Enhanced Generalized Regression Neural Network With Backward Sequential Feature Selection for Machine-Learning-Driven Soil Moisture Estimation: A Case Study Over the Qinghai-Tibet Plateau

Ling Zhang, Zhaohui Xue, *Member, IEEE*, Huan Liu, and Hao Li

***Abstract*—Soil moisture (SM) is affected by many factors, such as soil characteristics, land cover, and meteorological conditions, making accurate remote sensing SM estimation a tough task. To fully explore the complementary information of multisource remote sensing data in SM estimation, it is necessary to explore the multiple feature variable selection method. Traditional filter methods may lead to feature redundancy and low accuracy, and embedding methods usually require complex parameter optimization. To overcome the above issues, we propose an enhanced generalized regression neural network with backward sequential feature selection (EBSFS) method for SM estimation. By using *k*-fold cross-validation to obtain the training set and validation set, and using the Pearson correlation coefficient to design evaluation criteria and an objective function, EBSFS searches for feature variables that minimize the objective function and updates the feature subset during iteration. EBSFS can adaptively obtain the optimal number of feature variables based on the evaluation criteria. Moreover, EBSFS does not require parameter optimization and can be flexibly and conveniently embedded into ensemble learning framework. Experiments conducted over the Qinghai-Tibet Plateau (QTP) from April 2015 to March 2016 demonstrate that EBSFS greatly reduces the feature redundancy, produces a more compact feature subset, and achieves higher estimation accuracy. Precisely, EBSFS presents better performance with R = 0.9544 and RMSE = 0.0310 under 13 input feature variables.**

***Index Terms*—Enhanced generalized regression neural network (EGRNN), feature selection, Qinghai-Tibet Plateau (QTP), soil moisture (SM) estimation.**

Manuscript received 30 March 2023; revised 3 June 2023 and 16 July 2023; accepted 21 July 2023. Date of publication 26 July 2023; date of current version 8 August 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42201406, Grant 41971279, and Grant 42271324, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20221506, in part by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under Grant 22KJD420001, and in part by the Sichuan Provincial Key Research and Development Program under Grant 2023YFS0432. *(Corresponding authors: Zhaohui Xue; Huan Liu.)*

Ling Zhang is with the School of Naval Architecture & Ocean Engineering, Jiangsu Maritime Institute, Nanjing 211100, China (e-mail: zhangling_jmi@163.com).

Zhaohui Xue and Hao Li are with the School of Earth Sciences and Engineering, Hohai University, Nanjing 211100, China, and also with the Jiangsu Province Engineering Research Center of Water Resources and Environment Assessment Using Remote Sensing, Hohai University, Nanjing 211100, China (e-mail: zhaohui.xue@hhu.edu.cn; lihao@hhu.edu.cn).

Huan Liu is with the Sichuan Academy of Safety Science and Technology, Chengdu 610045, China, and also with the Sichuan Anxin Kechuang Technology Co., Ltd., Chengdu 610045, China (e-mail: liuhuanraul@126.com).

Digital Object Identifier 10.1109/JSTARS.2023.3298946

I. Introduction

SOIL moisture (SM) is the link between surface water and groundwater, and it is one of the key variables in surface water cycle and atmospheric energy balance. It plays a key role in the global water and energy cycle, and directly affects hydrology cyclic processes, such as precipitation, runoff, infiltration, and evapotranspiration [1], [2]. Currently, with the development of satellite remote sensing technology, it has become possible to remotely sense SM over large areas, long time series, and high spatial resolutions.

The mainstream studies of SM estimation are based on satellite microwave remote sensing. Microwave and optical remote sensing are usually integrated for SM estimation. For example, an integration of microwave data obtained from the Soil Moisture Active Passive (SMAP) and Advanced Microwave Scanning Radiometer 2 (AMSR2) satellite radiometers was attempted to achieve an accurate SM estimation [3]. Similarly, an algorithm was developed for retrieving SM at the plateau scale by combined usage of Aquarius active and passive *L*-band observations [4]. The reflectance of Landsat-8 OLI and the backscattering coefficient of Sentinel-1 were combined with the Dubois model to jointly retrieve SM [5]. SM in agricultural fields was retrieved by the integration of Sentinel-1 A data and the Moderate Resolution Imaging Spectroradiometer (MODIS) data [6].

Recently, machine-learning-driven SM estimation methods can integrate multisource remote sensing data, topography and landform, vegetation and soil quality, without considering the complex physical mechanism of the inversion process, avoiding tedious model assumptions, and realizing complex nonlinear valid mapping [7], [8], [9]. The general idea of machine-learning-driven methods is to establish the mapping relationship between multiple feature variables and SM. According to the source of feature variables for SM estimation, they can be divided into visible light-near infrared features, thermal infrared features, microwave features, and other auxiliary features such as hydroclimate, topography, and soil properties. In [10], a nonlinear relationship between ground-based SM measurements from sparse network stations and passive microwave observations from the SMAP satellite was modeled. To evaluate the performance of different *L*-band SM products over the

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

tropical and the rainforests areas, a comprehensive evaluation was conducted [11]. Recently, a new research stream is to integrate the machine learning methods and the physical or semi-physical models for boosting the performance of SM estimation or downscaling [12], [13], [14], [15].

In order to fully utilize the advantages of multisource remote sensing data and explore the complementary information for SM estimation, it is necessary to explore multiple feature variables selection methods. Currently, feature variable selection algorithms can be roughly divided into three categories: filter; embedded; and wrapper methods.

1) Filter feature selection method is based on different indicators to measure the importance of features, which is independent of the model used. It scores each feature according to divergence or correlation, and ranks features by setting thresholds or the number of features to be selected. For regression, this type of method mainly includes Regression ReliefF (RReliefF) [16], Laplacian Score (LS) [17], Pearson Correlation Coefficient (PCC) [18], Minimum Redundancy Maximum Relevance (MRMR) [19], etc. Generally speaking, the filter algorithms are simple, suitable for the situation with many feature variables, and can quickly remove feature variables that are not sensitive to SM. However, the evaluation criteria are independent of the specific machine learning algorithm, which may cause feature redundancy.
2) Embedded feature selection method first obtains the weight coefficients of each variable based on model training, and selects variables according to the order of coefficients. This method is similar to the filter method, but it uses training to determine the quality of different features. For example, mean decrease in accuracy (MDA) [20], least absolute shrinkage and selection operator (Lasso) [21], etc. The embedded feature selection method is relatively fast and effective, but the parameter setting is complicated.
3) Wrapper feature selection method selects feature subsets based on the target function, which is usually the prediction performance score. It trains the model by selecting feature subsets using forward or backward selection. It adds or removes features one by one according to the custom evaluation criteria, and then sorts them based on the prediction performance until the evaluation index no longer decreases. Examples of wrapper feature selection methods include recursive feature elimination (RFE) [22] and sequential feature selection (SFS) [23]. Compared with filter and embedded methods, wrapper methods are task-oriented and tailor sensitive feature subsets for specific prediction models, resulting in higher accuracy and stronger specificity.

Currently, there are few studies on the impact of multisource remote sensing feature variables on SM estimation. Our previous work used ten feature selection algorithms and a Random Forest (RF) estimation model to analyze the impact of 29 features variables on SM estimation in the continental United States [24]. This study showed that the importance ranking of feature variables obtained by different feature selection methods was different. Overlaying the importance of different feature selection methods is cumbersome to implement, and the participation of too many feature selection methods inevitably introduces more uncertainty.

The Qinghai-Tibet Plateau (QTP) is the highest and largest plateau in the world, which has the most prominent and complex terrain on Earth and is one of the most sensitive regions to global climate change. Studies have shown that SM changes in the QTP have a significant impact on the climate system and water cycle in China, Asia, and even the world [25]. However, the scarcity and uneven distribution of ground measurement stations in the region, coupled with the complex natural and geographical environment, make large-scale and long-term remote sensing SM monitoring in the QTP very challenging. In particular, how to overcome the problems of machine learning models falling into local optima and overfitting under the condition of sparse samples needs to be further explored.

In the research of machine-learning-based SM estimation in the QTP, the relevant methods mainly include the use of back propagation neural network (BPNN), extreme gradient boosting (XGBoost), least squares regression (LSR), RF models, and so on.

1) In single-model-based estimation research, Zhang et al. [26] conducted SM estimation research using TerraSAR-X and Sentinel-1 data, combined with LSR and multitemporal change detection methods. To address the shortcomings of the traditional generalized regression neural network (GRNN) SM estimation method, i.e., using full map construction, which leads to high computational complexity and difficulty in effectively expressing the spatiotemporal local features, our previous work presented an enhanced generalized regression neural network (EGRNN) SM estimation model [27]. Further, GRNN is integrated with the physical and theoretical scale change (DisPATCh) algorithm to overcome the spatially discontinuous issue of current SM products [28].
2) In ensemble-learning-based estimation research, He et al. [29] used the Stacking method to integrate the "Trapezoid" model, RF, and XGBoost models, and used MODIS reflectance, DEM, land surface temperature (LST), vegetation index, and other variables to construct a feature variable space for SM estimating in the QTP. In addition, Zhang et al. [30] used Landsat 8 optical and thermal infrared data, SMAP, ECMWF Reanalysis v5 (ERA5), terrain, soil texture, and precipitation data to construct a feature variable space, and conducted SM estimation research based on XGBoost and RF models. To improve the generalization performance of a single model, Xue et al. [31] designed two novel ensemble learning models based on the Gaussian process regression (GPR) for SM estimation. In the research of SM products downscaling, Shangguan et al. [32] evaluated and intercompared the downscaling performances of six machine/deep learning approaches and further proposed a hybrid downscaling method based on Bayesian three-cornered hat merging (MATCH).

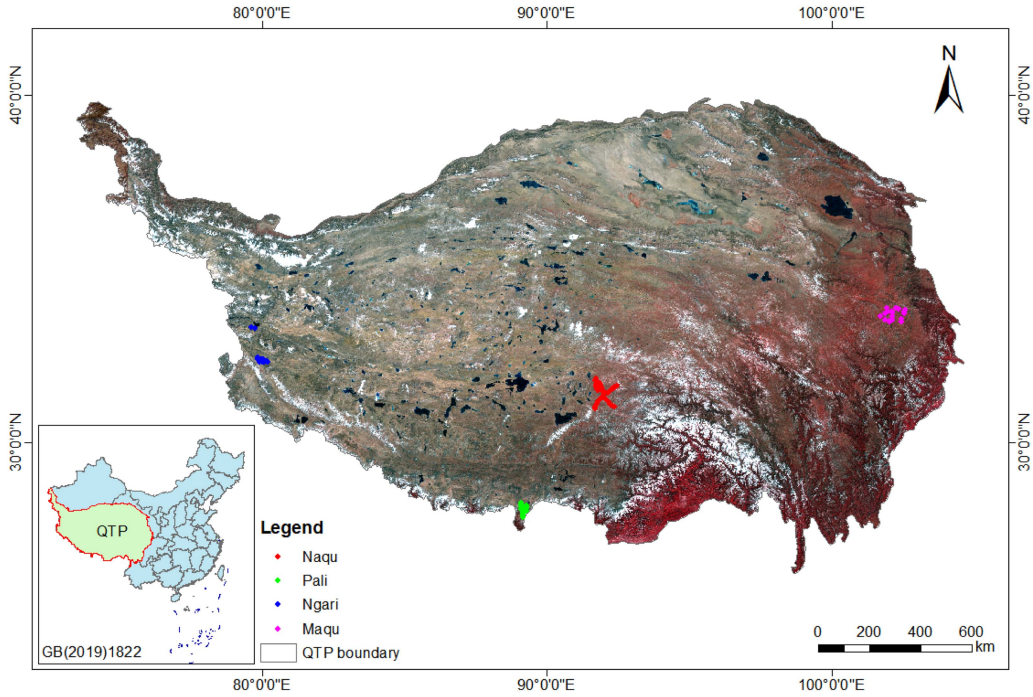

Fig. 1. Geographic location of the Qinghai-Tibet Plateau (QTP) and the spatial distribution of the ground stations from different networks. The China base map with the approval number GS(2019)1822 has not been modified.

To sum up, in the comprehensive machine-learning-driven SM estimation using multiple feature variables, the main considered variables include soil characteristics, vegetation characteristics, meteorological conditions, terrain features, underlying surface type, geographic location attributes, and temporal features, etc. Different feature variables may be independent of each other or may contain redundant information. Analyzing the sensitivity of feature variables to SM estimation can help improve the estimation accuracy of the model. Therefore, establishing an effective feature variables screening mechanism is an important issue that needs to be urgently addressed.

In this article, to address the shortcomings of the filter method, which may lead to feature redundancy and low accuracy, and the embedded method, which requires complex algorithm parameter optimization, an EGRNN with backward sequential feature selection (EBSFS) method is proposed for SM estimation. The original data is cross-validated by dividing the training samples into k equal parts to prevent overfitting. The Pearson correlation coefficient is used to calculate the prediction deviation of all cross-validation sets based on the given feature subset, and design evaluation criteria and objective function. In the iterative loop, a certain feature variable is sequentially removed from the current feature subset, and the objective loss of all cross-validation sets is calculated to find the feature variable that minimizes the objective function and update the feature subset. The iteration process is terminated when the objective function no longer decreases after removing a certain feature variable, and the optimal feature subset can be obtained.

The main novelty and contribution of our work are as follows.

1) We design a novel EBSFS method that adaptively obtains the optimal number of feature variables based on evaluation criteria, and the selected feature variables are more compact and have lower information redundancy. In addition, it can be flexibly and conveniently embedded into an ensemble learning framework.
2) We analyze the sensitivity of 19 feature variables for SM estimation over QTP, which can provide reference value for machine-learning-driven SM estimation studies in the literature.

II. Study Area and Data

A. Study Area

The QTP is located in the Central Asia, covering an area of more than 2.5 million square kilometers (see Fig. 1). With an average elevation exceeding 4500 m, QTP is often referred to as the "Roof of the World." QTP is also called the "water tower of Asia" since it contains the headwaters of the Yellow, Yangtze, and Mekong drainage basins. In addition, QTP is sometimes termed as the "Third Pole" because it contains the largest reserve of fresh water in the ice fields. Therefore, QTP can serve as a good indicator for the global climate change, greatly attracting the research interest of domestic and overseas scientists. Especially, SM changes in QTP have a great influence on the climate system and water cycle of the world. However, the mechanism of remote sensing SM monitoring is hard to be accurately modeled due to the complex natural and geographical environments. In addition, limited and unbalanced in situ measurements pose great challenges for large-scale and long-term SM monitoring.

TABLE I
DETAILS OF DIFFERENT IN-SITU SM MONITORING NETWORKS OVER THE QTP

Network	Ngari	Naqu	Maqu	Pali
Extent	32.3°-33.5°N; 79.5°-79.75°E	31°-32°N; 91.6°-92.5°E	33.5°-34.25°N; 101.75°-102.75°E	27.75°-28°N; 89°-89.25°E
Location	Western	Central	Eastern	Southern
# Stations	19	56	20	25
Depth (cm)	5	0-5	5	5
Collection interval	daily, 6:00 A.M.*	daily, 6:00 A.M.*	daily, 6:00 A.M.*	daily, 6:00 A.M.*
Acquired time	April 2015 to March 2016	April 2015 to March 2016	June 2015 to March 2016	April 2015 to December 2015
Climate	Arid	Semi-humid	Humid	Humid
Land cover	Bareland, Grassland	Grassland	Grassland, Wetland	Grassland

Note that: the time marked with * indicates the UTC time.

TABLE II
MULTIPLE FEATURE VARIABLES CONSIDERED FOR SM ESTIMATION ACQUIRED FROM APRIL 2015 TO MARCH 2016

Data name	Variable index	Variable name	Spatial resolution
MODIS	1	NDVI	500 m
	2	LST	1 km
	3	LC_Type1	500 m
	4	Sur_Ref	250 m
SMAP	5	Albedo	9 km
	6	Latitude	
	7	Longitude	
	8	SMAP_TBH	
	9	SMAP_TBV	
ERA-Interim	10	ERA_SR	0.125°
ERA5-Land	11	ERA_Eva	0.125°
	12	ERA_Runoff	
	13	ERA_TP	
SRTM	14	Elevation	90 m
HWSD	15	T_Clay	1 km
	16	T_Sand	
	17	T_Silt	
	18	SOC	
DOY	19	DOY	–

B. In Situ Data

We acquired the in situ measurements at 6:00 A.M. (UTC time) from April 2015 to March 2016, with a daily interval on the central Tibetan Plateau (CTP-SMTMN) [33] and the Tibetan Plateau Observatory (Tibet-Obs) [25] SM and temperature monitoring networks. As shown in Fig. 1, there are four in situ networks represented by different circles: the Naqu network (red circle) and Pali network (green circle) from CTP-SMTMN, and the Ngari network (blue circle) and Maqu network (magenta circle) from Tibet-Obs. Table I lists the details of different in situ networks.

C. Multiple Feature Variables

The 19 feature variables, obtained from five multisource data acquired from April 2015 to March 2016, listed in Table II are used in this study. These variables have been matched to the same time period, and they are also transformed into a uniform projection coordinate system and resampled to a common spatial resolution of 0.125° for consistency.

1) MODIS/Terra 16-Day vegetation indices products (MOD13A1), MODIS/Terra 8-day land surface temperature products (MOD11A2), MODIS/Terra land cover type products (MCD12Q1), and MODIS/Terra 8-day surface reflectance products (MOD09Q1) are used in this study. NDVI is a simple but good indicator of SM, because water stress can lead to spectral variation of the canopy in red and NIR bands. LST can be recognized as another SM indicator, given that the soil thermal conductivity changes with fluctuating moisture level. The MODIS datasets are available at https://modis.gsfc.nasa.gov.
2) SMAP enhanced L3 radiometer global daily 9-km EASE-grid SM V004 descending overpass datasets, including albedo, latitude and longitude, brightness temperatures at horizontal and vertical polarization (SMAP_TBH, SMAP_TBV), are used in this study. Existing empirical methods are commonly used to develop the regression relationships between passive microwaves measured brightness temperature (TB at both H and V polarizations) and ground SM over bare soils. SMAP datasets are available from https://nsidc.org/data/SPL3SMP.
3) The daily surface roughness (ERA_SR), daily evapotranspiration (ERA_Eva), daily runoff (ERA_Runoff), and total precipitation (ERA_TP) are used with a spatial resolution of 0.125°. ERA-Interim datasets can be obtained from http://apps.ecmwf.int/datasets.
4) The DEM data used in this study is provided by NASA's shuttle radar topography mission (SRTM) with a spatial resolution of 90 m. It started in February 2000, and covers an area of more than 119 million square kilometers between 60°N and 56°S. SRTM data can be obtained from http://srtm.csi.cgiar.org.
5) The soil texture data come from the 1 km harmonized world soil database version 1.2 (HWSD), in which the main soil classification system adopted is FAO-90. HWSD can be obtained from http://www.fao.org.
6) The day of year (DOY) should also be considered as a necessary feature to reflect time variations.

III. PROPOSED METHOD

Let $(\mathbf{X},\mathbf{Y})$ be n pairs of training samples with features $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ ($\mathbf{x}_i \in \mathbb{R}^B$ is a B-dimension feature) and targets $\mathbf{Y} = \{y_1, y_2, \ldots, y_n\}$. Let $\mathbf{X}^u = \{\mathbf{x}_1^u, \mathbf{x}_2^u, \ldots, \mathbf{x}_m^u\}$ be a set of unknown data. A graphical illustration of the proposed EBSFS scheme for SM estimation is shown in Fig. 2.

A. Initialization

To avoid overfitting, we equally split the training set into k subsets, i.e., $\{\mathbf{X}_1, \mathbf{Y}_1\}$, $\{\mathbf{X}_2, \mathbf{Y}_2\}, \ldots,$ $\{\mathbf{X}_k, \mathbf{Y}_k\}$, and each

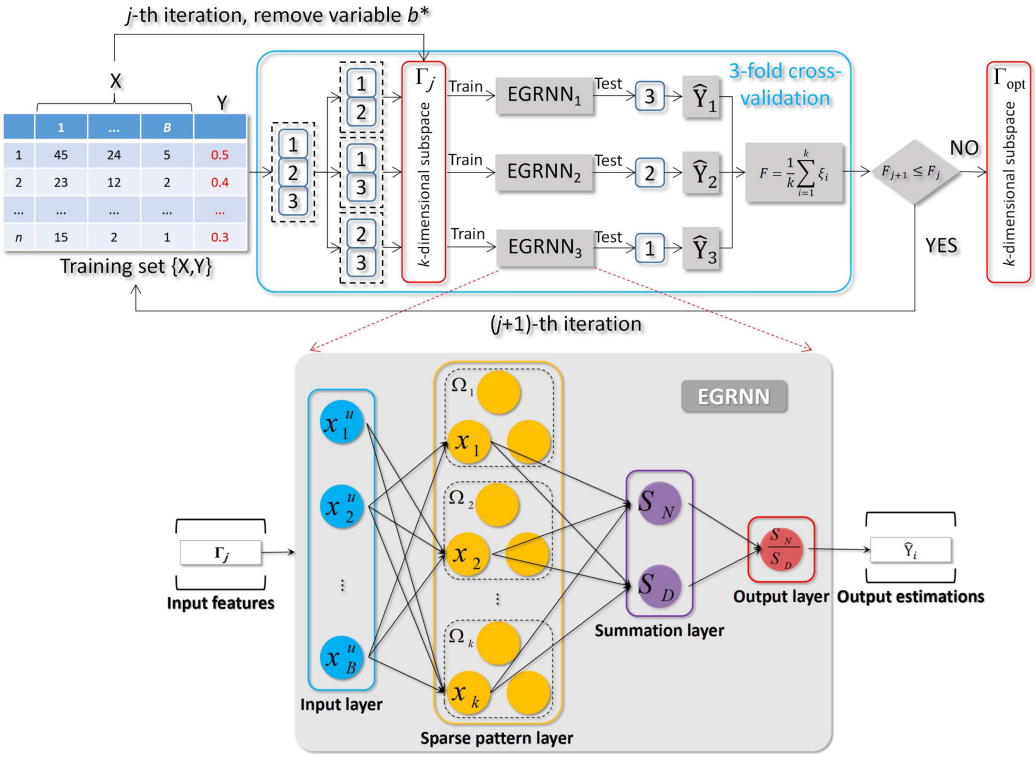

Fig. 2. Graphical illustration of the proposed EGRNN [27] with backward sequential feature selection (EBSFS) algorithm for SM estimation (Taking threefold cross-validation as an example). In each iterative loop, a certain feature variable (b^*) is sequentially removed from the current feature subset (Γ_j), and the objective loss (F) of all cross-validation sets is minimized to find the optimal feature subset (Γ_{opt}).

subset composes training and validation sets. Then, we initialize the K-dimensional feature subset as Γ_0 (where $K == B$).

B. Define Criterion and Objective Function

Given the feature subset, we can calculate the prediction bias (ξ) of all cross-validation sets as

$$\hat{\mathrm{Y}}_i = \mathrm{EGRNN}\left(\mathrm{X}_i, \mathrm{Y}_i, \mathrm{X}_i\right)$$
$$\xi_i = 1 - \mathrm{corr}\left(\hat{\mathrm{Y}}_i, \mathrm{Y}_i\right) \tag{1}$$

where function "corr" represents the Pearson coefficient. We then calculate the averaged prediction bias of all cross-validation sets based on the given feature subset. The objective function can be defined as

$$F = \frac{1}{k}\sum_{i=1}^{k}\xi_i. \tag{2}$$

C. Iterative Optimization

In the jth iteration, sequentially remove a feature variable from the current subset of features Γ_j, calculate the loss F_j, and find the feature variable b^* that minimizes the objective function, which takes the form

$$b^* = \arg\min_{b\in\Gamma_j} F_j. \tag{3}$$

Then, we need to remove b^* from the current subset of features, update the optimal subset of features Γ_j, and repeat the optimization given in (3)

$$\Gamma_j = \Gamma_j / b^* \tag{4}$$

where the symbol "A/a" means removing element "a" from "A."

If the loss F_j no longer reduces the performance after removing a feature variable, then terminate the optimization process. Otherwise, continue to the next iteration

$$\begin{aligned} &if \quad F_{j+1} \geq F_j, && end \\ &else, && continue. \end{aligned} \tag{5}$$

Finally, the optimal feature subset can be obtained after iterative optimization, i.e., $\Gamma_{\text{opt}} = \Gamma_j$.

With the above formulations at hand, the pseudocode of EBSFS can be summarized in Algorithm 1. EBSFS overcomes the drawbacks of existing feature selection algorithms. 1) It avoids the problem of feature redundancy that occurs in feature selection methods such as filter feature selection and forward sequential feature selection. 2) It avoids the problem of obtaining an optimal feature subset that may not be suitable for a specific task when filter feature selection is detached from the prediction model. 3) It eliminates the issue of complex parameter setting in embedded feature selection algorithms.

The advantages of EBSFS are as follows: 1) It adaptively obtains the optimal number of feature variables based on the evaluation criteria, enhancing the operability and targeting of the algorithm; and 2) The algorithm does not require parameter optimization and can be flexibly and conveniently embedded in ensemble learning framework.

Algorithm 1: EGRNN With Backward Sequential Feature Selection (EBSFS).

Input: Training data pairs with features $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and targets $\mathbf{Y} = \{y_1, y_2, \ldots, y_n\}$; Cross-validation times k.
Output: The optimal feature subset Γ_{opt}.
Main loop:
for $j = 1, 2, \ldots, B$ **do**
  for $i = 1, 2, \ldots, k$ **do**
    $\hat{\mathrm{Y}}_i = \mathrm{EGRNN}(\mathrm{X}_i, \mathrm{Y}_i, \mathrm{X}_i)$
    $\xi_i = 1 - \mathrm{corr}(\hat{\mathrm{Y}}_i, \mathrm{Y}_i)$
  end for
  $F_j = \frac{1}{k}\sum_{i=1}^{k} \xi_i$
  $b^* = \arg\min_{b \in \Gamma_j} F_j, \quad j \in \{1, 2, \ldots, B\}$
  $\Gamma_j = \Gamma_j / b^*$
  if $F_{j+1} \geq F_j$ end; else, move to Step 4.
end for
return $\Gamma_{\text{opt}} = \Gamma_j$

IV. Experimental Results

A. Experimental Settings

1) A total of 5582 labeled samples are randomly split into two parts: one for training (60%) and the other for test (40%).
2) For EGRNN, we set $k = 5$ as recommended in our previous work [27].
3) RReliefF [16], LS [17], MDA [20], PCC [18], and MRMR [19] are compared to evaluate the superiority of the proposed method.
4) Pearson correlation coefficient (R), root mean square error (RMSE), mean bias (bias), and unbiased RMSE (ubRMSE) are used to evaluate the performance.
5) The experiments are implemented on Intel Core i9-9900 K CPU, 3.6 GHz, and 64 GB RAM using MATLAB R2022a software. For stability, we conducted 20 independent runs and reported the averaged results.[1]

B. Feature Importance and Rank

We analyze the feature importance and rank of different feature selection methods, and we normalize the feature importance to [0,1] for comparability. As shown in Fig. 3, a total of 18 feature variables except for DOY are included for analysis. Since DOY reflects time variations, it is considered as a necessary feature. The higher feature importance represents more sensitivity to SM estimation, thus more forward of the feature rank.

According to the results, the feature importance and rank of different feature selection methods have significant difference or even opposite performance. For example, MODIS_NDVI is ranked first, first, and third by RReliefF, LS, and PCC, respectively. However, it is ranked fifth and tenth by MDA and MRMR, respectively. ERA_SR is ranked 1st by MDA and MRMR, but it is ranked third, 11th, and 16th by RReliefF, LS, and PCC, respectively. SRTM_Elevation is ranked second and third by RReliefF and LS, whereas it is ranked eighth, eighth, and 15th by MDA, PCC, and MRMR, respectively. Therefore, the above observations indicate that different feature selection methods have significant bias on specific feature variables, leading to large uncertainty for SM estimation. The reason behind this is that RReliefF, LS, PCC, and MRMR are filter-based methods, the evaluation criteria are independent of the specific machine learning method. MDA is an embedded feature selection method that obtains the weight coefficients of different variables based on model training.

[1]The source code will be available at https://github.com/ZhaohuiXue/EBSFS.

C. Stacked Feature Importance

Since different feature selection method yields variant results, we stack the feature importance to validate the composite result. As shown in Fig. 4, SMAP_TBH > MODIS_LST > MODIS_LST > ERA_SR > SMAP_TBV are the top five important feature variables. Whereas, HWSD_T_Silt < ERA_Runoff < ERA_TP < HWSD_T_Sand < HWSD_T_Clay are the top five insensitive feature variables.

By stacking the feature importance obtained by different methods, we can obtain more comprehensive feature variables. However, it is difficult in practical use since we need to implement multiple feature selection methods and composite their results. In addition, the involvement of multiple feature selection methods will unavoidably introduce additional uncertainty.

D. Optimal Number of Feature Variables

To validate the optimal number of feature variables for different feature selection methods, we plot the stairs of R as a function of the number of features. As shown in Fig. 5, we conclude the following observations. First, the SM estimation accuracy increases as the number of features also increases. Second, the optimal number of features for different methods is different. RReliefF and MDA need 16 features to obtain the highest R, PCC and MRMR need 18 features, whereas EBSFS only need 12 features to achieve the best R. Third, the performance with very few features is quite different. For example, the values of R for EBSFS, MDA, and MRMR can reach to 0.9 with only two features, whereas the value of R is around 0.85 for RReliefF and LS, but it is less than 0.8 for PCC. Finally, the increasing speed of R along with the number of features is variant in different cases. For example, the increasing speed for EBSFS is fastest, but the other methods are not stable. The above observations indicate that the features selected by most of the methods have certain information redundancy, but EBSFS can select more compact and representative feature subset for SM estimation.

E. SM Estimation Results

We then validate the SM estimation performance using the selected features. First, we present the scatter plots for different methods to demonstrate the linear relationship between predictions and in situ measurements. As shown in Fig. 6, EBSFS obviously exhibits the best fitting performance since the fitting line is more close to the 1:1 line, and the scatter points are

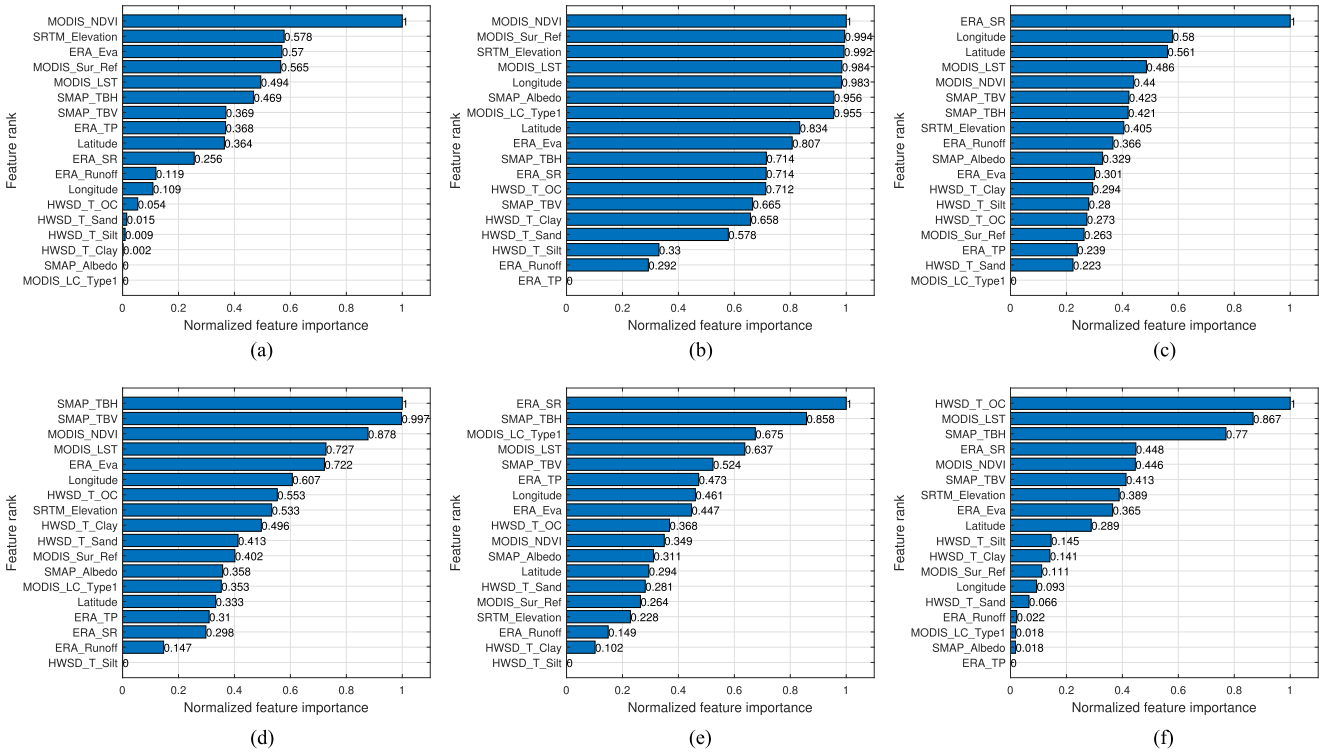

Fig. 3. Feature importance and rank obtained by different methods. (a) RReliefF, (b) LS, (c) MDA, (d) PCC, (e) MRMR, (f) EBSFS.

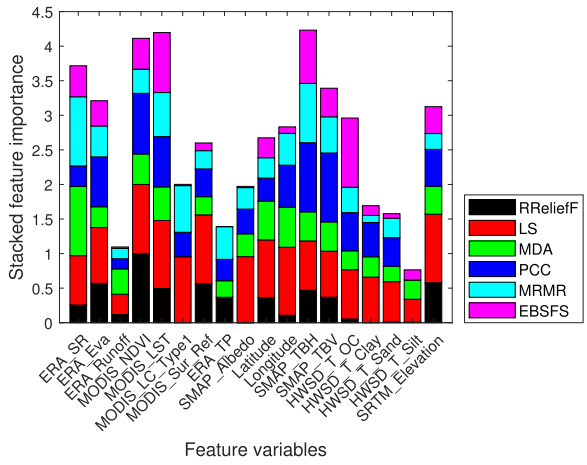

Fig. 4. Histograms of stacked feature importance.

more compact. As for the other methods, their fitting lines deviate from the 1:1 line with lower R values. As for specific accuracy improvement, we can observe that EBSFS significantly outperforms the other counterparts with R = 0.9544 and RMSE = 0.0310 cm^3/cm^3, and the performance improvements are 0.0026–0.0269 and 0.0008–0.0078 in terms of R and RMSE, respectively. Therefore, the above observation verifies that the proposed method has better regression performance compared to the other methods.

For visual inspection, we present the annual mean SM maps in Fig. 7. For different method, SM in the middle of QTP increases gradually from west to east, and north to south, according to the climate distribution. RReliefF overestimates SM in the central part, MDA also overestimates SM in the east of QTP, and PCC underestimates SM in most of the regions. By comparing the other methods, EBSFS captures more specific spatial dynamics, which indicates that our model has powerful regression performance.

V. Discussion

A. Temporal Dynamics

As shown in Fig. 8, we illustrate the time-series of estimated SM for one year. We can observe that the predicted SM values based on EBSFS are more consistent with the true values in the time range marked A, whereas the predicted results of other compared methods are either higher or lower than the true value. For example, MDA and PCC underestimate the SM values, and MRMR overestimates the SM values. In the time range marked B, the predictions of MDA and PCC are underestimated, whereas the other methods are quite similar. In the time range marked C, the predictions of the MDA, PCC, and EBSFS are lower than the real value, but RReliefF and LS provide the least differences between the estimated values and the in situ measurements. In general, the above experimental results demonstrate that the proposed method can well capture spatial-temporal dynamics of SM in the study area.

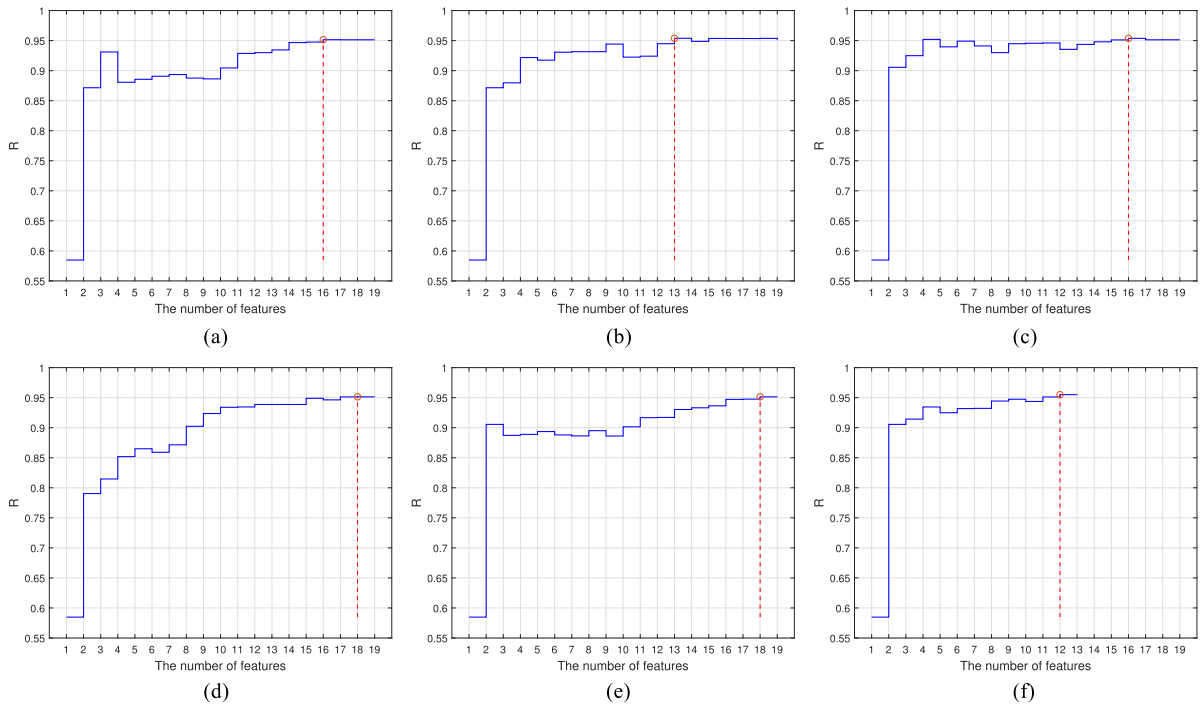

Fig. 5. R as a function of the number of features. The red dotted lines indicate the optimal number of features in each case. (a) RReliefF, (b) LS, (c) MDA, (d) PCC, (e) MRMR, (f) EBSFS. (a) Opt = 16. (b) Opt = 13. (c) Opt = 16. (d) Opt = 18. (e) Opt = 18. (f) Opt = 12.

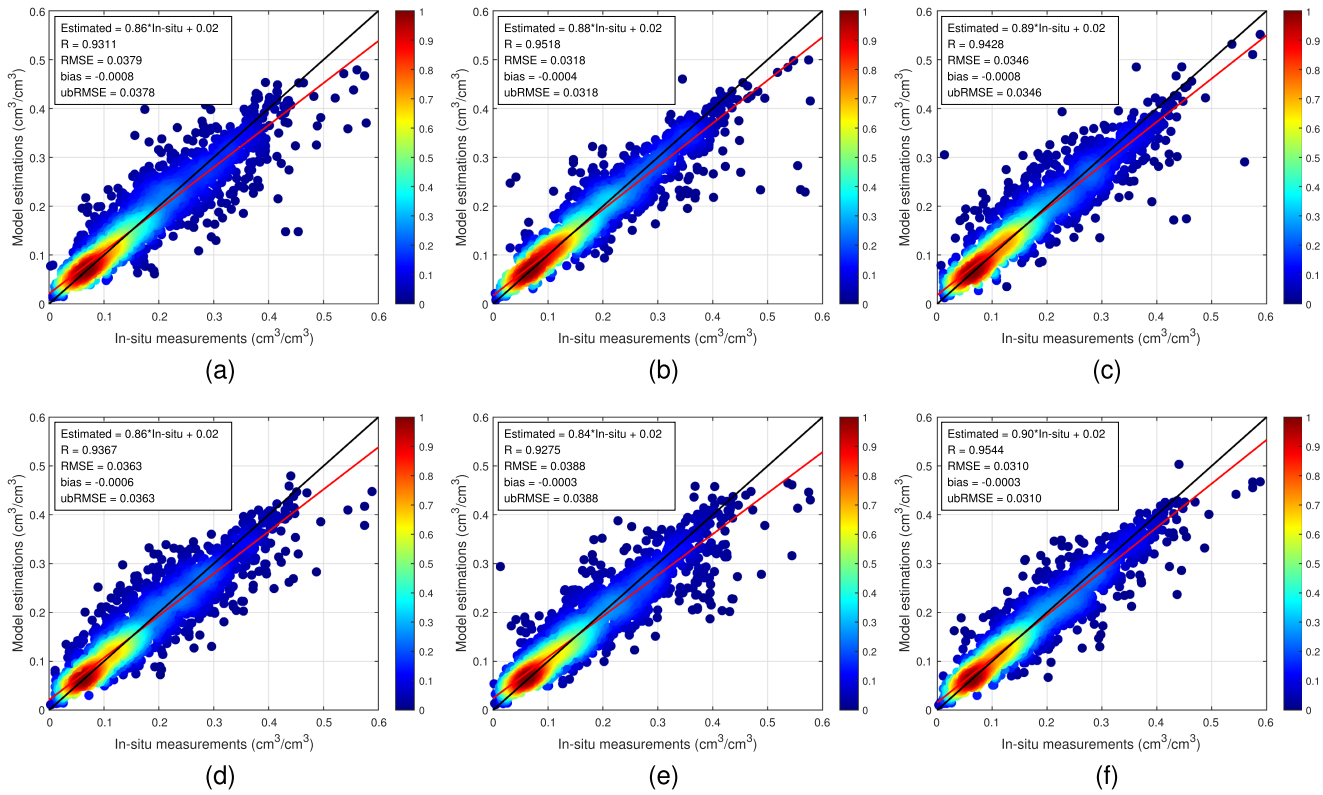

Fig. 6. Scatter plots of different methods by fitting SM estimations with in situ measurements. The red lines indicate the fitting line, and the black diagonal lines represent 1:1 line for reference. (a) RReliefF, (b) LS, (c) MDA, (d) PCC, (e) MRMR, (f) EBSFS.

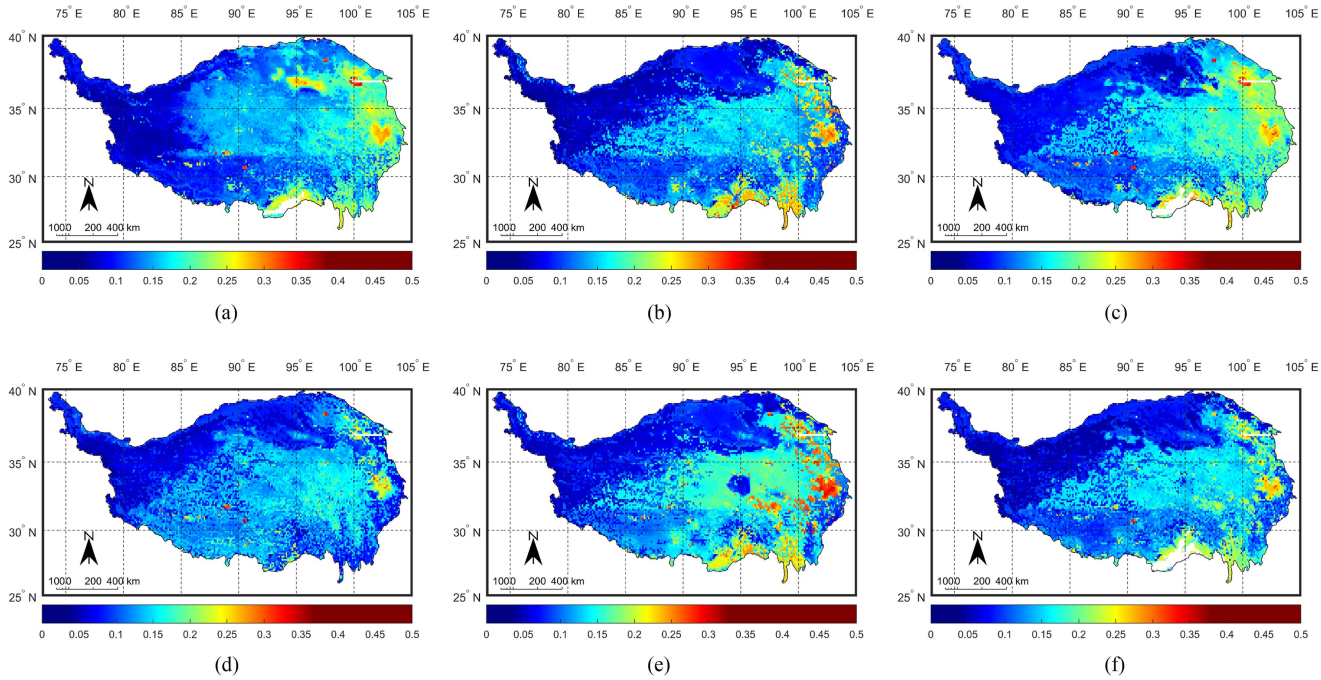

Fig. 7. SM maps obtained by different methods. (a) RReliefF, (b) LS, (c) MDA, (d) PCC, (e) MRMR, (f) EBSFS.

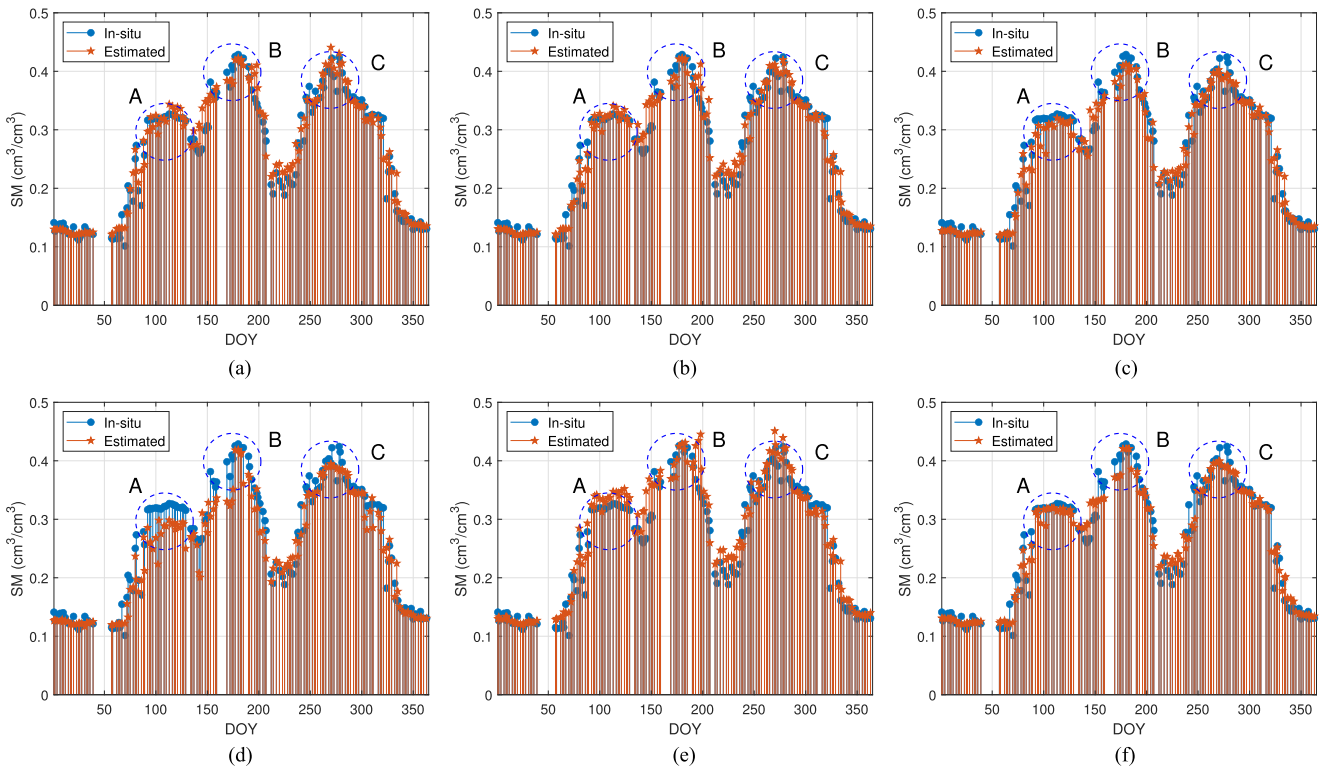

Fig. 8. Temporal dynamics of daily estimated SM at an arbitrary location: 33°38′N, 101°53′E. (a) RReliefF, (b) LS, (c) MDA, (d) PCC, (e) MRMR, (f) EBSFS.

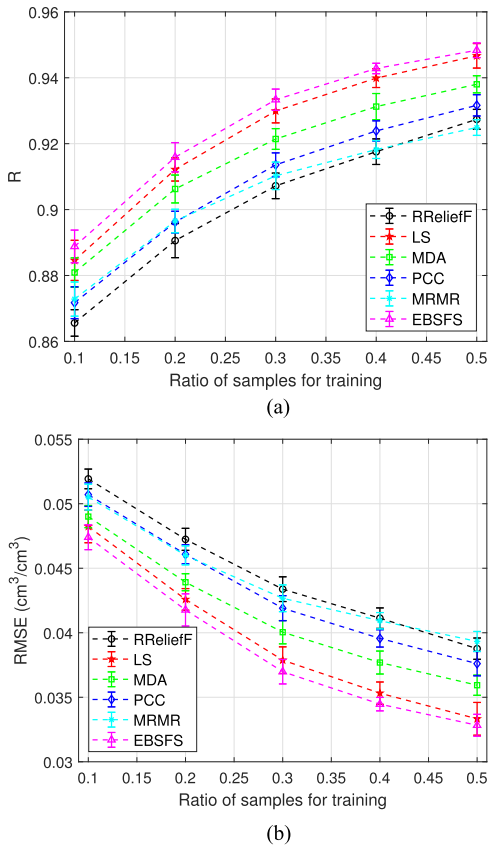

Fig. 9. Estimation accuracy as a function of different ratio of training samples. The error bars indicate standard deviations. (a) R and (b) RMSE.

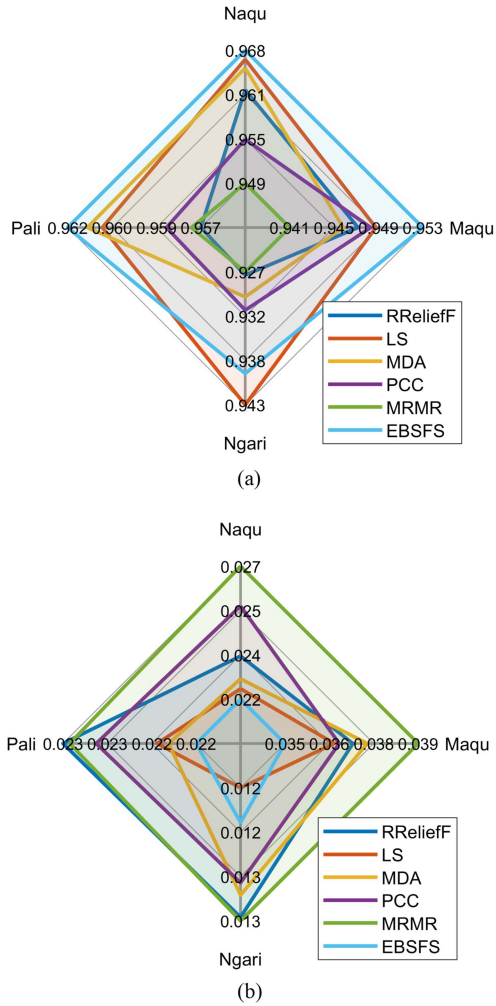

Fig. 10. Radar plots for different feature selection methods on the four SM monitoring networks. (a) R and (b) RMSE.

B. Generalization Performance

1) Evaluation of Different Initial Training Samples: As we known, machine-learning-driven SM estimation results are greatly affected by the initial training set. To evaluate the generalization performance of the proposed method in terms of different initial training set, Fig. 9 illustrates the evolution of R and RMSE as a function of different ratio of training samples for different feature selection methods. As depicted in the figure, EBSFS always obtains the highest accuracy among all considered methods when the ratio ranges from 10% to 50%. It's worth noting that, when the training set is very small, e.g., with only 10% samples used for training, the proposed method also performs very well with R higher than 0.88. The above observations demonstrate that EBSFS has better generalization performance in terms of the initial training set than the other counterparts.

2) Evaluation of Different In Situ Networks: To evaluate the generalization performance of the proposed method on different ground stations, Fig. 10 depicts the radar plots for different feature selection methods based on four ground stations, i.e., Naqu, Pali, Ngari, and Maqu. It can be observed that the proposed method performs best on three of four different ground stations, including Naqu, Pali, and Maqu. The reason behind may due to the fact that our method is more sensitive in humid or semihumid regions. It's interesting to note that LS exhibits the best prediction performance in Ngari ground stations. In short, the proposed method has the best stability under most of the ground stations.

C. Uncertainty Analysis

We use the triple collocation (TC) [34] to evaluate the uncertainty of different feature selection methods. TC method does not require known ground truth as a reference and can directly evaluate the uncertainty of three or more sets of observed sequences. To prepare the triple input for TC, we combine the SM products from SMAP and ERA5-Land with the SM estimations results obtained by each feature selection method. As shown in Fig. 11, the averaged uncertainty of SM estimations results obtained from different feature selection methods ranges from 0.8 to 0.85, and the spatial differences are not significant. EBSFS has the lowest relative uncertainty, i.e., $R_{TC} = 0.8576$, while ReliefF and MDA feature selection methods have the highest relative uncertainty.

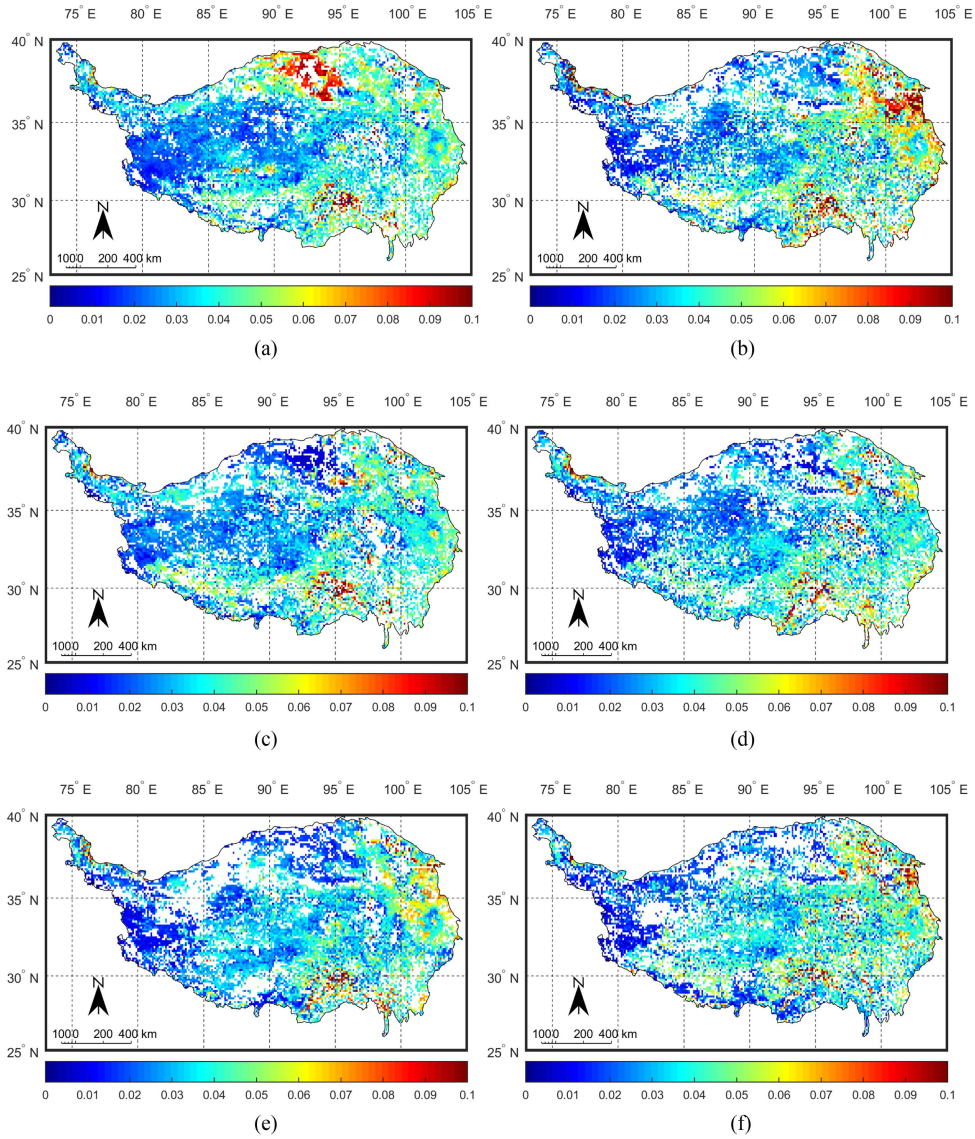

Fig. 11. Spatial uncertainty of different feature selection methods. (a) RReliefF, (b) LS, (c) MDA, (d) PCC, (e) MRMR, (f) EBSFS. The parentheses indicate the average uncertainty measured by TC. (a) $R_{TC} = 0.8083$. (b) $R_{TC} = 0.8536$. (c) $R_{TC} = 0.8152$. (d) $R_{TC} = 0.8249$. (e) $R_{TC} = 0.8200$. (f) $R_{TC} = 0.8576$.

VI. Conclusion

In this article, we propose a novel EGRNN with backward sequential feature selection method for SM estimation, namely, EBSFS. First, k-fold cross-validation is used to divide the training samples into training set and validation set. Second, Pearson correlation coefficient is used to design evaluation criteria and objective function. Finally, iterative optimization is conducted to find the feature variables that minimize the objective function and update the feature subset during iteration.

The proposed method is tested in the QTP during April 2015 to March 2016 based on 19 feature variables. In the comparison with four other different feature selection methods, we conclude the following: 1) Different feature selection methods have significant differences in the importance ranking of different feature variables, which leads to great uncertainty for SM estimation, i.e., PCC underestimates SM and MRMR overestimates SM; 2) The information redundancy among the feature variables extracted by EBSFS is relatively low, and the accuracy can maintain steady growth as the number of feature variables increases; 3) At the optimal number of features, the estimation accuracy obtained by EBSFS is significantly better than other comparative methods; and 4) The most sensitive feature variables for SM estimation over QTP include SMAP_TBH, MODIS_LST, MODIS_NDVI, ERA_SR, SMAP_TBV, ERA_Eva, SRTM_Elevation. Meanwhile, the SM map produced by EBSFS has richer details in space and fits better with the measured data.

In the future, we will focus on integrating the current method with ensemble learning framework. In addition, our method has not been well coupled with the physical SM inversion mechanism and underlying surface characteristics, which also deserves further exploration since physical model can well improve the interpretablility of machine-learning-driven methods.

REFERENCES

[1] Z. L. Li, P. Leng, C. H. Zhou, K. S. Chen, F. C. Zhou, and G. F. Shang, "Soil moisture retrieval from remote sensing measurements: Current knowledge and directions for the future," *Earth- Sci. Rev.*, vol. 218, 2021, Art. no. 103673.

[2] J. Peng et al., "A roadmap for high-resolution satellite soil moisture applications–confronting product characteristics with user requirements," *Remote Sens. Environ.*, vol. 252, 2021, Art. no. 112162.

[3] E. Santi, S. Paloscia, S. Pettinato, L. Brocca, L. Ciabatta, and D. Entekhabi, "Integration of microwave data from SMAP and AMSR2 for soil moisture monitoring in Italy," *Remote Sens. Environ.*, vol. 212, pp. 21–30, 2018.

[4] Q. Wang, R. van der Velde, P. Ferrazzoli, X. L. Chen, X. J. Bai, and Z. B. Su, "Mapping soil moisture across the Tibetan Plateau plains using aquarius active and passive L-band microwave observations," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 77, pp. 108–118, 2019.

[5] S. Huang et al., "The capability of integrating optical and microwave data for detecting soil moisture in an Oasis region," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1358.

[6] Y. Han, X. Bai, W. Shao, and J. Wang, "Retrieval of soil moisture by integrating Sentinel-1A and modis data over agricultural fields," *Water*, vol. 12, no. 6, 2020, Art. no. 1726.

[7] Q. Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, 2020, Art. no. 111716.

[8] I. P. Senanayake, I. Y. Yeo, J. P. Walker, and G. R. Willgoose, "Estimating catchment scale soil moisture at a high spatial resolution: Integrating remote sensing and machine learning," *Sci. Total Environ.*, vol. 776, 2021, Art. no. 145924.

[9] A. S. Abowarda et al., "Generating surface soil moisture at 30m spatial resolution using both data fusion and machine learning toward better water resources management at the field scale," *Remote Sens. Environ.*, vol. 255, 2021, Art. no. 112301.

[10] Q. Q. Yuan, H. Z. Xu, T. W. Li, H. F. Shen, and L. P. Zhang, "Estimating surface soil moisture from satellite observations using a generalized regression neural network trained on sparse ground-based measurements in the continental U.S.," *J. Hydrol.*, vol. 580, 2020, Art. no. 124351.

[11] H. L. Ma et al., "An assessment of L-band surface soil moisture products from SMOS and SMAP in the tropical areas," *Remote Sens. Environ.*, vol. 284, 2023, Art. no. 113344.

[12] H. Sun and Y. Cui, "Evaluating downscaling factors of microwave satellite soil moisture based on machine learning method," *Remote Sens.*, vol. 13, no. 1, 2021, Art. no. 133.

[13] H. Sun, X. Zhang, and X. Zhao, "Series or parallel? an exploration in coupling physical model and machine learning method for disaggregating satellite microwave soil moisture," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[14] H. Sun and Z. Zhao, "Effects of the new Priestly-Taylor equation on determining the boundary of LST/FVC space for soil moisture monitoring," *Geocarto Int.*, vol. 37, no. 26, pp. 11534–11558, Dec. 2022.

[15] H. Sun and J. Gao, "A pixel-wise calculation of soil evaporative efficiency with thermal/optical remote sensing and meteorological reanalysis data for downscaling microwave soil moisture," *Agricultural Water Manage.*, vol. 276, 2023, Art. no. 108063.

[16] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, no. 1, pp. 23–69, 2003.

[17] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. 18th Int. Conf. Neural Inf. Process. Syst.*, 2005, pp. 507–514.

[18] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D dissertation, Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, 1999.

[19] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinf. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, 2005.

[20] H. Han, X. Guo, and H. Yu, "Variable selection using mean decrease accuracy and mean decrease gini based on random forest," in *Proc. 7th IEEE Int. Conf. Softw. Eng. Service Sci.*, 2016, pp. 219–224.

[21] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. Ser. B: Statist. Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[22] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1, pp. 389–422, 2002.

[23] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognit. Lett.*, vol. 15, no. 11, pp. 1119–1125, Nov. 1994.

[24] L. Zhang, Z. Zhang, Z. Xue, and H. Li, "Sensitive feature evaluation for soil moisture retrieval based on multi-source remote sensing data with few in-situ measurements: A case study of the continental U.S.," *Water*, vol. 13, no. 15, 2021, Art. no. 2003.

[25] P. Zhang et al., "Status of the Tibetan Plateau observatory (Tibet-Obs) and a 10-year (2009-2019) surface soil moisture dataset," *Earth System Sci. Data*, vol. 13, no. 6, pp. 3075–3102, Jul. 2021.

[26] X. Zhang et al., "Soil moisture estimation based on the distributed scatterers adaptive filter over the QTP permafrost region using Sentinel-1 and high-resolution TerraSAR-X data," *Int. J. Remote Sens.*, vol. 42, no. 3, pp. 902–928, Feb. 2020.

[27] L. Zhang, Z. Xue, Y. Zhang, J. Ma, and H. Li, "Enhanced generalized regression neural network for soil moisture estimation over the Qinghai-Tibet Plateau," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3815–3829, 2022.

[28] P. Leng et al., "A framework for estimating all-weather fine resolution soil moisture from the integration of physics-based and machine learning-based algorithms," *Comput. Electron. Agriculture*, vol. 206, 2023, Art. no. 107673.

[29] L. He, Y. Cheng, Y. X. Li, F. Li, K. L. Fan, and Y. Z. Li, "An improved method for soil moisture monitoring with ensemble learning methods over the Tibetan Plateau," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2833–2844, 2021.

[30] Y. F. Zhang, S. L. Liang, Z. L. 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.

[31] Z. Xue, Y. Zhang, L. Zhang, and H. Li, "Ensemble learning embedded with gaussian process regression for soil moisture estimation: A case study of the continental U.S.," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.

[32] Y. L. Shangguan, X. X. Min, and Z. Shi, "Inter-comparison and integration of different soil moisture downscaling methods over the Qinghai-Tibet Plateau," *J. Hydrol.*, vol. 617, 2023, Art. no. 129014.

[33] K. Yang et al., "A multiscale soil moisture and Freeze–Thaw monitoring network on the Third Pole," *Bull. Amer. Meteorological Soc.*, vol. 94, no. 12, pp. 1907–1916, 2013.

[34] A. Stoffelen, "Toward the true near-surface wind speed: Error modeling and calibration using triple collocation," *J. Geophysical Res.-Oceans*, vol. 103, no. C4, pp. 7755–7766, 1998.

Ling Zhang received the B.S. degree in geomatics engineering from Shandong Agricultural University, Tai'an, China, in 2010, the M.E. degree in remote sensing from the China University of Mining and Technology, Beijing, China, in 2013, and the Ph.D. degree in surveying and mapping from Hohai University, Nanjing, China, in 2022.

Currently, she is a Lecturer with the Jiangsu Maritime Institute, Nanjing, China. Her research interests include using ensemble learning and deep learning for SM estimation.

Zhaohui Xue (Member, IEEE) received the B.S. degree in geomatics engineering from Shandong Agricultural University, Tai'an, China, in 2009, the M.E. degree in remote sensing from the China University of Mining and Technology, Beijing, China, in 2012, and the Ph.D. degree in cartography and geographic information system from Nanjing University, Nanjing, China, in 2015.

He is currently a Professor (Ph.D. supervisor) with the School of Earth Sciences and Engineering, Hohai University, Nanjing, China. He has authored more than 50 scientific papers including more than 30 Science Citation Index (SCI) papers. His research interests include hyperspectral image classification, time-series image analysis, pattern recognition, and machine learning.

Dr. Xue was the recipient of the National Scholarship for Doctoral Graduate Students granted by the Ministry of Education of the People's Republic of China in 2014. He was awarded the Best Reviewer for the *IEEE Geoscience and Remote Sensing Society*. He is an Editorial Board Member in *National Remote Sensing Bulletin* (2020–2024). He has been a reviewer for more than ten famous remote sensing journals including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and *ISPRS Journal of Photogrammetry and Remote Sensing*.

Huan Liu received the B.S. degree in surveying and mapping from the Guilin University of Technology, Guilin, China, in 2009, and the M.E. degree in geodesy and survey engineering from the China University of Mining and Technology, Beijing, China, in 2012.

Currently, he is a Scientific Researcher with the Sichuan Academy of Safety Science and Technology, Chengdu, China, and the Sichuan Anxin Kechuang Technology Company Ltd., Chengdu, China. His research interest includes the application of aerospace remote sensing technology in the fields of natural disaster emergency monitoring, disaster prevention, reduction and relief.

Hao Li received the B.S. degree in aerial photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1985, the M.E. degree in geodesy and survey engineering and the Ph.D. degree in hydraulic structure engineering from Hohai University, Nanjing, China, in 1995 and 2005, respectively.

He is currently a Professor with the School of Earth Sciences and Engineering, Hohai University. He is the author of almost 100 journal papers and five books. His research interests include photogrammetry and remote sensing, computer vision, and geographic information system.

Dr. Li has been honored as an recipient of more than ten science and technology awards. He was a leader of more than 50 research projects hold by the Country and local government. Now, he serves as the Vice Chairman of the Photogrammetry and Remote Sensing Special Committee of Jiangsu Society of Surveying and Mapping.