VH
M Ⅱ	S-1_VV
MⅢ	S-1_VV+VH
MⅣ	S-1_VV+VH+ SRD
M Ⅴ	S-1_VV+VH+ SRD+DEM

2) Feature Selection and Inversion Method: The soil measured data included SM, correlation length and root mean square height, which ranged from 6 to 34 cm and 0.4 to 3.4 cm, respectively. The range of SM was 2 to 40 vol.%. We also calculated slope, elevation and aspect from ASTERGDEM. Therefore, we selected S-1 radar signal, SRD and DEM as features to retrieve SM. To select the most favorable features for SM inversion, we calculated the Pearson correlation coefficients of different feature parameters with SM and observed whether they had a significant correlation. We then used an ensemble learning approach based on boosting algorithms to retrieve SM. During the modeling process of SM retrieval, the remote sensing feature factor assumes a crucial role in capturing the correlation between remote sensing data and SM. Moreover, the effectiveness of combining these feature factors is a significant determinant when choosing suitable feature combinations. The degree of alignment between the feature factor subset and the estimation model further impacts the accuracy of SM estimation. Therefore, it is imperative to combine and carefully select the characteristic factors to identify the optimal feature combination for SM retrieval research.

We performed different feature parameter combinations with the established system dataset and compared and analyzed, which feature parameter gave the highest accuracy results when used as input. The feature combinations were as follows: MI: S-1_VH; MII:S-1_VV; MIII: S-1_VV+VH; MIV:S-1_VV+VH+ SRD; MV:S-1_VV+VH+ SRD+DEM. The sampling of soil moisture dataset had 917 points, 60% of which were used for training and 40% for validation. The specific methodology employed in this article is outlined as follows.

1) The WCM model was utilized to calculate the backscattering coefficient (VV, VH) of the soil.
2) In conjunction with other multisource data, a feature selection process was conducted to identify parameters that exhibited a strong correlation with soil moisture.
3) The selected feature factors from Step 2 were then combined.
4) The recombined dataset from Step 3 was employed as input for four distinct Boosting algorithms, facilitating soil water retrieval in each case.

Fig. 2 shows the flowchart of SM retrieval using ensemble learning method. Table I presents the five scenarios.

D. Model Performance Evaluation

To evaluate the accuracy of SM retrieval based on ensemble learning, we used four metrics: mean absolute error (MAE),

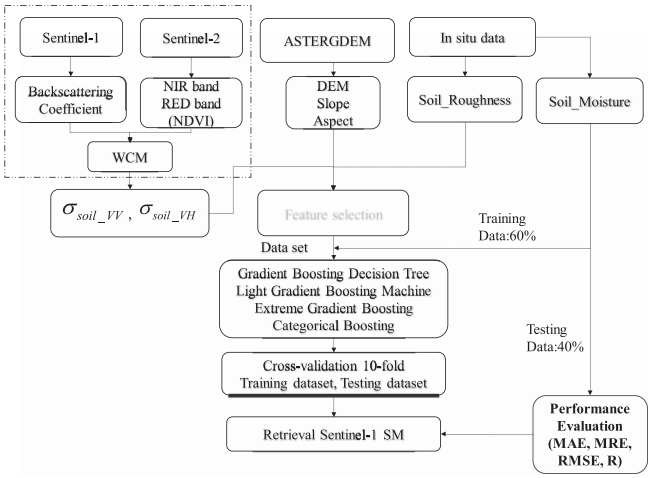

Fig. 2. Flowchart of SM retrieval.

mean relative error (MRE), root-mean-square error (RMSE), and Pearson correlation coefficient R [39]. The formulas for these metrics are as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{y}_i - y_i\right| \tag{5}$$

$$\mathrm{MRE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\widehat{y}_i - y_i\right|}{\widehat{y}_i} \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\widehat{y}_i - y_i\right)^2} \tag{7}$$

$$R = \frac{E\left[\left(\widehat{y}_i - E\left[\widehat{y}_i\right]\right)\left(y_i - E\left[y_i\right]\right)\right]}{\widehat{\sigma}\sigma} \tag{8}$$

where $\widehat{y}_i$ is the predicted SM value and y_i is the true SM value. E[.] is the mean operator. $\widehat{\sigma}$ and σ are the standard deviation of the *in situ* SM and estimated SM, respectively.

III. Results

A. Correlation Analysis of Predictor Indicators and SM

We used Pearson correlation coefficient to compute the relationship between the input variables derived from S-1, SRE, and DEM for each of the five conditions. SPSS was used for correlation and significance analysis. Table II shows that radar signal VV and VH, CL, and DEM had positive correlations with SM, while RMSH had a negative correlation. These features were significantly correlated with SM. However, ASPECT and SLOPE had low significant correlations with SM. Based on these results, we selected VV, VH, RMSH, CL, and DEM as the input parameters of the integrated learning algorithm.

TABLE II
Pearson Correlation Coefficient Analysis of Input Variables and SM

Input variables	Pearson correlation	Sig. (2-tailed)	
VV&SM	-.159**	0.000	**.Correlation is significant at the 0.01 level (2-tailed).
VH&SM	-.164**	0.000	**.Correlation is significant at the 0.01 level (2-tailed).
RMSH&SM	.174**	0.000	**.Correlation is significant at the 0.01 level (2-tailed).
CL&SM	-.093*	0.000	**.Correlation is significant at the 0.01 level (2-tailed).
DEM&SM	-.103**	0.000	**.Correlation is significant at the 0.01 level (2-tailed).
ASPECT&SM	0.029	0.383	-
SLOPE&SM	-0.047	0.152	-

B. Evaluation and Comparison of Scenarios and Different Combinations of Predictor Variables Models

This study explored the utility of four boosting ensemble models for SM retrieval using different combinations of predictor variables. We designed five combinations based on variable characteristics for inverse evaluation analysis. The feature combinations were as follows: MI: S-1_VH; MII:S-1_VV; MIII: S-1_VV+VH; MIV:S-1_VV+VH+SRD; MV:S-1_VV+VH+SRD+DEM. We estimated SM from Sentinel-1 SAR data using these ensemble learning methods. Figs. 3–6 and Table III show the results of the SM retrievals from Sentinel-1 using different four ensemble learning methods.

Fig. 3 shows the inversion results of GBDT method, which "n_estimators," "learning_rate," "subsample," "loss" of the frame parameter we set are 57, 0.12, 0.8, ls, respectively. a1–a5 are the inversion results obtained when different parameters are input. When the system data set has only VH polarization mode radar signal, we can get MAE = 3.10 vol.%, MRE = 0.22 vol.%, RMSE = 4.24 vol.%, and R = 0.55 (see Fig. 3, a1, Table III). In MII mode, MAE is 3.19 vol.%, MRE is 0.21 vol.%, RMSE is 4.21 vol.%, and R = 0.54 (a2). The MV produced a highest prediction accuracy with MAE is 2.91 vol.%, MRE is 0.21 vol.%, RMSE is 3.92 vol.%, and R = 0.60 (a5), followed by MIV with MAE is 3.08 vol.%, MRE is 0.21 vol.%, RMSE is 4.00 vol.%, and R = 0.60 (a4), and MIII with MAE is 3.14 vol.%, MRE is 0.21 vol.%, and RMSE is 4.09 vol.% and R = 0.56 (a3).

The results of LightGBM are shown in Fig. 4 and Table III. Similarly, we set the parameters for the LightGBM algorithm, where num_leaves = 1200, learning_rate = 0.17, n_estimators

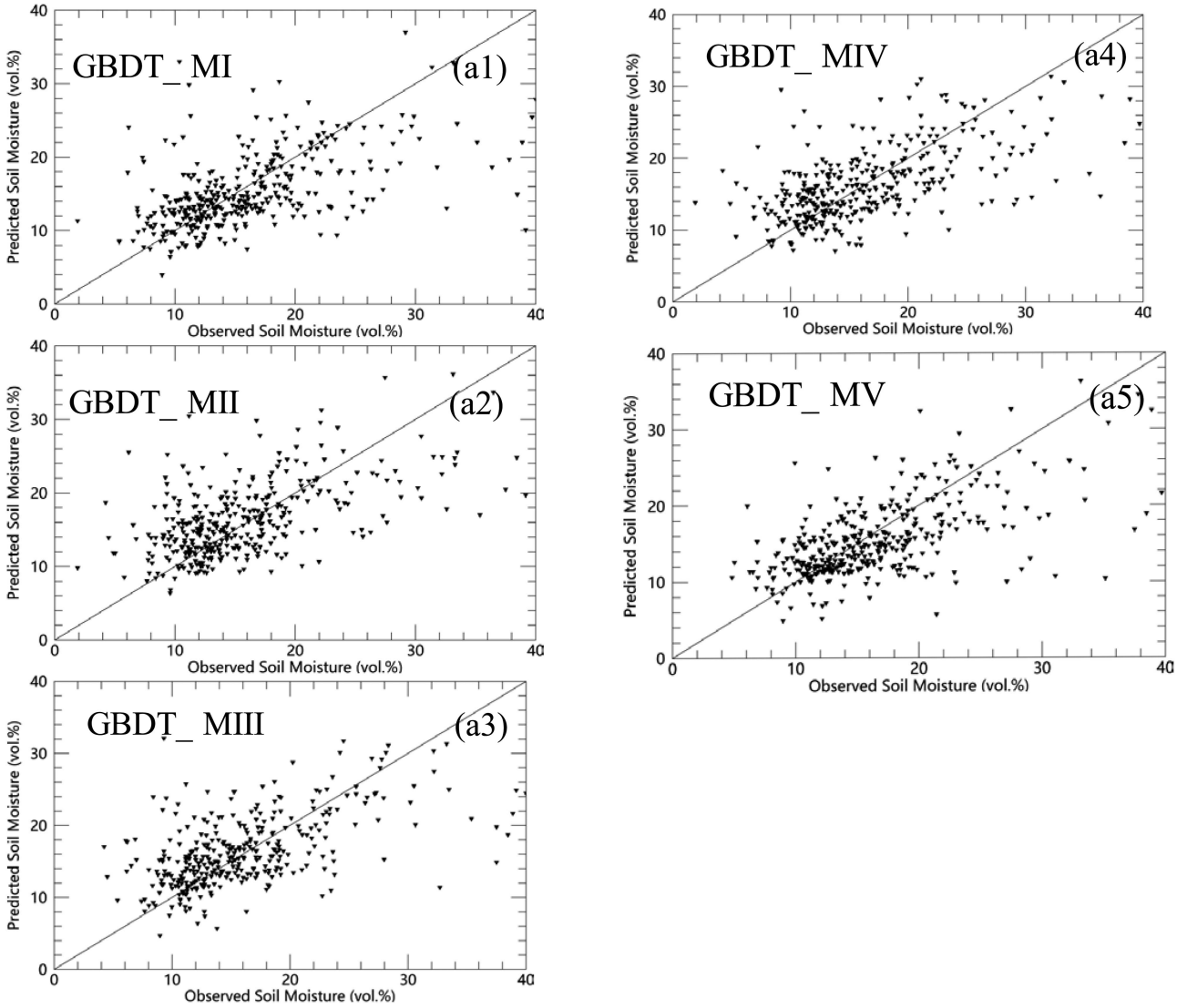

Fig. 3. Inversion results of GBDT. (a1) MI: S-1_VH; (a2) MII:S-1_VV; (a3) MIII: S-1_VV+VH; (a4) MIV:S-1_VV+VH+ SRD; (a5) MV:S-1_VV+VH+ SRD+DEM.

= modelcount, metric = "rmse," bagging_fraction = 0.8, feature_fraction = 0.8, reg_lambda = 0.9. The results indicated, MI achieved a lowest prediction accuracy with MAE is 2.86 vol.%, MRE is 0.19 vol.%, RMSE is 3.60 vol.%, and R is 0.48 (b1). When all the parameters (MV) are the input set, we get the optimal result with MAE is 2.59 vol.%, MRE is 0.17 vol.%, and RMSE is 3.32 vol.% and R = 0.56 (b5). When the input set MV is compared to MII (b2), MAE increases by 0.29 vol.%, MRE increases by 0.02 vol.%, and RMSE increases by 0.33 vol.%. When the input mode is MIII (b3) and MIV (b4), MAE, MRE, RMSE, R results are 2.77 vol.%, 0.18 vol.%, 3.59 vol.%,0.55 and 2.70 vol.%, 0.18 vol.%, 3.52 vol.%, 0.58, respectively.

Fig. 5 and Table III revealed the forecast results of XGBoost algorithm. First, we also confirm the parameter settings of the XGBoost structure with learning_rate = 0.1, n_estimators = 80, silent = True, objective = "reg:gamma". When the input characteristics is only VH polarization, MAE is 3.00 vol.%, MRE is 0.21 vol.%, RMSE is 4.09 vol.%, and R is 0.54 (see Fig. 5(c1), Table III). When the input features of is only VV polarization, MAE is 2.88 vol.%, MRE is 0.23 vol.%, RMSE is 4.02 vol.%, and R is 0.53 (see Fig. 5(c2), Table III). When the input dataset are VV and VH polarization, MAE is 2.95 vol.%, MRE is 0.21 vol.%, RMSE is 3.97 vol.%, and R is 0.56 (see Fig. 5(c3), Table III). While MIV mode, the MAE is 2.56 vol.%, MRE is 0.18 vol.%, and RMSE is 3.35 vol.% and R = 0.55(c4). When using the XGBoost algorithm, the best prediction results are still achieved in MV, where MAE = 2.50 vol.%, MRE = 0.17 vol.%, RMSE = 3.32 vol.%, and R = 0.61 (c5).

For the CatBoost algorithm, as depicted in Fig. 6 and Table III, which get the optimal results compared with the other three algorithms. Setting the structural parameters of the CatBoost regression model, while iterations = 50, depth = 8, learning_rate = 0.8, loss_function = "RMSE". Overall, we found that in five modes, using the CatBoost algorithm gives approximate results with highest prediction accuracy. The results of the five modes are summarized in Table III. The MV mode showed the

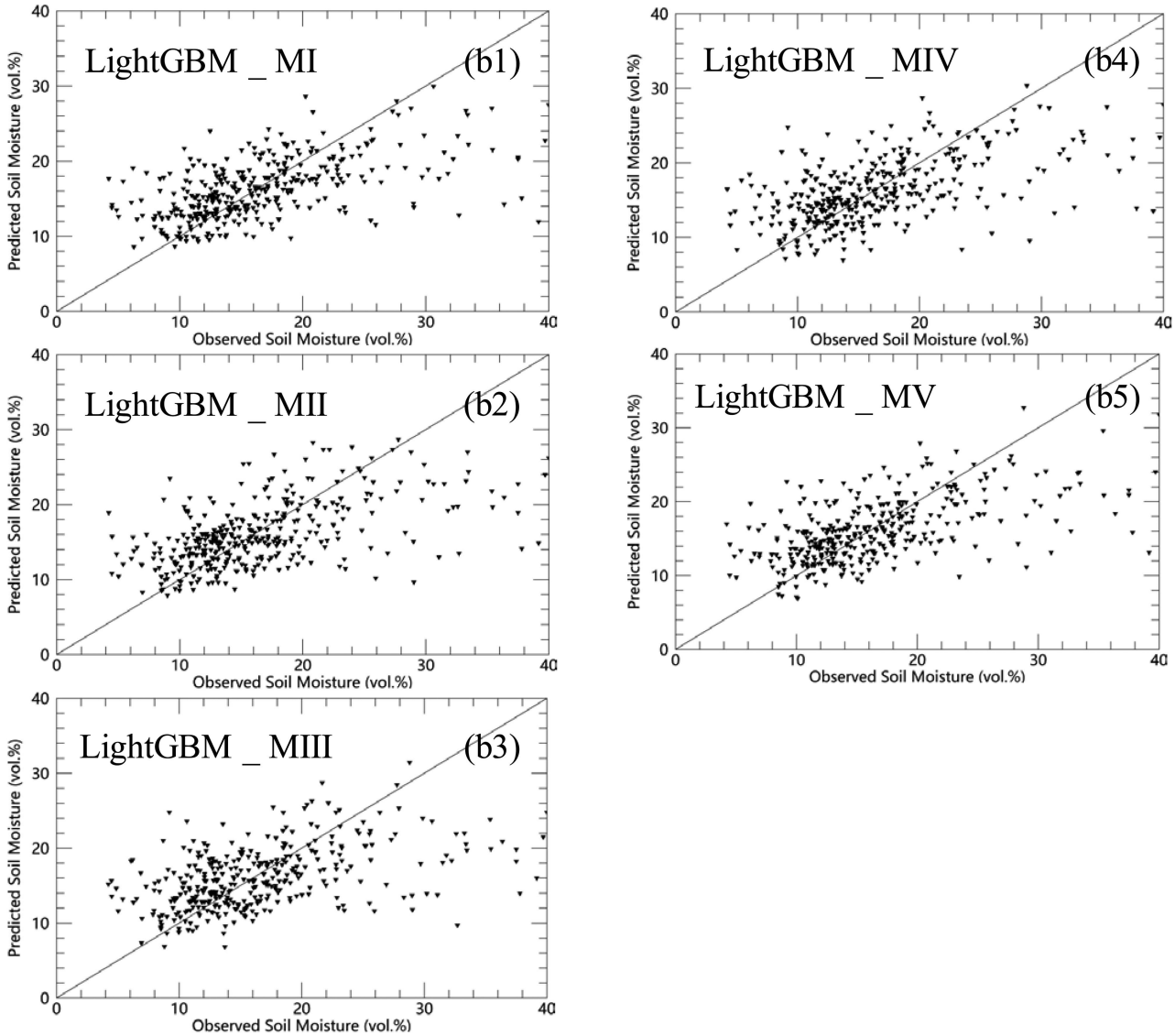

Fig. 4. Inversion results of LightGBM. (b1) MI: S-1_VH; (b2) MII:S-1_VV; (b3) MIII: S-1_VV+VH; (b4) MIV:S-1_VV+VH+ SRD; (b5) MV:S-1_VV+VH+ SRD+DEM.

highest estimation performance, with the lowest MAE, MRE, and RMSE values and a high R value [MAE = 2.40 vol.%, MRE = 0.16 vol.%, RMSE = 3.26 vol.%, and R = 0.73 (d5)]. The MIV mode was slightly worse than the MV mode, with a higher RMSE value but a slightly higher R value (d4). The other three modes had similar prediction results, especially when the input data set consisted of only radar signals (MI, MII, and MIII modes).

C. Inversion Results in Different SM Ranges

C-band radar signals have a limitation in high vegetation cover areas. This limitation stems from the low penetration depth of C-band signals into the canopy in areas with more developed vegetation cover. In densely vegetated areas, the radar signal may not reach the soil through the plant canopies, resulting in less accurate estimation of SM. Some studies have shown that, the C-band radar signal is insensitive to SM when NDVI > 0.7 [48], [49], [50].

In this study area, there are grassland, crops, and other vegetation. Fig. 7 shows the types of vegetation cover and the range of NDVI in the study area. In Section III-B, we analyzed the whole area without considering the effect of vegetation cover level on soil sensitivity. Therefore, in this section, we reanalyze the different sensitivity of C-band to soil when NDVI is in different range.

First, we divided the range of measured SM (2 to 40 vol.%), which we divided into two parts: dry to slightly moist (0–25 vol.%), based on experience; and very wet (25–40 vol.%) [48]. Second, we perform inversion analysis in different NDVI ranges using four methods.

According to the scope of SM and the scope of NDVI, we are divided into four cases, namely SM0-25 vol.%NDVI0.7-1, SM0-25 vol.% NDVI0-0.7, SM25-40 vol.% NDVI0.7-1, and

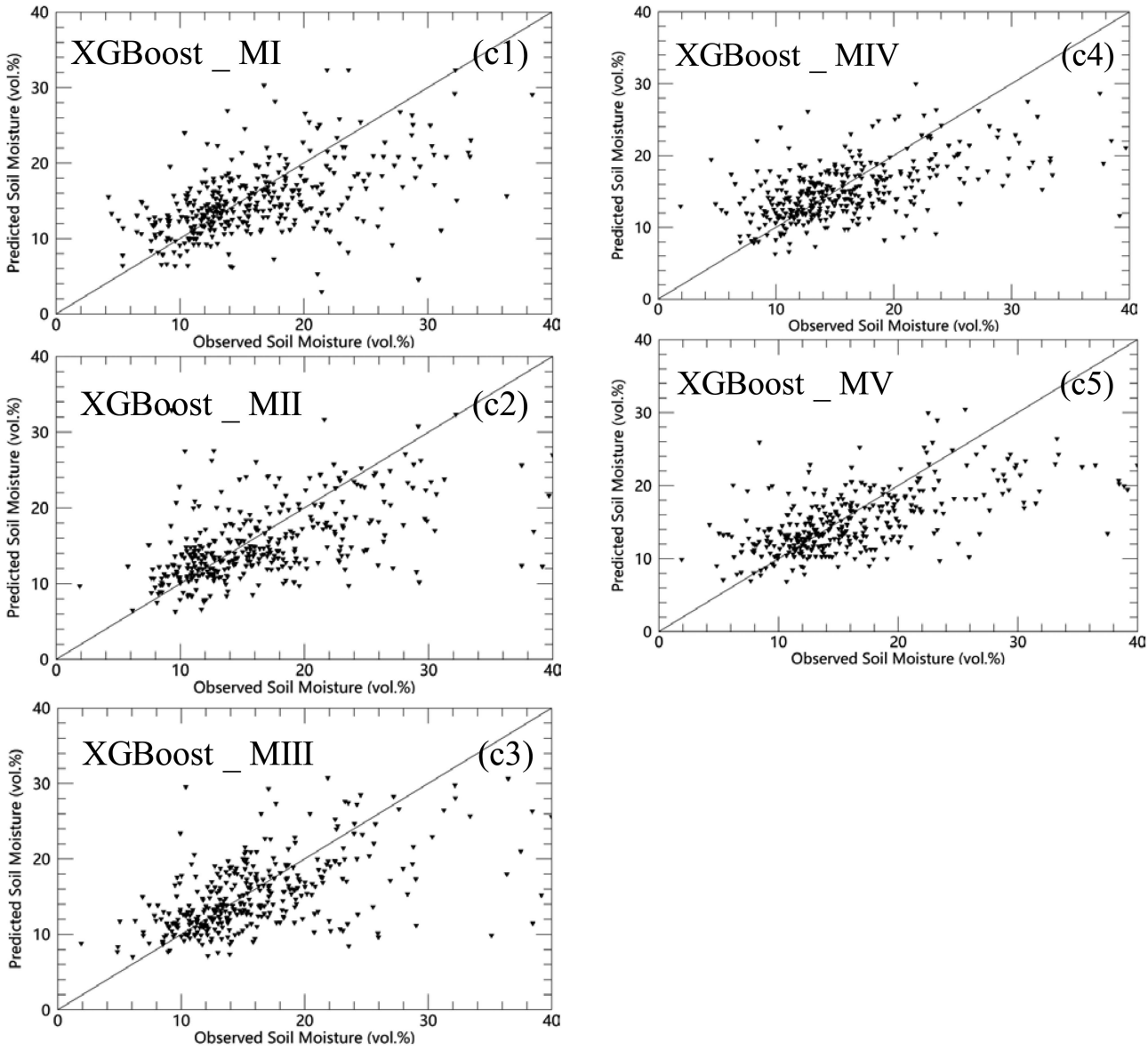

Fig. 5. Inversion results of XGBoost. (c1) MI: S-1_VH; (c2) MII:S-1_VV; (c3) MIII: S-1_VV+VH; (c4) MIV:S-1_VV+VH+ SRD; (c5) MV:S-1_VV+VH+ SRD+DEM.

SM25-40 vol.% NDVI0-0.7. At the same time, according to the method established in the second part, four methods were used to analyze the prediction results under five models, respectively. For the results, MAE, MRE, RMSE, and R were calculated, and the results were shown in Fig. 8 and Tables IV–VII.

From Fig. 8 and Tables IV–VII, we can see that CatBoost algorithm is obviously superior to the other three algorithms in terms of accuracy under any case and mode. In the case of SM0-25 vol.% NDVI0.7-1 (see Table IV), at the mode MI, we can find that the MAE, MRE, and RESE values of LightGBM algorithm are lowest (MAE = 0.72, MRE = 0.05, RMSE = 0.81), but the R value is also very low (R = 0.14). At this point, the CatBoost algorithm has the highest R with a value of 0.69. In mode MII, the GBDT algorithm has the highest MAE (3.71), MRE (0.28), RMSE (4.51), and lowest R(-0.14). LightGBM has the lowest MAE (0.6), MRE (0.04), RMSE(0.74), and R is 0.1. Under the pattern MIII, MIV, and MV, with the same rule, the LightGBM algorithm has lower MAE, MRE, and RMSE values, but its R value is also lower. The MAE, MRE, and RMSE values of CatBoost algorithm are slightly higher than LightGBM, but have higher R values (MIII:0.8, MIV:0.78, and MV:0.78).

In the case of SM25-40 vol% NDVI0.7-1 (see Table V), Within this range, we found that all four methods have this relatively good accuracy. At the mode MI, the MAE is 2.35 vol.%, MRE is 0.17 vol.%, RMSE is 3.06 vol.%, and R = 0.46 of GBDT. The MAE is 1.6 vol.%, MRE is 0.12 vol.%, RMSE is 2.19 vol.%, and R = 0.76 with the method of CatBoost. And MAE is 1.81 vol.%, MRE is 0.13 vol.%, RMSE is 2.24 vol.%, and R = 0.5 of LightGBM. MAE is 2.3 vol.%, MRE is 0.18 vol.%, RMSE is 2.93 vol.%, and R = 0.5 of XGBoost. At the mode

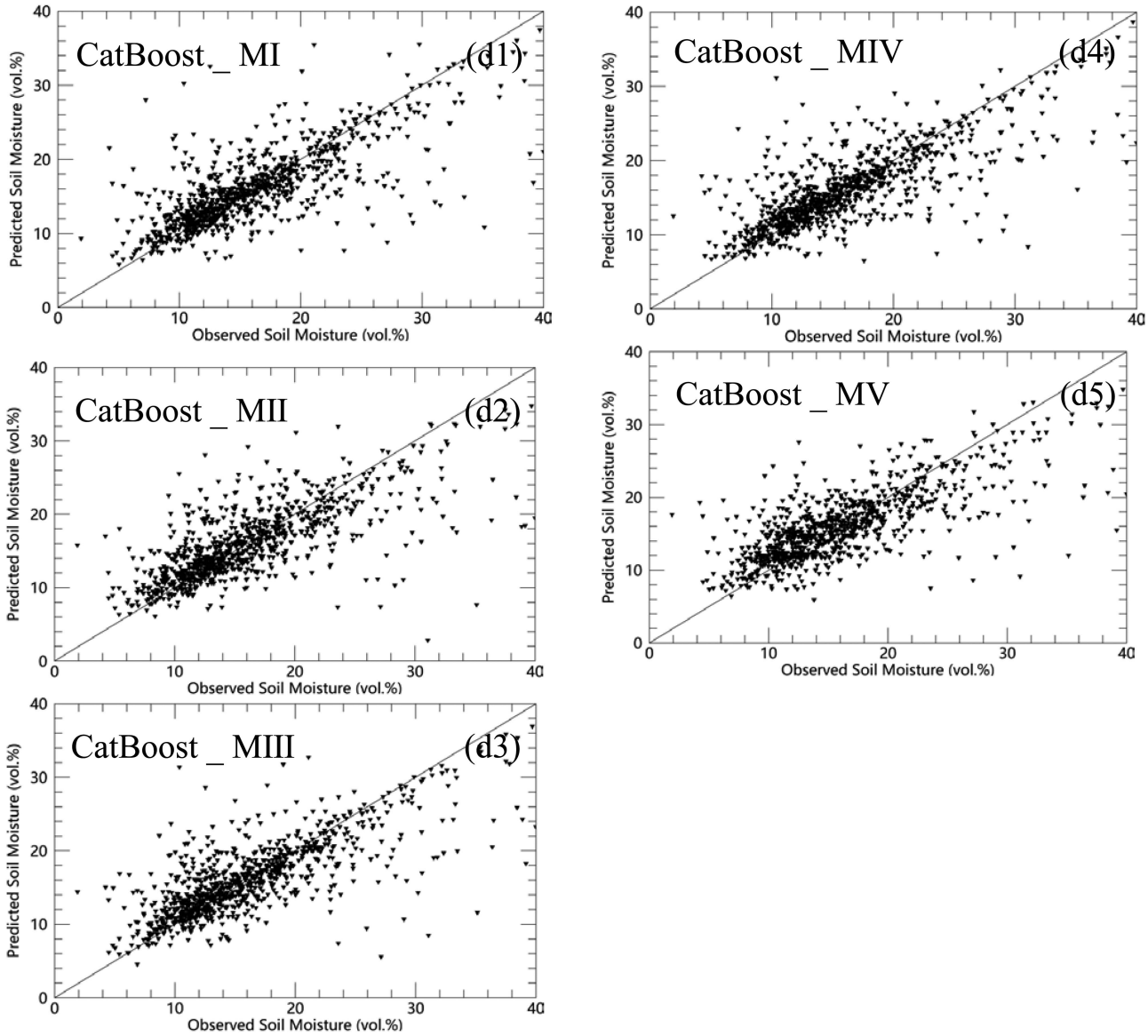

Fig. 6. Inversion results of CatBoost. (d1) MI: S-1_VH; (d2) MII:S-1_VV; (d3) MIII: S-1_VV+VH; (d4) MIV:S-1_VV+VH+ SRD; (d5) MV:S-1_VV+VH+ SRD+DEM.

MII, with the methods of GBDT, LightGBM, XGBoost, and CatBoost, the MAE is 1.67 vol.%, 2.00 vol.%, 2.47 vol.%,1.67 vol.%, the MRE is 0.17 vol.%, 0.14 vol.%, 0.2 vol.%, 0.12 vol.%, the RMSE is 2.11, 2.51, 3.12, 2.11, and the R is 0.66, 0.47, 0.46, 0.66. Under the MIII, the results of CatBoost is better than other algorithm, MAE is 1.62, MRE is 0.12 vol.%, RMSE is 2.16 vol.%, and R = 0.72. Under the MIV, and MV of CatBoost, the MAE, MRE, RMSE, and R is 1.55 vol.%, 0.12 vol.%, 2.13 vol.%, 0.8 and 1.52 vol.%, 0.12 vol.%, 2.11 vol.%, 0.81 severally.

Table VI shows the results of case SM25-40, NDVI0.7-1, we find that CatBoost has a highest R than the other three algorithms. In the five modes, the R value is 0.68, 0.76, 0.79, 0.76, and 0.68. Table VII shows the results of case SM25-40, NDVI0-0.7. In this case, the result is similar to MV25-40, NDVI0-0.7.

Through this part of analysis, we find that CatBoost algorithm can always get the best result in any case. When NDVI is not so sensitive to soil, the results are relatively poor, which indicates that in this range, the accuracy of soil water inversion results is poor, and other ways should be sought to overcome the influence of insensitivity to soil under high NDVI.

D. Inversion Results in Different Modes

Through Sections III-B and III-C, we found that using the CatBoost algorithm can achieve the optimal results compared with other three algorithms, and at the same time when MV ranges from 0 to 25 vol.% and NDVI ranges from 0 to 0.7, the predicted SM results are the most accurate and meaningful. When MV ranges from 0 to 25 vol.% and NDVI ranges from

TABLE III
PREDICTION PERFORMANCE OF ENSEMBLE LEARNING METHOD

Algorithm	Mode	MAE (vol.%)	MRE (vol.%)	RMSE (vol.%)	R
GBDT	MI	3.10	0.22	4.24	0.55
	MII	3.19	0.21	4.21	0.54
	MIII	3.14	0.21	4.09	0.56
	MIV	3.08	0.21	4.00	0.60
	MV	2.91	0.21	3.92	0.60
LightGBM	MI	2.86	0.19	3.60	0.48
	MII	2.85	0.19	3.65	0.55
	MIII	2.77	0.18	3.59	0.55
	MIV	2.70	0.18	3.52	0.58
	MV	2.59	0.17	3.32	0.56
XGBoost	MI	3.00	0.21	4.09	0.54
	MII	2.88	0.23	4.02	0.53
	MIII	2.95	0.21	3.97	0.56
	MIV	2.56	0.18	3.35	0.55
	MV	2.50	0.17	3.32	0.61
CatBoost	MI	2.45	0.16	3.56	0.75
	MII	2.53	0.18	3.52	0.72
	MIII	2.46	0.17	3.45	0.74
	MIV	2.40	0.16	3.39	0.75
	MV	2.40	0.16	3.26	0.73

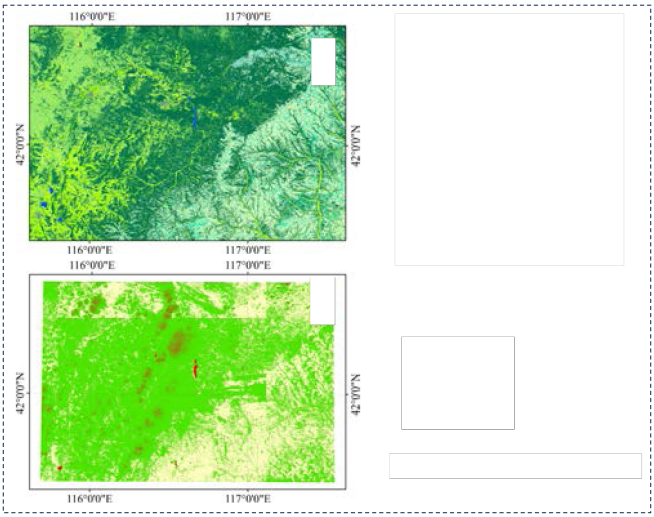

Fig. 7. Land cover map and NDVI range diagram.

0 to 0.7, we analyzed the SM inversion results using CatBoost algorithm under five input modes (mentioned in Section II-C2).

Fig. 9 and Table V shows the SM inversion results under different modes. Overall, under MV mode, the optimal result was still obtained. The MAE was 1.52 vol.%, the MRE was 0.12 vol.%, the RMSE was 2.11 vol.%, and the R was 0.81 compared with MI, MII, MIII, and MIV modes. Compared with the highest MAE under MII mode, the MAE of MV dropped by 0.08vol.%. Compared with the highest RMSE under MI mode,

TABLE IV
PREDICTION PERFORMANCE OF ENSEMBLE LEARNING METHOD WITH RANGE OF SM0-25 VOL.%, NDVI0.7-1

	Mode	MAE (vol.%)	MRE (vol.%)	RMSE (vol.%)	R
GBDT	MI	3.71	0.28	4.51	-0.14
	MII	4.06	0.28	4.79	0.22
	MIII	3.57	0.29	4.41	-0.11
	MIV	2.98	0.19	3.56	0.38
	MV	2.53	0.17	2.96	0.54
LightGBM	MI	0.72	0.05	0.81	0.14
	MII	0.6	0.04	0.74	0.1
	MIII	0.63	0.04	0.71	0.18
	MIV	1.99	0.13	2.35	0.38
	MV	2.04	0.13	2.45	0.46
XGBoost	MI	1.09	0.09	1.36	0.32
	MII	0.96	0.08	1.15	0.16
	MIII	3.28	0.22	3.6	-0.08
	MIV	2.98	0.19	3.56	0.44
	MV	2.51	0.18	3.4	0.25
CatBoost	MI	2.43	0.17	3.2	0.69
	MII	2.78	0.19	3.79	0.62
	MIII	1.9	0.12	2.74	0.8
	MIV	2.11	0.14	2.81	0.78
	MV	2.06	0.15	3.01	0.78

its value dropped by 0.08vol.%. The highest R increased by 0.15 compared with the lowest.

Fig. 10 shows the comparison error between predicted and real SM under five modes. The blue scatter chart displays the predicted SM values, the red line chart shows the real SM values, and the black line shows the error between the real and predicted values. Under MI mode, the error between the predicted SM values and the true SM values ranges from −4 vol.% to 4 vol.% [see Fig. 10(a)]. Fig. 9(b) shows that under MII mode, the error between the predicted SM value and the true SM value ranges from −3 vol.% to 4 vol.%. Under MIII mode, the error between the true values and the predicted values ranges from −3 vol.% to 6 vol.% [see Fig. 10(c)]. Fig. 9(d) shows that under MIV mode, the error between the predicted SM value and the true SM value ranges from −2 vol.% to 8 vol.%. Under MIV mode, the error between the true values and the predicted values ranges from −2 vol.% to 3 vol.% [see Fig. 10(e)].

IV. DISCUSSION

In this study, we demonstrated that CatBoost algorithm has great potential in soil moisture retrieval. Many studies have used radar polarization data, soil roughness data and DEM data to invert SM [19], [40]. DEM, CL, RMSH together with VV polarization and VH polarization, have a good correlation with SM and perform well in SM inversion. Soil roughness is widely used in SM inversion. However, it is not easy to obtain soil

Fig. 8. (a1)–(a4) MV0-25 NDVI0.7-1; (b1)–(b4) MV0-25NDVI0-0.7; (c1)–(c4) MV25-40NDVI0.7-1; (d1)–(d4) MV25-40NDVI0-0.7. Where 1 is GBDT, 2 is LightGBM, 3 is XGBoost, 4 is CatBoost.

roughness data in areas with vegetation. Moreover, DEM data also showed a strong correlation with SM in a Pearson correlation analysis. The study area we selected is located in the upper Luanhe River area. The terrain of this area is complex, with high elevations in the northwest and low elevations in the southeast; the altitude ranges from 750 to 1829 m. We designed different input data combination modes (MI: S-1_VH; MII:S-1_VV; MIII: S-1_VV+VH; MIV:S-1_VV+VH+SRD; MV:S-1_VV+VH+SRD+DEM) and obtained the following conclusion. By adding soil roughness data and DEM data (MV), we can achieve higher accuracy in SM inversion. Hence, the DEM and soil roughness factors influence the accuracy of SM inversion. These factors are very important characteristic parameters in model inversion.

NDVI is one of the more important parameters in soil moisture inversion, especially in vegetation cover areas. This is because the vegetation cover affects the signal of the soil. Also NDVI is used as an important parameter when removing the influence of vegetation signal using water cloud model, which reduces the influence of vegetation cover on soil signal [41]. The WCM presents a distinct advantage in the context of areas characterized by low vegetation coverage. As a result, its application for the purpose of vegetation removal is particularly advantageous in study areas where vegetation is predominantly low in height. This approach provides an effective means to mitigate the potential impact of vegetation on soil signal detection, thereby allowing for more accurate analyses of the underlying data.

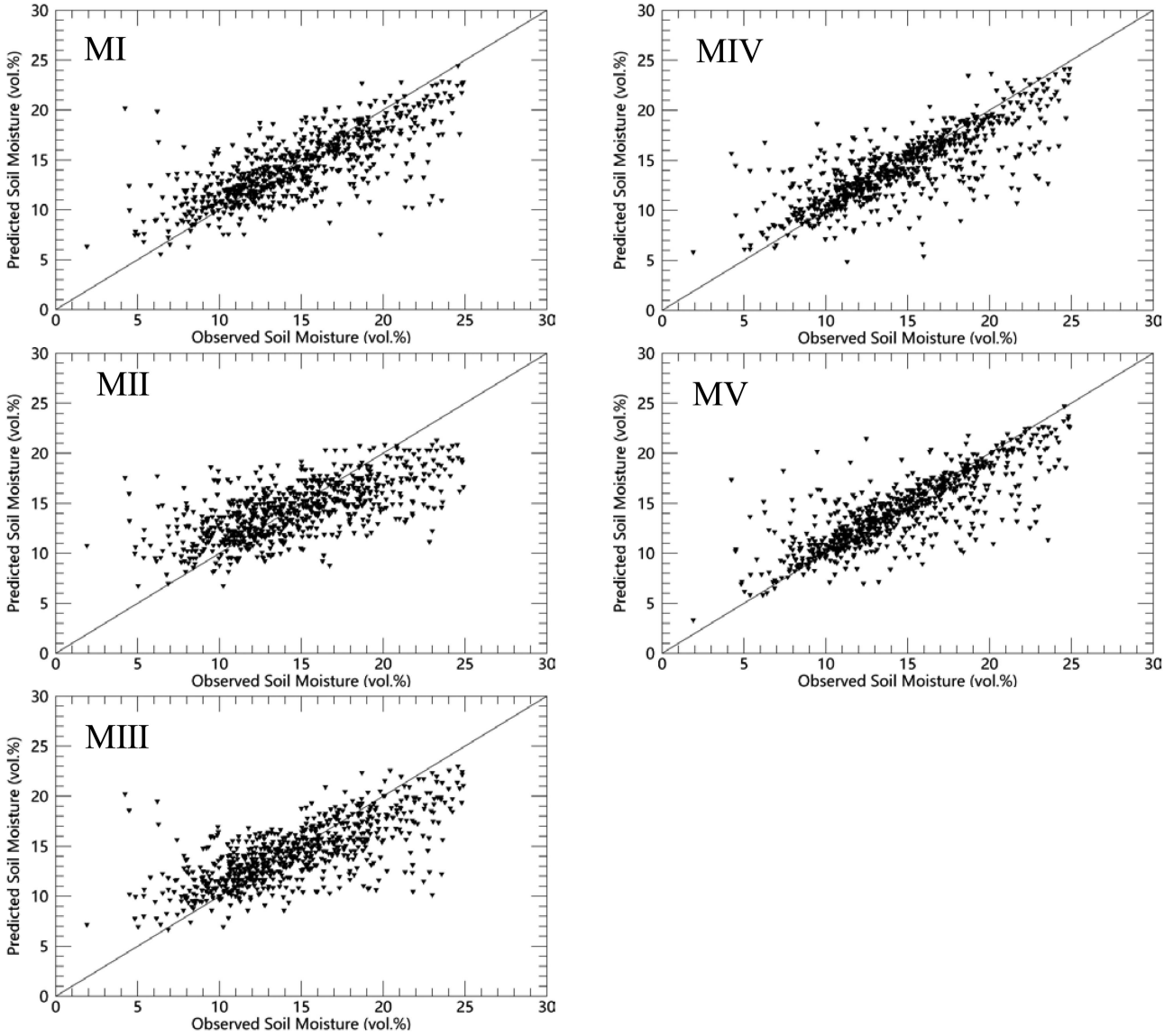

Fig. 9. Inversion results of CatBoost in different modes.

Studies by many scholars have shown that radar signals VV and VH together give better soil moisture prediction results than VV or VH signals alone [42] .In particular, DEM can be used as reliable data for inversion of soil moisture when topography has some control on soil moisture movement [43], [44]. Soil roughness is also an important parameter for soil moisture inversion, but the actual measurement of soil roughness is a difficult task [45], [46]. Especially at the time that has passed, we cannot get the soil roughness value at that time. In this study, we found that the CatBoost algorithm can get the optimal soil moisture inversion results, and also we used different parameter combinations and found that we got the optimal results when there is soil roughness, but the CatBoost algorithm can still get good soil moisture prediction results when there is no soil roughness. This brings more possibilities to our research. Due to the complexity of the soil moisture inversion process and the many parameters, we often cannot get all the parameters for soil moisture prediction. This is where the selection of the appropriate method and parameters is particularly important.

Our study highlights the importance of selecting appropriate methods and parameters in soil moisture predictions, given the complexity of soil moisture inversion processes. By dividing our dataset into different combinations for analysis, this study provides insights into the method's versatility in various data availability situations that aid in soil moisture inversion.

Simultaneously, it is crucial to acknowledge that different algorithms possess inherent strengths and weaknesses of their own. The GBDT algorithm exhibits the capability to flexibly handle various types of data and requires less tuning time. However, an obstacle associated with this algorithm is its difficulty in parallel training. The XGBoost algorithm employs multiple strategies to mitigate the risk of overfitting. Nevertheless, this algorithm is accompanied by a multitude of parameters, rendering the tuning process relatively intricate. The LightGBM

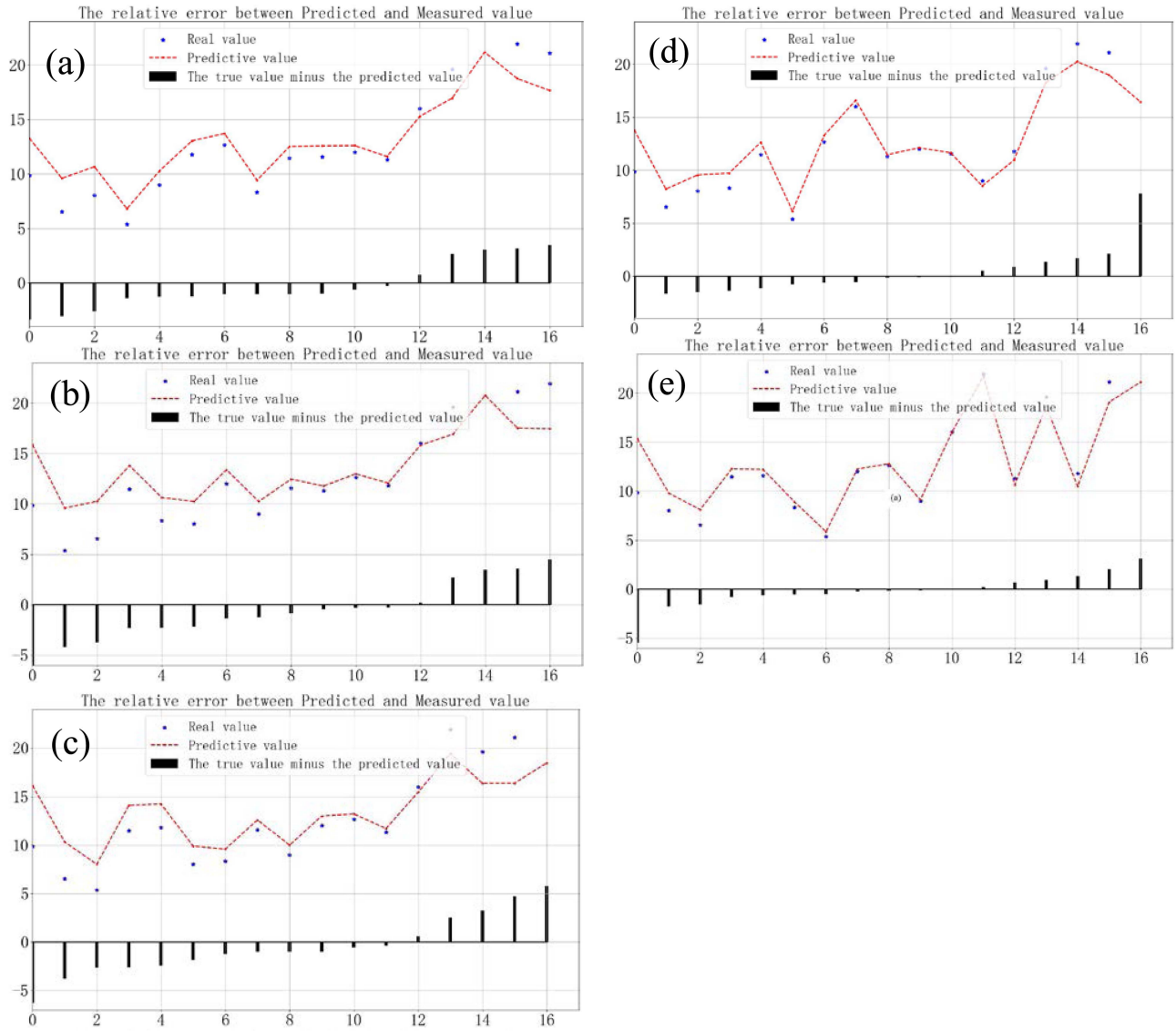

Fig. 10. Relative error between Predicted and Measured value of CatBoost in different modes under SM: 0–25 vol.%, NDVI:0-0.7. (The horizontal coordinate represents the number. Due to the large amount of data, we randomly selected 16 groups of data for display and analysis).

algorithm is distinguished by its rapid training speed and low memory consumption. Nevertheless, it also presents a challenge due to the excessive number of parameters and the intricacy of parameter adjustment. However, the CatBoost algorithm exhibits remarkable robustness, thereby minimizing the requirement for extensive hyperparameter tuning and reducing the risk of overfitting. This characteristic enhances the versatility of the model, enabling it to effectively handle diverse types of data. It is plausible that this attribute contributes to the superior outcomes observed in this study's soil water inversion using the CatBoost algorithm.

Although there are important discoveries revealed by these studies, there are also limitations. First, the use of NDVI in the water cloud model aims to eliminate the signal interference caused by vegetation. However, it is necessary to validate whether applying different indices in the water cloud model may have varying effects, depending on the specific study area. This warrants further investigation and verification to ensure accurate results. Second, we evaluated the efficacy of the CatBoost algorithm for estimating soil water content using various input datasets and different ranges of NDVI and SM. The selected region for our study is relatively small and we were initially unaware of the potential for the CatBoost algorithm to achieve high levels of accuracy in large regions. This represents a limitation of our study and highlights the importance of thoroughly investigating suitable algorithms to ensure optimal results. In summary, the CatBoost algorithm has demonstrated exceptional accuracy in predicting soil water content through rigorous analysis and verification. This breakthrough carries significant implications for future research. Nonetheless, due to some identified limitations, further exploration and implementation of CatBoost must be conducted to expand its spectrum of applications. Simultaneously, it is imperative to compare various algorithms to validate their suitability in diverse research domains and data scenarios. This process aims to attain more optimal outcomes.

TABLE V
PREDICTION PERFORMANCE OF ENSEMBLE LEARNING METHOD WITH RANGE OF SM 0–25 VOL.%, NDVI0–0.7

	Mode	MAE (vol.%)	MRE (vol.%)	RMSE (vol.%)	R
GBDT	MI	2.35	0.17	3.06	0.46
	MII	1.67	0.12	2.11	0.66
	MIII	2.36	0.17	2.98	0.49
	MIV	2.23	0.17	2.85	0.52
	MV	2.27	0.17	2.92	0.51
LightGBM	MI	1.81	0.13	2.24	0.5
	MII	2	0.14	2.51	0.47
	MIII	1.89	0.14	2.39	0.41
	MIV	2.06	0.15	2.58	0.48
	MV	2.13	0.16	2.66	0.5
XGBoost	MI	2.3	0.18	2.93	0.5
	MII	2.47	0.2	3.12	0.46
	MIII	2.08	0.16	2.61	0.45
	MIV	2.21	0.17	2.83	0.5
	MV	2.08	0.17	2.75	0.51
CatBoost	MI	1.6	0.12	2.19	0.76
	MII	1.67	0.12	2.11	0.66
	MIII	1.62	0.12	2.16	0.72
	MIV	1.55	0.12	2.13	0.8
	MV	1.52	0.12	2.11	0.81

TABLE VI
PREDICTION PERFORMANCE OF ENSEMBLE LEARNING METHOD WITH RANGE OF SM 25–40 VOL.%, NDVI0.7-1

	Mode	MAE (vol.%)	MRE (vol.%)	RMSE (vol.%)	R
GBDT	MI	1.66	0.05	2.02	-0.29
	MII	3.1	0.1	3.68	0.4
	MIII	2.02	0.07	2.82	-0.07
	MIV	1.1	0.04	1.29	-0.32
	MV	3.74	0.12	4.32	-0.03
LightGBM	**MI**	6.49	0.18	7.04	0
	MII	6.49	0.18	7.04	0
	MIII	6.49	0.18	7.04	0
	MIV	6.49	0.18	7.04	0
	MV	6.49	0.18	7.04	0
XGBoost	**MI**	2.19	0.07	2.46	-0.21
	MII	0.95	0.03	1.27	0.45
	MIII	2.74	0.09	3.12	-0.49
	MIV	2.37	0.08	2.56	-0.16
	MV	0.91	0.03	1.2	0.05
CatBoost	**MI**	2.04	0.07	2.68	0.68
	MII	1.76	0.06	2.32	0.76
	MIII	1.7	0.05	2.21	0.79
	MIV	1.76	0.06	2.29	0.76
	MV	2.01	0.06	2.66	0.68

TABLE VII
PREDICTION PERFORMANCE OF ENSEMBLE LEARNING METHOD WITH RANGE OF SM 25–40 VOL.%, NDVI0-0.7

	Mode	MAE (vol.%)	MRE (vol.%)	RMSE (vol.%)	R
GBDT	MI	3.52	0.11	4.3	0.04
	MII	2.17	0.07	2.89	0.57
	MIII	1.8	0.06	2.11	-0.08
	MIV	1.87	0.06	2.36	0.12
	MV	2.93	0.09	3.48	0.23
LightGBM	MI	3.7	0.13	4.26	0
	MII	3.7	0.13	4.26	0
	MIII	3.7	0.13	4.26	0
	MIV	3.7	0.13	4.26	0
	MV	3.7	0.13	4.26	0
XGBoost	MI	1.6	0.06	1.98	0.46
	MII	1.5	0.05	1.96	0.26
	MIII	1.64	0.06	1.95	0.09
	MIV	2.18	0.07	2.52	0.31
	MV	2.16	0.07	2.52	0.33
CatBoost	MI	1.73	0.06	2.72	0.75
	MII	1.47	0.05	2.1	0.81
	MIII	1.42	0.05	2.21	0.8
	MIV	1.36	0.05	2.14	0.84
	MV	1.43	0.05	2.21	0.82

V. CONCLUSION

In this study, an ensemble learning approach is used to predict SM using the five modes (MI: S-1_VH; MII:S-1_VV; MIII: S-1_VV+VH; MIV:S-1_VV+VH+SRD; MV:S-1_VV+VH+SRD+DEM) from Sentinel-1and other remote sensing images. Ensemble learning is a meta-learning method that combines multiple base learners into a single strong learner to improve the predictive performance of a machine learning task. Boosting is one of the types of ensemble learning, and in this article, we choose four boosting algorithms. The purpose of this study is to establish a SM inversion algorithm based on Sentinel-1 and multisource data by using the four boosting algorithms of ensemble learning, and to explore their applicability for SM inversion. We use Sentinel-1, Sentinel-2, and other multisource data to build training and validation sets for boosting algorithms. We can come to the following conclusion.

1) We use Pearson correlation analysis to select the feature factors for inversion. The results show that VV, VH, RMSH, CL, and DEM are significantly correlated with SM. Therefore, we can select these features as inputs for the boosting algorithm to invert SM.
2) All four boosting ensemble learning algorithms can invert SM, but CatBoost algorithm achieves higher accuracy than GBDT, XGBoost, and LightGBM. CatBoost is a gradient boosting algorithm that handles categorical features well and reduces overfitting by using ordered

boosting. GBDT, XGBoost, and LightGBM are also gradient boosting algorithms that use different tree structures and regularization methods.

3) We designed five different combinations of feature parameters to invert SM (MI: S-1_VH; MII:S-1_VV; MIII: S-1_VV+VH; MIV:S-1_VV+VH+SRD; MV:S-1_VV+VH+ SRD+DEM). We found that the best inversion results were obtained by using all features as algorithm inputs, with MAE = 2.40 vol.%, MRE = 0.16 vol.%, RMSE = 3.26 vol.%, and R = 0.73.
4) We also analyzed the SM inversion results under different SM and NDVI conditions. We found that higher accuracy results can be obtained when MV ranges from 0 to 25 vol.% and NDVI ranges from 0 to 0.7. The reason is that C-band radar signals are more sensitive to SM in this range.
5) We analyzed the inversion results of CatBoost algorithm in five modes when MV ranges from 0 to 25 vol.% and NDVI ranges from 0 to 0.7. We found that the best results were still obtained by M5 mode, with MAE = 1.52 vol.%, MRE = 0.12 vol.%, RMSE = 2.11 vol.%, and R = 0.81.

This study demonstrates the efficacy of a boosting algorithm utilizing C-band radar data in effectively inverting SM and achieving improved accuracy. However, it is important to note that the study only utilized September data from a single year and focused on a single study area. Future investigations will involve expanding the research to encompass multiple months and study areas, thereby enhancing the applicability of the boosting ensemble learning algorithm. Moreover, future research endeavors will incorporate additional parameters that exhibit sensitivity to soil moisture, aiming to attain even higher accuracy in SM inversion results.

ACKNOWLEDGMENT

The data set is provided by National Tibetan Plateau Data Center (http://data.tpdc.ac.cn).

REFERENCES

[1] A. Al-Yaari et al., "Assessment and inter-comparison of recently developed/reprocessed microwave satellite soil moisture products using ISMN ground-based measurements," *Remote Sens. Environ.*, vol. 224, pp. 289–303, 2019, doi: 10.1016/j.rse.2019.02.008.
[2] S. K. Chaudhary, D. K. Gupta, P. K. Srivastava, D. K. Pandey, A. K. Das, and R. Prasad, "Evaluation of radar/optical based vegetation descriptors in water cloud model for soil moisture retrieval," *IEEE Sens. J.*, vol. 21, no. 18, pp. 21030–21037, Sep. 2021, doi: 10.1109/JSEN.2021.3099937.
[3] H. Wang et al., "Soil moisture retrieval over a site of intensive agricultural production using airborne radiometer data," *Int. J. Appl. Earth Observation Geoinformation*, vol. 97, 2021, Art. no. 102287, doi: 10.1016/j.jag.2020.102287.
[4] H. Wang, R. Magagi, and K. Goita, "Comparison of different polarimetric decompositions for soil moisture retrieval over vegetation covered agricultural area," *Remote Sens. Environ.*, vol. 199, pp. 120–136, 2017, doi: 10.1016/j.rse.2017.07.008.
[5] A. Amazirh et al., "Retrieving surface soil moisture at high spatio-temporal resolution from a synergy between sentinel-1 radar and landsat thermal data: A study case over bare soil," *Remote Sens. Environ.*, vol. 211, pp. 321–337, 2018, doi: 10.1016/j.rse.2018.04.013.
[6] X. Shen et al., "Soil moisture retrieval depth of P- and L-band radiometry: Predictions and observations," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6814–6822, Aug. 2021, doi: 10.1109/TGRS.2020.3026384.
[7] J. Zhao, C. Zhang, L. Min, Z. Guo, and N. Li, "Retrieval of farmland surface soil moisture based on feature optimization and machine learning," *Remote Sens.-Basel*, vol. 14, no. 20, 2022, Art. no. 5102.
[8] M. El Hajj, N. Baghdadi, M. Zribi, and H. Bazzi, "Synergic use of sentinel-1 and sentinel-2 images for operational soil moisture mapping at high spatial resolution over agricultural areas," *Remote Sens.*, vol. 9, no. 12, 2017, Art. no. 1292, doi: 10.3390/rs9121292.
[9] M. El Hajj, N. Baghdadi, and M. Zribi, "Comparative analysis of the accuracy of surface soil moisture estimation from the C- and L-bands," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 82, 2019, Art. no. 101888, doi: 10.1016/j.jag.2019.05.021.
[10] E. P. W. Attema and F. T. Ulaby, "Vegetation modeled as a water cloud," *Radio Sci.*, vol. 13, no. 2, pp. 357–364, Mar./Apr. 1978.
[11] S. H. Alemohammad, J. Kolassa, C. Prigent, F. Aires, and P. Gentine, "Global downscaling of remotely-sensed soil moisture using neural networks," *Hydrol. Earth Syst. Sci.*, vol. 22, no. 10, pp. 5341–5356, 2018, doi: 10.5194/hess-22-5341-2018.
[12] G. Ayehu, T. Tadesse, B. Gessesse, Y. Yigrem, and A. M. Melesse, "Combined use of sentinel-1 SAR and landsat sensors products for residual soil moisture retrieval over agricultural fields in the upper blue Nile Basin, Ethiopia," *Sensors*, vol. 20, no. 11, 2020, Art. no. 3282, doi: 10.3390/s20113282.
[13] L. Zhang, X. Lv, Q. Chen, G. Sun, and J. Yao, "Estimation of surface soil moisture during corn growth stage from SAR and optical data using a combined scattering model," *Remote Sens.*, vol. 12, no. 11, 2020, Art. no. 1844, doi: 10.3390/rs12111844.
[14] F. Greifeneder, C. Notarnicola, and W. Wagner, "A machine learning-based approach for surface soil moisture estimations with google earth engine," *Remote Sens.*, no. 13, 2021, doi: 10.3390/rs13112099.
[15] R. Hansch, T. Jagdhuber, and B. Fersch, "Soil-permittivity estimation under grassland using machine-learning and polarimetric decomposition techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 2877–2887, Apr. 2021, doi: 10.1109/TGRS.2020.3010104.
[16] V. P. Yadav, R. Prasad, R. Bala, and A. K. Vishwakarma, "An improved inversion algorithm for spatio-temporal retrieval of soil moisture through modified water cloud model using C-band sentinel-1A SAR data," *Comput. Electron. Agriculture*, vol. 173, 2020, Art. no. 105447, doi: 10.1016/j.compag.2020.105447.
[17] Z. Yang, J. Zhao, J. Liu, Y. Wen, and Y. Wang, "Soil moisture retrieval using microwave remote sensing data and a deep belief network in the Naqu region of the Tibetan plateau," *Sustainability*, vol. 13, 2021, Art. no. 12635.
[18] Y. Gao, M. Gao, L. Wang, and O. Rozenstein, "Soil moisture retrieval over a vegetation-covered area using ALOS-2 L-band synthetic aperture radar data," *Remote Sens.-Basel*, vol. 13, no. 19, 2021, Art. no. 3894, doi: 10.3390/rs13193894.
[19] H. Cui et al., "The potential of ALOS-2 and sentinel-1 radar data for soil moisture retrieval with high spatial resolution over agroforestry areas, China," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 4402617, doi: 10.1109/TGRS.2021.3082805.
[20] N. Jarray, A.B. Abbes, and I. R. Farah, "A novel teacher–student framework for soil moisture retrieval by combining sentinel-1 and sentinel-2: Application in arid regions," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Apr. 2022, Art. no. 3006705, doi: 10.1109/LGRS.2022.3168982.
[21] L. Wang and Y. Gao, "Soil moisture retrieval from sentinel-1 and sentinel-2 data using ensemble learning over vegetated fields," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1802–1814, Feb. 2023, doi: 10.1109/JSTARS.2023.3242264.
[22] Y. Zhang, S. Liang, Z. Zhu, H. Ma, and T. He, "Soil moisture content retrieval from landsat 8 data using ensemble learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 185, pp. 32–47, 2022, doi: 10.1016/j.isprsjprs.2022.01.005.
[23] M. Jamei, M. Karbasi, A. Malik, M. Jamei, O. Kisi, and Z. M. Yaseen, "Long-term multi-step ahead forecasting of root zone soil moisture in different climates: Novel ensemble-based complementary data-intelligent paradigms," *Agricultural Water Manage.*, vol. 269, 2022, Art. no. 107679, doi: 10.1016/j.agwat.2022.107679.
[24] Z. Yang, Q. He, S. Miao, F. Wei, and M. Yu, "Surface soil moisture retrieval of China using multi-source data and ensemble learning," *Remote Sens.*, vol. 15, 2023, Art. no. 2786.
[25] Y. Liu et al., "Downscaling satellite retrieved soil moisture using regression tree-based machine learning algorithms over southwest France," *Earth Space Sci.*, vol. 7, no. 10, 2020, Art. no. 250120, doi: 10.1029/2020EA001267.

[26] T. Zhao et al., "Soil moisture experiment in the Luan River supporting new satellite mission opportunities," *Remote Sens. Environ.*, vol. 240, 2020, Art. no. 111680, doi: 10.1016/j.rse.2020.111680.
[27] G. Peng, "Synchronous observation data set of soil surface roughness in the upstream of Luan River," in *Book Synchronous observation Data Set of Soil Surface Roughness in the Upstream of Luan River (2018), Series Synchronous Observation Data Set of Soil Surface Roughness in the Upstream of Luan River (2018)*. Tibetan Plateau, China: National Tibetan Plateau Data Center, 2021.
[28] G. L. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, pp. 1–9.
[29] C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, "An up-to-date comparison of stateof-the-art classification algorithms," *Expert Syst. Appl.*, vol. 82, pp. 128–150, 2017.
[30] G. Ke, Z. Xu, J. Zhang, J. Bian, and T. Liu, "DeepGBM: A deep learning framework distilled by GBDT for online prediction tasks," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 384–394.
[31] A. V. Dorogush and V. Ershov, "CatBoost: Gradient boosting with categorical features support," 2018, *arXiv:1810.11363*.
[32] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Advances in neural information processing systems," 31, 2018, *arXiv:1706.09516*.
[33] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: An interdisciplinary review," *J. Big Data*, vol. 7, no. 1, 2020, Art. no. 94, doi: 10.1186/s40537-020-00369-8.
[34] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 785–794.
[35] W. Liang, S. Luo, G. Zhao, and H. Wu, "Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms," *Mathematics*, vol. 8, no. 5, 2020, Art. no. 765, doi: 10.3390/math8050765.
[36] Q. Meng, "LightGBM: A highly efficient gradient boosting," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, pp. 1–9.
[37] M. Zribi et al., "Soil moisture mapping in a semiarid region, based on ASAR/wide swath satellite data," *Water Resour. Res.*, vol. 50, no. 2, pp. 823–835, 2014, doi: 10.1002/2012WR013405.
[38] M. El Hajj et al., "Soil moisture retrieval over irrigated grassland using X-band SAR data," *Remote Sens. Environ.*, vol. 176, pp. 202–218, 2016, doi: 10.1016/j.rse.2016.01.027.
[39] N. Baghdadi, M. El Hajj, M. Zribi, and S. Bousbih, "Calibration of the water cloud model at C-band for winter crop fields and grasslands," *Remote Sens.*, vol. 9, no. 9, 2017, Art. no. 969, doi: 10.3390/rs9090969.
[40] Y. Bao, L. Lin, S. Wu, K. A. Kwal Deng, and G. P. Petropoulos, "Surface soil moisture retrievals over partially vegetated areas from the synergy of Sentinel-1 and Landsat 8 data using a modified water-cloud model," *Int. J. Appl. Earth Observation Geoinformation*, vol. 72, pp. 76–85, 2018, doi: 10.1016/j.jag.2018.05.026.
[41] S. Bousbih et al., "Soil moisture and irrigation mapping in a semi-arid region, based on the synergetic use of sentinel-1 and sentinel-2 data," *Remote Sens.*, vol. 10, no. 12, 2018, Art. no. 1953, doi: 10.3390/rs10121953.
[42] A. Sekertekin, A. M. Marangoz, and S. Abdikan, "ALOS-2 and sentinel-1 SAR data sensitivity analysis to surface soil moisture over bare and vegetated agricultural fields," *Comput. Electron. Agriculture*, vol. 171, 2020, Art. no. 105303, doi: 10.1016/j.compag.2020.105303.
[43] Y. Liu, J. Qian, and H. Yue, "Combined sentinel-1A with sentinel-2A to estimate soil moisture in farmland," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1292–1310, Dec. 2020, doi: 10.1109/JSTARS.2020.3043628.
[44] H. R. Mirsoleimani, M. R. Sahebi, N. Baghdadi, and M. El Hajj, "Bare soil surface moisture retrieval from sentinel-1 SAR data based on the calibrated IEM and Dubois models using neural networks," *Sensors-Basel*, vol. 19, no. 14, p. 3209, 2019, doi: 10.3390/s19143209.
[45] M. El Hajj, N. Baghdadi, and M. Zribi, "Comparative analysis of the accuracy of surface soil moisture estimation from the C- and L-bands," *Int. J. Appl. Earth Observation Geoinformation*, vol. 82, 2019, Art. no. 101888, doi: 10.1016/j.jag.2019.05.021.
[46] E. H. Hegazi, L. Yang, and J. Huang, "A convolutional neural network algorithm for soil moisture prediction from sentinel-1 SAR images," *Remote Sens.*, vol. 13, no. 24, 2021, Art. no. 4964, doi: 10.3390/rs13244964.
[47] T. Zhou, Y. Geng, J. Chen, J. Pan, D. Haase, and A. Lausch, "High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, sentinel-1 and sentinel-2 data based on machine learning algorithms," *Sci. Total Environ.*, vol. 729, 2020, Art. no. 138244, doi: 10.1016/j.scitotenv.2020.138244.
[48] Z. Zhao, Q. Yang, X. Ding, and Z. Xing, "Model prediction of the soil moisture regime and soil nutrient regime based on DEM-derived topo-hydrologic variables for mapping ecosites," *Land*, vol. 10, no. 5, 2021, Art. no. 449, doi: 10.3390/land10050449.
[49] Z. Yang, J. Zhao, J. Liu, Y. Wen, and Y. Wang, "Soil moisture retrieval using microwave remote sensing data and a deep belief network in the Naqu region of the Tibetan plateau," *Sustainability*, vol. 13, 2021, Art. no. 12635.
[50] L. Dong, W. Wang, R. Jin, F. Xu, and Y. Zhang, "Surface soil moisture retrieval on Qinghai–Tibetan plateau using sentinel-1 synthetic aperture radar data and machine learning algorithms," *Remote Sens.*, vol. 15, 2023, Art. no. 153.

Ya Gao is currently working toward the Ph.D. degree in research on soil moisture inversion based on machine learning and multi-source remote sensing with Harbin Engineering University, Harbin, China.

Her research interests include agricultural drought monitoring and early warning, remote sensing image processing, and radar soil moisture inversion, machine learning, and neural networks.

Liguo Wang received the M.A. and Ph.D. degrees in signal and information processing from Harbin Institute of Technology, Harbin, China, in 2002 and 2005, respectively.

From 2006 to 2008, He was Postdoctoral Researcher with the College of Information and Communications Engineering, Harbin Engineering University. He is currently a Professor with the College of Information and Communications Engineering, Dalian Minzu University, Dalian, China, and a Doctor Supervisor with the College of Information and Communications Engineering, Harbin Engineering University, Harbin, China. His research interests include remote sensing image processing and machine learning.

Geji Zhong received the bachelor's degree in science from the Luoyang normal university, Luoyang, China, in 2014, and the master's degree in management from the School of Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing, China, in 2019. She is currently working toward the Ph.D. degree in remote sensing resource utilization with the Minzu University of China, Beijing, China.

His research interests include spatial analysis, agricultural remote sensingagricultural resources utilization, and spatial sampling.

Yitong Wang received the bachelor's degree in communication engineering from the Hebei University of Technology, Tianjin, China, in 2020, and the master's degree in electronic information from the School of Information and Communication Engineering, Harbin Engineering University, Harbin, China, in 2023.

His research interests include array signal processing, radar mainlobe interference suppression, adaptive beamforming algorithm, and soil moisture inversion.

Jinghui Yang received the B.S., M.S., and Ph.D. degrees in information and communication engineering from Harbin Engineering University, Harbin, China, in 2010, 2013, and 2016, respectively.

She is currently an Associate Professor with the School of Information Engineering, China University of Geosciences, Beijing, Beijing, China. Her research interests include hyperspectral imagery, remote sensing, pattern recognition, and signal processing.