Temporal-Spatial Soil Moisture Estimation from CYGNSS Using Machine Learning Regression With a Preclassification Approach

Yan Jia, *Member, IEEE*, Shuanggen Jin, *Senior Member, IEEE*, Haolin Chen, Qingyun Yan, *Member, IEEE*, Patrizia Savi, *Senior Member, IEEE*, Yan Jin, and Yuan Yuan, *Member, IEEE*

***Abstract*—Global navigation satellite system-reflectometry (GNSS-R) can retrieve Earth's surface parameters, such as soil moisture (SM) using the reflected signals from GNSS constellations with advantages of noncontact, all-weather, real-time, and continuity, particularly the space-borne cyclone GNSS (CYGNSS) mission. However, the accuracy and efficiency of SM estimation from CYGNSS still need to improve. In this article, the global SM is estimated using machine learning (ML) regression aided by a preclassification strategy. The total observations are classified by land types and corresponding subsets are built for constructing ML regression submodels. Ten-fold cross-validation technique is adopted. The overall performance of SM estimation with/without preclassification is compared, and the results show that the SM estimations using different ML algorithms all have substantial improvement with the preclassification strategy. Then, the optimal XGBoost predicted model with root-mean-square error (RMSE) of 0.052 cm^3/cm^3 is adopted. In addition, the satisfactory daily and seasonal SM prediction outcomes with an overall correlation coefficient value of 0.86 and an RMSE value of 0.056 cm^3/cm^3 are achieved at a global scale, respectively. Furthermore, the extensive temporal and spatial variations of CYGNSS SM predictions are evaluated. It shows that the reflectivity plays a main role among the predictors in SM estimation, and the next is vegetation. In some extremely dry places, the roughness may become more important. The value of SM is positively correlated with RMSE and also another limit condition that will constrain the variation of predictors, thus affecting correlation coefficient R and RMSE. Also, we compare both SMAP and CYGNSS SM predictions against *in situ* SM measurements from 301 stations. Similar low-median unbiased RMSEs are obtained, and the daily averaged CYGNSS-based SM against the *in situ* networks is 0.049 cm^3/cm^3. The presented approach succeeds in providing SM estimation at a global scale with employing the least ancillary data with superior results and this article reveals the spatio-temporal heterogeneity for SM estimation using CYGNSS data.**

***Index Terms*—CYGNSS, GNSS-Reflectometry, preclassification, SMAP, soil moisture, XGBoost.**

I. Introduction

SOIL moisture (SM) is an important indicator in the fields of climate, hydrology, ecology, and agriculture [1]–[3]. The spatio-temporal change and distribution of SM have a significant impact on the Earth-atmosphere energy balance, atmospheric circulation, and soil temperature [4], [5]. Hence, the monitoring of SM on a large scale is an important part of agricultural research and the evaluation of environmental factors. It is of great significance in improving the global climate and predicting regional precipitation events [6]–[7].

Many passive microwave sensors have been used to observe surface SM (<5 cm), such as NASA's the Advanced Microwave Scanning Radiometer-Earth Observing System [8] (AMSR-E), the Soil Moisture Passive and Active [9] (SMAP), and the Soil Moisture and Ocean Salinity [10] (SMOS) of the European Space Agency. Although microwave sensors can be used to obtain high-precision SM products, the 2–3 days revisit period (SMAP) restricts its applications on higher time resolution. Additionally, some active platforms, e.g., Sentinel-1 [11] and ERRASAR-X [12], can also provide SM estimation through radar backscattering measurements, but the time resolution is lower with around 6 days.

In recent years, the technology of Global Navigation Satellite System-Reflectometry (GNSS-R) has been receiving much attention due to its free, 24-h, and flexible properties [13]. The GNSS-R signal is very sensitive to the properties of the reflecting surface. The reflected signal is especially related to the permittivity that can be detected depending on the strength of the reflected signal [14]. GNSS-R was first proposed for ocean remote sensing and further extended to the land surface [15]–[17]. Many institutions, for example, Jet Propulsion Laboratory (JPL), the University of Colorado at Boulder, Institut d'Estudis Espacials de Catalunya-Universitat Politècnica de Catalunya

Manuscript received January 20, 2021; revised March 31, 2021 and April 22, 2021; accepted April 24, 2021. Date of publication April 29, 2021; date of current version May 26, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 42001375, Grant 42001362, Grant 41901356, and Grant 42001332, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20180765, in part by the Nanjing Technology Innovation Foundation for Selected Overseas Scientists under Grant RK032YZZ18003, in part by the NUPTSF under Grant 219066, and in part by the Shanghai Leading Talent Project under Grant E056061, and in part by the Strategic Priority Research Program Project of the Chinese Academy of Sciences under Grant XDA23040100. *(Corresponding author: Shuanggen Jin.)*

Yan Jia, Haolin Chen, Yan Jin, and Yuan Yuan are with the Department of Surveying and Geoinformatics, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: jiayan@njupt.edu.cn; 787284069@qq.com; jinyan@njupt.edu.cn; yuanyuan@njupt.edu.cn).

Shuanggen Jin and Qingyun Yan are with the School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China, and also with the Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China (e-mail: sgjin@shao.ac.cn; jayyqy@qq.com).

Patrizia Savi is with the Department of Electronic and Telecommunication, Politecnico di Torino, 10129 Torino, Italy (e-mail: patrizia.savi@polito.it).

Digital Object Identifier 10.1109/JSTARS.2021.3076470

(IEEC-UPC), and Starlab in Spain carried out a series of theoretical studies and experiments using GNSS reflection signals, and successively developed soft/hardware GNSS-R receivers [18]–[20].

With the continuous development of GNSS-R SM retrieval, new constellation observation programs [21]–[26] with long-term observation data have become a new approach for GNSS-R SM retrieval. At present, some significant results have been found utilizing cyclone GNSS (CYGNSS) data for the SM application [27]–[32]. Among them, Kim and Venkat [27] proposed that the relative signal-to-noise ratio (rSNR) of CYGNSS can be used to retrieve SM, and the regional daily SM estimation was derived by combining the rSNR of CYGNSS and the SM of SMAP. In moderate vegetation condition areas, the correlation coefficient (R) between SM obtained by CYGNSS and SMAP is 0.77, but in high-density vegetation areas, the R drops to 0.68. Chew and Small [28] found that the change of CYGNSS reflectivity was related to the change of SM in SMAP, and explained this correlation by a linear regression method. An unbiased root-mean-square error (ubRMSE) of 0.045 cm^3/cm^3 was reported, thus further improving the accuracy of CYGNSS SM retrieval. Their CYGNSS-based SM data product, called "UCAR/CU," has been made publicly available. The performance was validated against point-scale *in situ* observations, with a median ubRMSE of 0.049 cm^3/cm^3 and a median R of 0.4. For the same station, the median ubRMSE between SMAP and *in situ* data sets was 0.045 cm^3/cm^3, with a median R of 0.69, showing that the SM product is complementary to SMAP [29]. Al-Khaldi *et al.* [30] used the maximum and minimum SM values of SMAP to limit the range of CYGNSS retrieval results, and the total root mean square error (RMSE) obtained for sites of interest was 0.04 cm^3/cm^3. Clarizia *et al.* [31] proposed an RVR algorithm, which uses CYGNSS reflectivity, roughness coefficient, and vegetation opacity (VO) in SMAP to perform a linear regression to obtain daily SM. The estimated RMSE is 0.07 cm^3/cm^3. Calabia *et al.* [32] proposed regional SM estimation using bistatic radar physical models and Fresnel reflection coefficients. An R-square of 0.6 and an RMSE of 0.05 cm^3/cm^3 were provided. Yan *et al.* [33] reported an effective schematic for estimating SM by utilizing the statistics of CYGNSS reflectivity. Then, the SM estimation was determined through the linear regression technique with an R of 0.80 and an RMSE of 0.07 cm^3/cm^3.

With the demand for higher accuracy and efficiency, different from the above traditional statistical regression methods, the intelligent retrieval based on machine learning (ML) algorithms has developed rapidly for CYGNSS SM estimation. Eroglu *et al.* [34] employed a fully connected artificial neural network (ANN) regression model to perform regional SM predictions through learning the nonlinear relations of SM and other land geophysical parameters to the CYGNSS observables. This learned network used eight input features. Three features are from CYGNSS data and the other five features are from other sources. As an extension work, Senyurek *et al.* [35] adopted three ML models, RF, ANN, and SVM, for comparison purposes. An overall RMSE value of 0.052, 0.061, and 0.065 cm^3/cm^3 are achieved for the RF, ANN, and SVM techniques, respectively. The RF method performed best and was then adopted to show the performance of CYGNSS-based SM estimates involving SMAP data. Mean unbiased ubRMSE of 0.055 and 0.054 cm^3/cm^3 were obtained with CYGNSS estimates and SMAP against *in situ* observations, respectively, with a higher R with the CYGNSS retrievals [36]. The number of ancillary data that are from other sources was not reduced [35], [36]. Yang *et al.* [37] also used backpropagation (BP)-ANN to compare and evaluate the SM estimation performance of the two spaceborne GNSS-R satellite missions (TDS and CYGNSS), which has quite a few (six) ancillary variables. The results showed that TDS-1 and CYGNSS agree and correlate very well with the SMAP SM in Mainland China.

The majority of the previous CYGNSS-based SM retrieval studies considered SMAP data as the reference SM and validated their performance with SMAP or point-scale *in situ* observations. Some of them achieved very high resolutions. However, it is difficult to directly compare the CYGNSS-based SM data products from these studies against each other because each paper shows differences in (1) time spans, (2) the number of data samples and spatial coverage, (3) validation and reference data sets, (4) assumptions regarding gridding, (5) ancillary data used, and (6) spatial resolutions. These factors all impact the performance of SM retrieval. Despite these differences, most approaches have shown a moderate performance in CYGNSS-based SM retrieval. Relevant information about the abovementioned SM estimates at a spatial resolution of 36 km is summarized in Table I.

Up until now, CYGNSS has proved to be a powerful tool for producing accurate SM estimates when applied at small and medium scales. However, alternative approaches are necessary for the following reasons: (1) Due to a large amount of input auxiliary data, the existing ML models are difficult to implement and are time-consuming; (2) there remains a need for a comprehensive large-scale SM estimation method based on ML since larger amounts of data put even stricter requirements on algorithm performance and efficiency; (3) although ML algorithms are commonly characterized by good universality and transferability, a single ML framework may not address all problems. Each algorithm has different advantages and applicable data structures that have to be examined.

This article proposes a novel preclassification aided strategy for global CYGNSS SM estimation. The performances of different ML regression models with and without the strategy are compared. Using the least ancillary data, the performance of the proposed ML regression with a preclassification approach is evaluated for all types of land surfaces, exhibiting its simple and effective property. This article is organized as follows. Section II describes the employed CYGNSS and SMAP data and observables. Section III presents the design of the proposed ML model. Section IV shows the results for SM estimation with analyzing on the spatio-temporal heterogeneity. Section V summarizes the conclusions.

II. Data Sets and Observables

A. Satellite Data

The CYGNSS mission has been a topic of interest since it was launched in Dec. 2016 [38]. The CYGNSS constellation

TABLE I
APPLICATIONS OF CYGNSS SM ESTIMATION METHODS WITH A RESOLUTION OF 36 KM

Source	Time Span	Spatial coverage	Reference SM	Validation SM	Adopted algorithms	Overall performance (cm^3/cm^3)	Spatial Resoluations
Chew and Small (2018)	1	Global	SMAP	SMAP, in situ	Multiple linear regression	0.045 (ubRMSD) (CYGNSS vs SMAP)	36km × 36 km
Clarizia et al. (2019)	0.5	Global	SMAP	SMAP	Physical, Trilinear regression	0.07 (RMSE) (CYGNSS vs SMAP)	36km ×36 km
Chew and Small (2020)	3	Global	SMAP	ISMN sites	Multiple linear regression	0.049 (ubRMSE), R=0.40 (CYGNSS vs in situ)(median) 0.045 (ubRMSE), R=0.69 (SMAP vs in situ)(median)	36km ×36 km
Yang et al. (2020)	2	Regional	SMAP	SMAP and in situ networks	BP-ANN	0.062 (ubRMSD), R=0.79 (CYGNSS vs SMAP) 0.053 (ubRMSE), R=0.72 (CYGNSS vs in situ)	36km × 36 km
Yan et al. (2020)	1	Global	SMAP	SMAP	Linear regression	0.07 (RMSE), R=0.80 (CYGNSS vs SMAP)	36km × 36 km
Proposed method	2 years	Global	SMAP	SMAP and in situ networks	XGBoost with pre-classification strategy	0.052 (RMSE), R=0.86 (CYGNSS vs SMAP) 0.049 (ubRMSE), R=0.753 (CYGNSS vs in situ) (median) 0.046 (ubRMSE), R=0.823 (SMAP vs in situ) (median)	36km ×36 km

is composed of eight micro-satellites, and each satellite can acquire data from four specular reflection points on the Earth's surface simultaneously. The orbit of the constellation is in the middle and low latitudes, and the specular reflection points are distributed within the range of ±37° (latitude). Therefore, it can provide data with extensive spatial coverage that can reach up to 7×0.5 km and high temporal resolution [28]. In terms of temporal resolution, despite the variety of other global SM satellite products, the revisit time of satellites and orbits is generally limited to 2–5 days or even longer. The revisit period of CYGNSS is effectively shortened by eight low-inclination satellites. With all satellites working, it can currently achieve a temporal resolution of several hours and make almost full global coverage in 1 day.

CYGNSS provides three levels of scientific data products to the public. The CYGNSS data used in this work is Level 1 (L1), version 2.1 data. It provides free access to metadata[1] that contains the bistatic radar cross section (BRCS or σ), and signal-to-noise ratio (SNR) as well as other geometry and navigation messages, such as the incident angle and the distance between the satellite and specular points.

The SMAP mission is an orbiting observatory that has measured the amount of water in the surface soil, since January 2015 [9]. It is a joint radar + radiometer operating at the L band (same band as CYGNSS). Radar ceased operation a few months after launch, but the radiometer is still working well. It provides measurements of the land surface SM and freeze–thaw state with near-global revisit coverage in almost 3 days. These data products are made available.[2] The L3 global daily Equal-Area Scalable Earth Grid (EASE-Grid) data with a spatial resolution of 36 km are employed [28], [31].

The SMAP daily data contain SM estimates, quality flags, roughness coefficient, VO, and other auxiliary information that can be gridded over the EASE-Grid. To facilitate the further

[1][Online]. Available: https://podaac.jpl.nasa.gov
[2][Online]. Available: https://nsidc.org/data/SPL3SMP

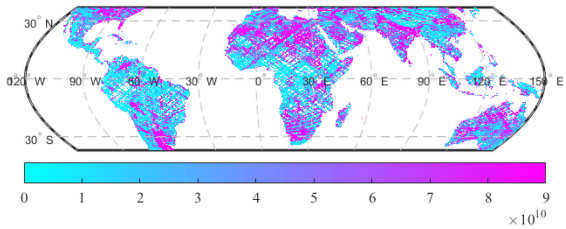

Fig. 1. CYGNSS plot figure (BRCS) before the data filtering on the day of 2020.1.1.

comparison and validation, the CYGNSS data (e.g., *brcs* in Fig. 1) are also resampled into 964 × 406 grids over the EASE-Grid, based on the longitude and latitude of the SMAP data and the CYGNSS observables at the specular reflection point [28], [31]. Thus, the SM data product is 36 × 36 km since it was trained with data at this resolution.

B. Data Quality Control

The total observation period was two years, from June 2018 to June 2020. The quality of CYGNSS and SMAP data was needed to be evaluated before modeling since the quality of data directly related to the performance of the ML predictions. Hence, the obtained CYGNSS observables calculated for each specular point (SP) acquisition and SMAP data gridded daily over the EASE-grid were first filtered according to the following rules: (1) The CYGNSS reflectivity had to be positive and smaller than 0.1 to remove the anomalies [33]; (2) the incident angle of CYGNSS data was above than 60° are commonly disregarded (a degradation in data quality often occurs at larger incidence angles [39]) [30], [37]; (3) the negative antenna gain in the direction of the specular point (corresponding to uncertainties reported in the measured antenna gain patterns) was removed [28],[34], [35], [37]; (4) to ensure that the error in the CYGNSS SP location estimation is within a reasonable range, only the BRCS data with a Delay Doppler Map (DDM) peak position

between the 5th and 11th bins in the delay axis were preserved [35], [40]; (5) the SMAP "retrieval successful/unsuccessful" quality flag was used as well to filter the SMAP data to ensure the quality of SM estimation [31], [41].

C. CYGNSS Data Observables

In this work, the CYGNSS data were employed to obtain surface reflectivity. The SMAP roughness coefficient as well as VO was taken as ancillary data for SM estimation, and SMAP SM was considered as the reference ground-truth data. It has been reported the method for obtaining the CYGNSS reflectivity Γ that was readily derived from CYGNSS *brcs* σ [see (1)] or an approximately substituting the CYGNSS *ddm_snr* into P^{coh} [see (2)] as shown below [34]–[40]:

$$\Gamma_{brcs} = \frac{\sigma(R_t + R_r)^2}{4\pi(R_r R_t)^2} \tag{1}$$

$$\Gamma_{snr} = \left(\frac{4\pi}{\lambda}\right)^2 \frac{P^{coh}(R_r + R_t)^2}{P_t G_t G_r} \tag{2}$$

where R_t and R_r are the distances from the transmitter and receiver to specular points, respectively. P^{coh} denotes the bistatic coherently received power. P_t is the transmitted signal power, G_t is the transmitter antenna gain, and G_r is the receiver antenna gain.

The CYGNSS reflectivity can be obtained from (1) or (2). Moreover, another observable Γ_{ratio} that was derived by the ratio of the reflected SNR (*ddm_snr*) and the direct SNR (*direct_snr*) was also obtained as a comparison [34], [42].

As mentioned before, the adopted SMAP data is published and distributed in the EASE-Grid format with a spatial resolution of 36 km. Therefore, when using SMAP data as ancillary data for estimating SM, the data of CYGNSS reflectivity are grided and resampled on the same EASE-Grid. In this case, each grid cell contains multiple CYGNSS samples and one SMAP sample, due to a higher spatial resolution of the CYGNSS data. In this work, the CYGNSS reflectivity in one grid was averaged and used to calculate the SM together with the SMAP data.

III. Design of SM Estimation Using ML Models

A. Typical Traditional ML and Deep-Learning Algorithms

ML is used to attempt to construct intrinsically nonlinear relationships between input and output data [43], [44]. ML can automatically learn and evolve as the amount of available data increases. ML algorithms are not entirely based on rules. As experience progresses, they learn to give specific answers by evaluating large amounts of data.

Random Forest (RF) is one of the most popular and powerful ML algorithms. It is a type of ensemble ML algorithm called Bootstrap Aggregation or bagging, proposed by Breiman [45]. Its performance can compete with the popular boosting-based gradient boosting regression tree (GBRT) algorithm. RF is easy to parallelize and implement in "big data." Due to the use of random sampling, the trained model has a small variance and strong generalization ability. However, it also has some disadvantages: a) Over-fitting may occur in classification or regression problems when data sets exhibit relatively large noise; b) the number of features is likely to have a greater impact on decision-making, thereby affecting the performance of the fitted model.

The support vector machine (SVM) was established by Vapnik [46] on the basis of statistical learning theory. It is a typical ML algorithm, which was originally used for classification. Moreover, the SVM not only performs well in the classification but can also be used as a typical solution to the regression problem. It requires fewer samples and it is quite efficient. However, for large-scale training samples, SVMs may not be effective, and they are sensitive to the choice of parameters and kernel functions [47].

Extreme gradient boosting (XGBoost) is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements the ML algorithm under the gradient boosting framework. XGBoost provides a parallel tree boosting (also known as GBRT) that solves many data science problems in a faster and more accurate way. XGBoost is essentially a GBRT, but strives to maximize speed and efficiency. However, since boosting is naturally executed sequentially, they are difficult to parallelize [45].

ANN is a typical example in the traditional neuron network (NN) framework [45]. An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons. It is widely used in various disciplines and engineering fields since it has good function approximation performance and strong robustness. Meanwhile, computational and space complexity is quite high.

B. SM Estimation Using ML Regression Aided By the Preclassification Strategy

With the assumption that the signal over land is predominantly determined by the coherent reflection from the surface and is eventually reduced by the roughness and the vegetation, the SM estimation was obtained by considering the data product of the reflectivity, roughness, and vegetation [31]. The CYGNSS reflectivity was taken as the main predictor in the SM estimation model, and the SMAP roughness coefficient $\delta$and VO τ were employed as ancillary data.

The previous study [32]–[37] has tried to add alternative ancillary data to improve the accuracy of SM estimates. We found that most of the added ancillary data are related to terrain, such as topography and soil texture [32]–[37]. These ancillary data have shown the ability to improve the accuracy of the estimates but are heavy-loaded and technically difficult to obtain, especially in the case of global estimates. Hence, we proposed a novel preclassification strategy that employs resampling and submodeling procedures based on the traditional ML regression approach to minimize the influence of different land types and improve the SM estimation accuracy in an easy and practical way. The flowchart showing the training and validation for CYGNSS SM estimation by using ML regression with the preclassification strategy approach is shown in Fig. 2.

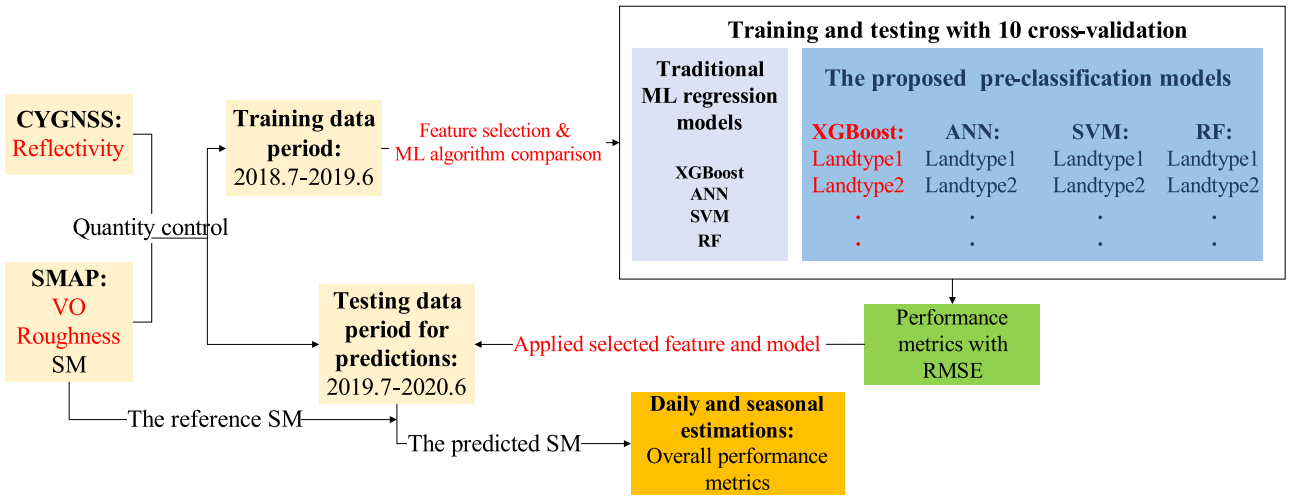

Fig. 2. Flowchart showing the training and validation for the CYGNSS SM estimation with the proposed approach.

Two years of CYGNSS and corresponding SMAP data were used for the analysis. The first-year data were used to build the model which was then used to predict the appropriate SM with the second-year data. The surface reflectivity, derived from CYGNSS BRCS σ or other variables, was calculated and regarded as the primary input with the other vectors (δ and τ) of the model (Fig. 2). The SMAP SM values were used as the output of the model and were also taken as the reference data for training and verifying the proposed ML approach.

The proposed preclassification strategy procedure that estimated SM by land type was compared with traditional ML regressions. The preclassification methodology trains a separate retrieval algorithm corresponding to each of the available IGBP land types, instead of presenting a single SM retrieval algorithm. When it refers to "traditional ML regressions," it means a single retrieval algorithm for all land types. First, the overall samples of different land types were grouped according to the International Geosphere-Biosphere Programme (IGBP) land-type classification provided by SMAP and then they were used to build several submodels (e.g., for land types 1, 2, …) for SM estimation.

The 10-fold cross-validation (CV) was adopted to train and verify the feasibility of the proposed regression with preclassification modeling and to select the optimal feature and algorithm. The 10-fold CV is commonly used and popular, it generally results in a less biased model compare to other methods. Because it ensures that every observation from the original dataset has the chance of appearing in the training and test set. The whole dataset is randomly partitioned into 10 folds (depending on the data size). Then we fit the model with nine folds that are used as a training set, and validate the model using the remaining set. Note down the RMSE as the performance metric. Repeat this process until every 10-fold has been served as the test set. The final evaluation performance metric is the average result of the recorded RMSE in each iteration. In general, the RMSE was calculated and showed for an entire dataset, except for the distribution map showing the performance of each pixel.

We extracted over 10 million groups of daily samples of CYGNSS and SMAP for the period ranging from July 2018 to June 2019. We shuffled the order of each group of data to ensure the accuracy of the 10-fold CV procedure. Additionally, sufficient training data were needed for obtaining a stable model. This also ensured that each submodel had enough data for training and testing (each submodel contained at least 300 thousand groups of data).

TABLE II
RMSE OF SM ESTIMATED BY XGBOOST MODEL USING 10-FOLD CV WITH DIFFERENT OBSERVABLES

RMSE (cm^3/cm^3)	Type 1: Γ_{brcs}	Type 2: Γ_{snr}	Type 3: Γ_{ratio}
	0.064	0.069	0.070

As shown in Fig. 2, the CYGNSS data of one year (between June 2018 and June 2019) under the 10-fold CV was used to evaluate the performance of the SM estimation model. Then, the optimal performing feature and model were applied to the SM prediction with "unseen" data (from July 2019 to June 2020) to validate its generalization ability and conduct further analysis. The SM predictions were compared with the reference SMAP data to verify the performance of the established model and to investigate the spatio-temporal variation of the SM estimates.

C. Examination of Different Input Observables

Reflectivity is the primary CYGNSS observable related to SM, which is taken as one of the input predictors in the regression model. Several different methods have been described in the previous section to compute the reflectivity. Combined with different deviations of reflectivity, three types of input vectors were investigated using the 10-fold CV to compare the performance of SM estimation: (1) $\Gamma_{\mathrm{brcs}}+\delta+\tau$; (2)$\Gamma_{\mathrm{snr}}+\delta+\tau$; (3) $\Gamma_{\mathrm{ratio}}+\delta+\tau$.

As an example shown in Table II, three types of observables were taken as inputs to build the XGBoost regression model, respectively. The Γ_{brcs} performed better than Γ_{snr} and Γ_{ratio}, which agrees with [34]. The reason could be the diverse levels

TABLE III
EVALUATION PERFORMANCE (RMSE) FOR SM ESTIMATION USING 10-FOLD CV AND DIFFERENT ML WITH/WITHOUT PRECLASSIFICATION STRATEGY

Land type	RMSE (cm^3/cm^3)							
	XGBoost		RF		ANN		SVM	
	No Pre-cla.	With Pre-cla.	No Pre-cla.	With Pre-cla.	No Pre-cla.	With Pre-cla.	No Pre-cla.	With Pre-cla.
Evergreen Broadleaf Forest	0.0884	0.0822	0.0918	0.0869	0.0901	0.0889	0.0925	0.0903
Deciduous Broadleaf Forest	0.0868	0.0718	0.0881	0.0759	0.0922	0.0798	0.0939	0.0879
Mixed Forest	0.0730	0.0613	0.0662	0.0653	0.0748	0.0723	0.0788	0.0782
Open Shrublands	0.0468	0.0423	0.0454	0.0441	0.0493	0.0488	0.0694	0.0650
Woody Savannas	0.0725	0.0657	0.0721	0.0657	0.0745	0.0725	0.0776	0.0757
Savannas	0.0706	0.0612	0.0712	0.0686	0.0723	0.0673	0.0768	0.0728
Grasslands	0.0720	0.0603	0.0668	0.0651	0.0778	0.0682	0.0805	0.0746
Croplands	0.0716	0.0635	0.0721	0.0689	0.0729	0.0717	0.0734	0.0707
Cropland/Natural Vegetation Mosaic	0.0761	0.0641	0.0731	0.0669	0.0779	0.0742	0.0835	0.0768
Barren or Sparsely Vegetated	0.0435	0.0370	0.0425	0.0415	0.0460	0.0455	0.0660	0.0700
Final results	0.064	0.052	0.063	0.060	0.070	0.063	0.079	0.068

of errors coming from changing calibration parameters in different reflectivity calculations. Hence, the term "reflectivity" was referred to Γ_{brcs} in the following sections.

Except for the input vector, as introduced previously, different ML algorithms search for different trends and patterns. The prediction performance directly depends on the model that was chosen. One algorithm is not the best across all data sets or for all use cases. So, it is so important to know how to match an ML algorithm to a particular problem. To select and know what kind of algorithm works best for the type of SM regression problem is quite critical. Hence, different ML regression models are built and validated also respectively by a 10-fold CV technique to access the performance of the SM prediction.

IV. RESULTS AND ANALYSIS

A. Comparisons of Different ML Regression Models With/Without a Preclassification Strategy

The SM estimates from the different traditional ML (RF, SVM, XGBoost) and DL (ANN) algorithms with and without preclassification strategies are shown in Table III. The optimal input vector ($\Gamma_{\mathrm{brcs}}+\delta+\tau$) was used to determine the best ML modeling method (Table III). The behavior of the proposed ML regression aided by preclassification was demonstrated in detail by using annual data to build the model. The result was also compared with the traditional regression models. The final evaluation results for the preclassification strategy are the weighted average results of each sub-model and show the overall good performance of the SM global estimation, indicating a clear drop in RMSE when using the preclassification strategy.

According to the IGBP land classification provided by SMAP, the CYGNSS samples contain a total of seventeen categories. Due to the quality control procedure for the CYGNSS and SMAP data, the number of data for certain categories were far from enough for building ML models. Therefore, (seven) categories with a data volume of less than 20 000 throughout the year were excluded and were not used in the modeling. Besides, according to the statistical results, these categories that were excluded are water bodies, permanent wetlands, ice, and snow, or areas with extremely thick vegetation, which is difficult to retrieve SM in these areas with current techniques.

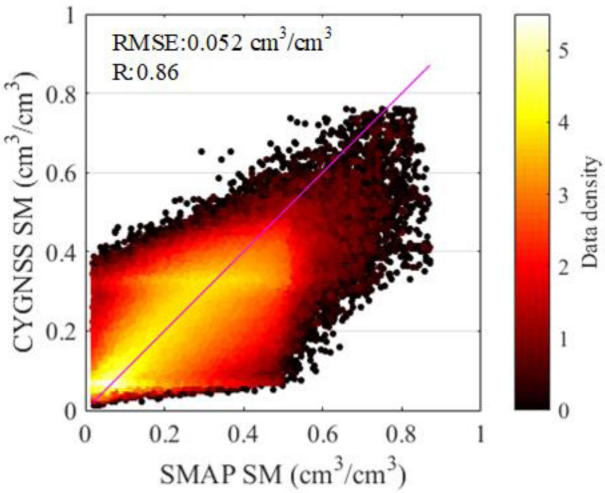

Fig. 3. An example of density plots (in log-scale) comparing CYGNSS and SMAP SM by using the proposed XGBoost with the 1:1 reference line.

The presented preclassification (aided by submodeling) strategy method yielded good results with a smaller RMSE in all algorithms (traditional ML and DL) and also in all land types. Moreover, comparing different ML algorithms, the RF outperformed ANN and SVM, which agrees with [35]. Furthermore, the XGBoost had the best performance with the least RMSE 0.052 cm^3/cm^3, which has not yet been reported on a global scale SM estimation.

In Fig. 3, an example of the density plots in log-scale showing the comparison between estimated CYGNSS SM and the reference SMAP SM data is presented. The density plot is shown to exhibit the performance of SM estimation using XGBoost with a preclassification strategy. The numbers of data for the first year with 10 million samples in total are shown. The density plot highlights an overall fairly good consistency between CYGNSS SM and SMAP SM, especially when the data are the densest. Each data cloud in density plots is centered along the 1:1 line. However, where the data density is lower, a tendency to deviate from the line is displayed. The slope shows a downward trend, which is less than 1, meaning that the CYGNSS tends to

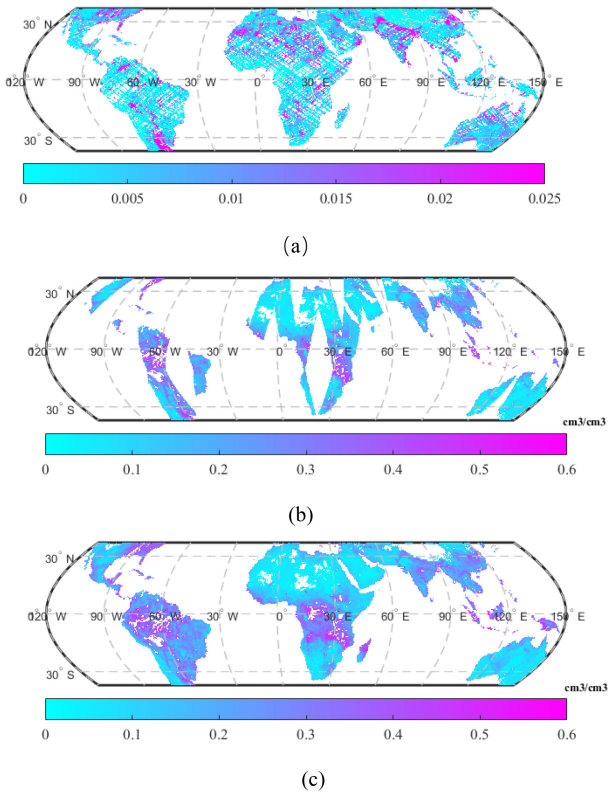

Fig. 4. Spatial coverage of CYGNSS observables and SMAP SM. (a) 1-day CYGNSS reflectivity, (b) 1-day SMAP SM on 2020.1.1, and (c) 3-day SMAP SM on 2020.1.1–2020.1.3.

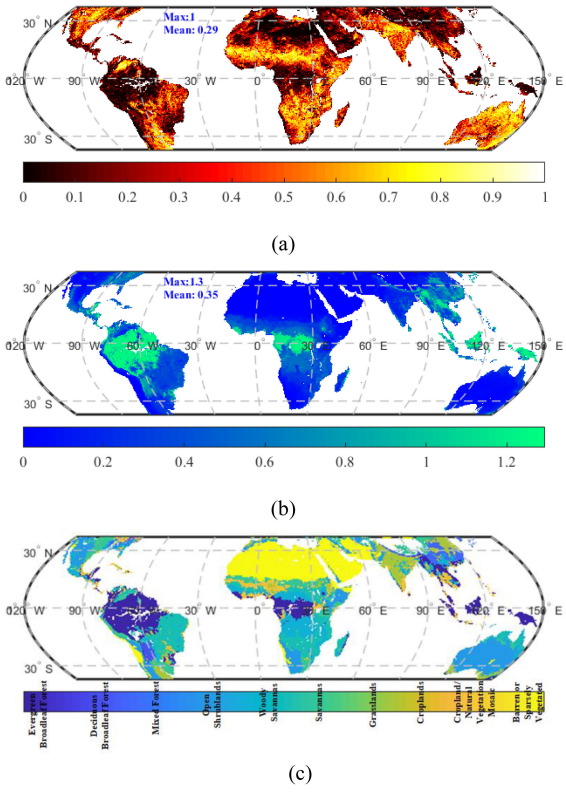

Fig. 5. CYGNSS predictors sensitivity to SMAP SM. (a) The spatial correlations ($R_{\text{reflectivity}}$) of CYGNSS reflectivity and SMAP SM. (b) The daily average of VO. (c) The dominant land type of each grid cell on 2020.1.1–2020.1.4.

underestimate SMC values to some degree. This phenomenon was also reported in [31] and explanation could be investigated in future work. In the next section, the XGBoost with preclassification strategy is adopted as the optimized prediction model (RMSE $= 0.052$ cm^3/cm^3) to demonstrate the performance of SM estimation on the aspect of spatio-temporal heterogeneity.

B. Sensitivity of CYGNSS Predictors to SM Estimation

The sensitivity analysis of CYGNSS predictors to SM estimation was analyzed in detail to investigate the relationship of the input predicting variables with SM estimations. Fig. 4 presents 1-day CYGNSS reflectivity, 1-day SMAP SM, and 3-day SMAP SM mapped into the EASE-Grid, respectively. The examples are based on the data of 2020.1.1 for illustration purposes.

The 1-day CYGNSS reflectivity in Fig. 4(a) was obtained from (1), and the spatial coverage was higher than that of 1-day SMAP SM [Fig. 4(b)]. Compared with Fig. 1, some grid cells not covered with CYGNSS data are present in Fig. 4(a), since some data has been removed due to insufficient data quality. The data in Fig. 4(a) still provided global coverage. Since the SMAP satellite can provide full global coverage within three days, the 3-day SMAP SM in Fig. 4(c) almost covers a global area, providing sufficient reference data for modeling and SM estimation. Most areas with higher CYGNSS reflectivity in Fig. 4(a) correspond to a higher SM in SMAP (Fig. 4c), and vice versa.

In order to further validate the capability of CYGNSS reflectivity for SM estimation, the spatial correlation ($R_{\text{reflectivity}}$) is calculated as the distribution of correlation coefficient of the CYGNSS reflectivity and SMAP SM for each grid cell (July 2019 to June 2020), which is shown in Fig. 5(a). It is clear that the consistency varies over different regions. High $R_{\text{reflectivity}}$ shows a strong correlation between CYGNSS reflectivity and SMAP SM for each grid, which agrees with the phenomenon observed in Fig. 4(a) and (c). Meanwhile, high correlations are observed in most of the areas, demonstrating that CYGNSS reflectivity plays the main role in SM estimation, which is consistent with the results of [31].

The daily average VO τ is also shown for further investigation in Fig 5(b). The areas with very high τ (bright green), e.g., the tropical areas of South America, the Congo Basin in Africa, and Southeast Asia, correspond to locations with very low $R_{\text{reflectivity}}$ in Fig 5(a). In addition to this, the other high τ areas (e.g., dark green) in Fig. 5(b) also correspond to regions with low $R_{\text{reflectivity}}$ in Fig 5(a). Hence, we infer that the $R_{\text{reflectivity}}$ is negatively correlated to the VO τ. This phenomenon can be explained by the fact that a higher τ leads to more diffuse scattering and weakens the signal reception of CYGNSS reflectivity [31]. In this case, the τ has a large weight among the input variables.

Few exceptions (e.g., the Sahara Desert) showing a very low τ were noticed [see Fig. 5(b)] corresponding to a low $R_{\text{reflectivity}}$

TABLE IV
STATISTICAL ANALYSIS OF PREDICTORS WITH DIFFERENT LAND TYPES

Land type	$R_{reflectivity}$	Vegetation Opacity	Roughness coefficient
Evergreen Broadleaf **Forest**	0.215	1.079	0.156
Deciduous Broadleaf **Forest**	0.263	0.751	0.155
Mixed **Forest**	0.404	0.706	0.152
Open Shrublands	0.329	0.061	0.113
Woody **Savannas**	0.312	0.507	0.128
Savannas	0.397	0.350	0.151
Grasslands	0.359	0.126	0.150
Croplands	0.569	0.239	0.112
Cropland/Natural Vegetation Mosaic	0.310	0.384	0.131
Barren or Sparsely Vegetated	0.176	0.001	0.149

TABLE V
STATISTICAL ANALYSIS OF DAILY SM ESTIMATION PERFORMANCE WITH DIFFERENT LAND TYPES

Land type	SM (cm^3/cm^3)	RMSE (cm^3/cm^3)	R_{sm}	S.D. SMAP (cm^3/cm^3)	τ	S.D. τ
Evergreen Broadleaf **Forest**	0.315	0.087	0.509	0.099	1.079	0.212
Deciduous Broadleaf **Forest**	0.206	0.075	0.706	0.103	0.751	0.143
Mixed **Forest**	0.257	0.060	0.736	0.088	0.706	0.145
Open Shrublands	0.090	0.043	0.639	0.056	0.061	0.040
Woody **Savannas**	0.246	0.068	0.713	0.097	0.507	0.110
Savannas	0.170	0.063	0.728	0.091	0.350	0.092
Grasslands	0.151	0.060	0.768	0.093	0.126	0.100
Croplands	0.213	0.066	0.779	0.104	0.239	0.124
Cropland/Natural Vegetation Mosaic	0.217	0.065	0.803	0.109	0.384	0.141
Barren or Sparsely Vegetated	0.080	0.036	0.633	0.047	0.001	0.006

in Fig. 5(a). This area corresponds to the barren or sparsely vegetated land type, as shown in Fig. 5(c). In Table IV, compared with the other land types, the $\tau$is quite small and the $R_{\text{reflectivity}}$ is greatly decreased, but the roughness coefficient δdid not change much. More importantly, a quite low SM (light blue) can be observed in this place in Fig. 4(c). In this case, it could be speculated that roughness may have a larger weight than τimpacting the CYGNSS reflectivity reception.

This phenomenon can be seen in the detailed statistics reported in Table IV. The $R_{\text{reflectivity}}$, τ, and $\delta$are shown and classified by land types. The IGBP land types with a data volume of less than 20 000 throughout the year were excluded from the modeling.

In forest regions (evergreen broadleaf, deciduous broadleaf, and mixed forest), the $R_{\text{reflectivity}}$ increased evidently with the decrease of τ, which agrees with the previous inference and the fact that vegetation mainly reduces coherent reflection [31]. In most of the areas, $R_{\text{reflectivity}}$ was negatively correlated with τ. This rule was found also in the savannas (woody savannas and savannas) and cropland (cropland and natural vegetation mosaic) regions. Moreover, there were variations to the rule of δ with respect to $R_{\text{reflectivity}}$ in those three regions and no pattern emerged, such as in the savannas region. The δ increased with the increases of $R_{\text{reflectivity}}$, which is opposite in the cropland region. This implies that the variation of δ did not impact the $R_{\text{reflectivity}}$ and thus δ had a much smaller weight than τ. Hence, in addition to reflectivity, vegetation played the main role in SM estimation; roughness had a much lower weight, as expected for most land types. Furthermore, as we have mentioned before, barren or sparsely vegetated areas show a very low τ corresponds to a low $R_{\text{reflectivity}}$. Sharp decreases were observed in VO [see Fig. 5(b) and Table IV], and the roughness coefficient did not change much compared to other cases (see Table IV). The $R_{\text{reflectivity}}$ was the lowest among different land types as the same to the very low SM [see Fig. 4(c) and Table V]. In this case, the reason can be explained by that the roughness in a very dry region (e.g., with a big sandhill or peculiar terrain) may become the main factor impacting the reflectivity reception among the predictors.

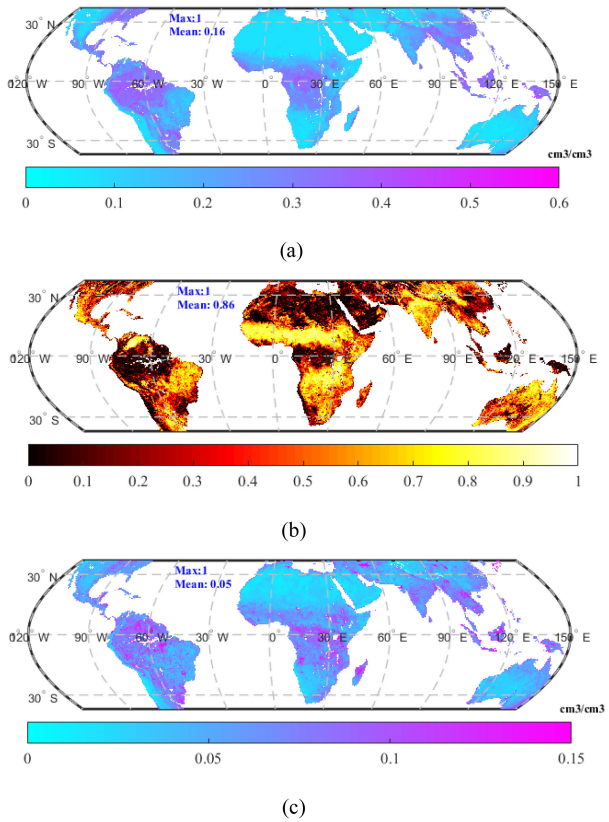

Fig. 6. Prediction performance: (a) Distribution of predicted daily averaged CYGNSS SM by the proposed XGBoost model, (b) R_{sm}, and (c) RMSE considering reference SMAP SM on global coverage.

C. Daily SM Estimation Performance

In this section, the CYGNSS SM predicted by the presented XGBoost model was compared with SMAP SM to demonstrate the applicability and feasibility of the proposed approach on a daily basis. The XGBoost prediction model was constructed by training the data of the first year (July 2018–June 2019), while the data of the second year (July 2019–June 2020) was used as the testing set. Fig. 6 shows an example of daily averaged CYGNSS SM predicted by the proposed model with global

coverage, with the corresponding RMSE and spatial correlation (or, R_{sm}) between the predicted CYGNSS SM and reference SM for each grid.

In Fig. 6(a), the predicted CYGNSS SM captures the main macroscopic features of the SMAP SM. The strong relationship of the predicted SM and SMAP SM can be observed also in Fig. 6(b), showing the high spatial correlation $R_{sm,}$ which is calculated as the correlation coefficient of the predicted CYGNSS SM and reference SM. Moreover, the pattern of R_{sm} is in good agreement with $R_{\text{reflectivity}}$ [Fig. 5(a)], confirming that reflectivity plays the main role in SM estimation. More than that, it has to be noted that higher values (brighter colors) in Fig. 6(b) than in Fig. 5(a) are present for each grid, indicating a stronger relationship for R_{sm} than $R_{\text{reflectivity}}$. Hence, it is conceivable that the adopted estimation model is capable of SM predictions and the other input predictors (including τ and δ) still produce improvements in SM predictions.

A higher RMSE is related to higher τ [see Figs. 5(b) and 6(c)], validating the conclusion that higher vegetation weakens the signal reception, impacting SM estimation and that vegetation is an important predictor. Other deviations could be the incidence angle of the specular point, mean elevation for each specular point, and the slope of the trailing edge of the reflectivity [34]–[37]. We also noticed that, in most areas, high SM was always accompanied by higher RMSE [see Fig. 6(a) and (c)]. The phenomenon can be further corroborated by the statistics in Table V.

The statistics of daily estimations are shown on the basis of predicted average SM, R_{sm}, and RMSE in Table V with the standard deviations (SD) of SMAP SM. Overall values of RMSE 0.056 cm^3/cm^3 and R_{sm} 0.86 were attained. Besides, a high spatial correlation R_{sm} always appears with a high value of SMAP SD, which agrees with the finding by [49]. Moreover, the averaged VO and the SD of VO are summarized and shown. The areas with high values of SD of $\tau$and τ are usually accompanied by high RMSE (e.g., forest region).

Furthermore, three region types were summarized for the analysis: sparse vegetation (e.g., open shrublands and barren or sparsely vegetated), dense forest (e.g., broadleaf forest areas), and moderately vegetated regions (e.g., grassland), according to the IGBP land types shown in Fig. 5(c) and Table V. We concluded that 1) for sparsely vegetated areas ($\tau$is very low), SM, RMSE, and the R_{sm} are all very low; 2) For forest areas, the τ and SM are very high, which leads to very high RMSE and the R_{sm} increases; 3) for grassland areas, the SM and the τ are in the middle of the former two cases, the RMSE is also between the two but the R_{sm} is a little higher. This rule can be summarized and confirms the conclusion that the τ is one of the main actors impacting the accuracy of SM predictions, which, as they increase, will lead to a higher RMSE. Also, the high SM always yields high RMSE, which has been mentioned with the comparison of Fig. 6(a) and (c).

At the same time, a decrease of SM decreases the soil permittivity and thus the Fresnel coefficient. It was observed that a very low SM corresponds to a low R_{sm} even when the τ is quite low (decreased diffuse scattering). One reason could be that a very low SM (low Fresnel coefficient) is more vulnerable to be influenced by other factors. The small variation of other input predictors (e.g., roughness) or other random factors will generate considerable impacts and errors, thus affecting the signal receptions. This could be another explanation for the exception observed between $R_{\text{reflectivity}}$ and τ before. Another possible reason could be due to the mechanism of the ML algorithm, since the values of samples for the low SM region hardly changed, which is not favorable to ML modeling and predictions. Hence, in some extreme cases (very dry regions), SM is also another limiting condition affecting spatial correlation *R*.

TABLE VI
DAILY AVERAGED VEGETATION OPACITY IN THE SOUTHERN AND NORTHERN HEMISPHERE

Vegetation opacity	1st season	2nd season	3rd season	4th season
Northern Hemisphere	0.259	0.251	0.248	0.238
Southern Hemisphere	0.343	0.354	0.373	0.364

D. Seasonal SM Estimation Performance

The seasonal predicted SM distributions obtained in four seasons using the XGBoost model are presented in Fig. 7. The distributions of the SM predictions are similar to each other. The CYGNSS SM captures the main macroscopic features of the SMAP SM. Further details of predicted SM and its related performance matrix for each grid cell are demonstrated in the temporal domain.

The predicted time series of daily CYGNSS SM, RMSE, and temporal correlations (or, T_{sm}) that are calculated as the correlation coefficients between the CYGNSS SM and SMAP SM for all grid cells in the northern and southern hemispheres are shown in Fig. 8. Due to the different seasonal distribution between the two hemispheres, the predicted seasonal SM time series were considered separately. The average predicted SM values in the four seasons are presented in Fig. 8(a) and (b). The time series was divided into four quarters, corresponding to autumn, winter, spring, and summer seasons with respect to the northern hemisphere. The SM predictions agree with SMAP SM and the trend of one-year SM predicted by CYGNSS in the northern hemisphere is reversed with the southern hemisphere, which conforms to the expectations.

The average one-year time series of RMSE and temporal correlation T_{sm} that is calculated as the correlation coefficient between daily CYGNSS SM and SMAP SM are also shown in Fig 8(c) and (d). From Fig. 8(d), it appears that there is a higher T_{sm} [see Fig. 8(d)] in the first and second seasons in both hemispheres. From Fig. 8(c), the RMSEs in the southern hemisphere, which are higher than in the northern hemisphere, increase with the seasonal variability (from the first season to the third season). This can be explained by Table VI, since the VO, which is higher in the southern hemisphere, also increases with time. Hence, higher RMSE always appears with high τ, which also confirms the conclusion that vegetation plays an important role in SM estimation except for CYGNSS reflectivity.

The statistical analysis for seasonal T_{sm} and RMSE is summarized and shown in Table VII, classified by IGBP land types

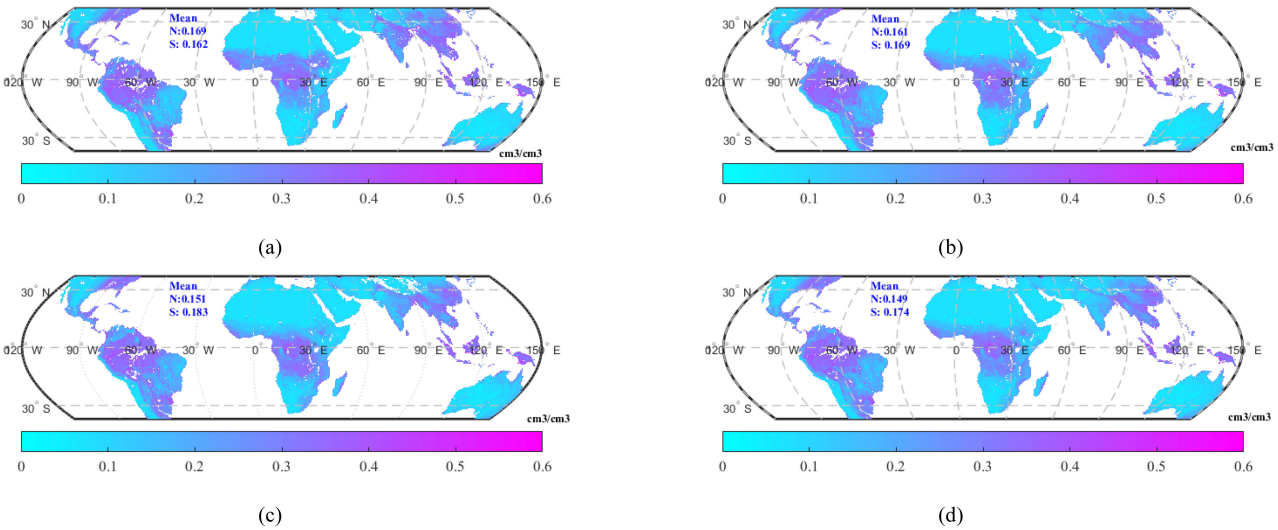

Fig. 7. Distribution of predicted seasonal CYGNSS SM. (a) First season (b) Second season. (c) Third season. (d) Fourth season.

TABLE VII
STATISTICAL ANALYSIS OF SEASONAL SM ESTIMATION PERFORMANCE WITH DIFFERENT LAND TYPES

	RMSE (cm^3/cm^3)				R_{sm}			
Time distributions / Land type	1st season	2nd season	3rd season	4th season	1st season	2nd season	3rd season	4th season
Evergreen Broadleaf Forest	0.082	0.082	0.092	0.086	0.535	0.524	0.497	0.492
Deciduous Broadleaf Forest	0.056	0.076	0.084	0.077	0.793	0.708	0.663	0.703
Mixed Forest	0.057	0.063	0.057	0.060	0.637	0.713	0.804	0.760
Open Shrublands	0.034	0.041	0.052	0.040	0.663	0.649	0.636	0.626
Woody Savannas	0.061	0.066	0.070	0.070	0.797	0.722	0.684	0.661
Savannas	0.052	0.063	0.069	0.062	0.849	0.727	0.643	0.696
Grasslands	0.060	0.058	0.063	0.057	0.797	0.774	0.718	0.777
Croplands	0.073	0.062	0.060	0.069	0.741	0.776	0.802	0.777
Cropland/Natural Vegetation Mosaic	0.073	0.062	0.057	0.069	0.696	0.816	0.852	0.779
Barren or Sparsely Vegetated	0.031	0.037	0.041	0.033	0.641	0.646	0.634	0.624

and four quarters in one year. Overall values of RMSE 0.056 cm^3/cm^3 and R_{sm} 0.86 can be achieved as the same with daily SM estimations. For most land types, in the third and fourth quarters, the RMSE is high, and T_{sm} is low. Very few land types show high RMSE in the first and second quarters, such as croplands, which could be that the cropland presents variability from cultivation practices.

E. Validation of CYGNSS-Based SM Estimation Results With In Situ Measurements

Although a similar CYGNSS-based SM model CV method using SMAP SM has been used in other publications [27], [28], [30],[31],[33], it is better to involve other data sources to cross-validate the SM results. Given the limitation of the dependence of SMAP data during the training and prediction stage of the model, an independent SM source was added. The CYGNSS-based SM product obtained by the proposed model was compared against *in situ* measurements derived from dense SM ground networks from Mainland China for further validation. The *in situ* SM validation data set was collected by China's automatic SM observation stations throughout the year of 2018. Given that each site provides hourly SM measurements from 0 to 100-cm depth below the soil surface with an interval of 10 cm, the SM data with the top 10 cm of soil in each day are utilized and regarded as the ground-truth value in this study [50]. For quality control purposes, some unrealistic SM values (e.g., SM $<$ 0 cm^3/cm^3 or SM $>$ 1 cm^3/cm^3) were removed before validation.

The overall and regional results of both the proposed CYGNSS-based and SMAP SM products against *in situ* SM sites from five networks are shown in Table VIII. Performance metrics such as RMSE and the unbiased-RMSE (ubRMSE) are shown to facilitate the comparison with other studies [29].

In Table VIII, the overall median ubRMSE for CYGNSS against *in situ* SM (0.049 cm^3/cm^3) is similar to SMAP (0.046 cm^3/cm^3). A similar performance can also be seen from the median RMSE with the results of CYGNSS-based SM against *in situ* data (0.059 cm^3/cm^3) and SMAP against *in situ*

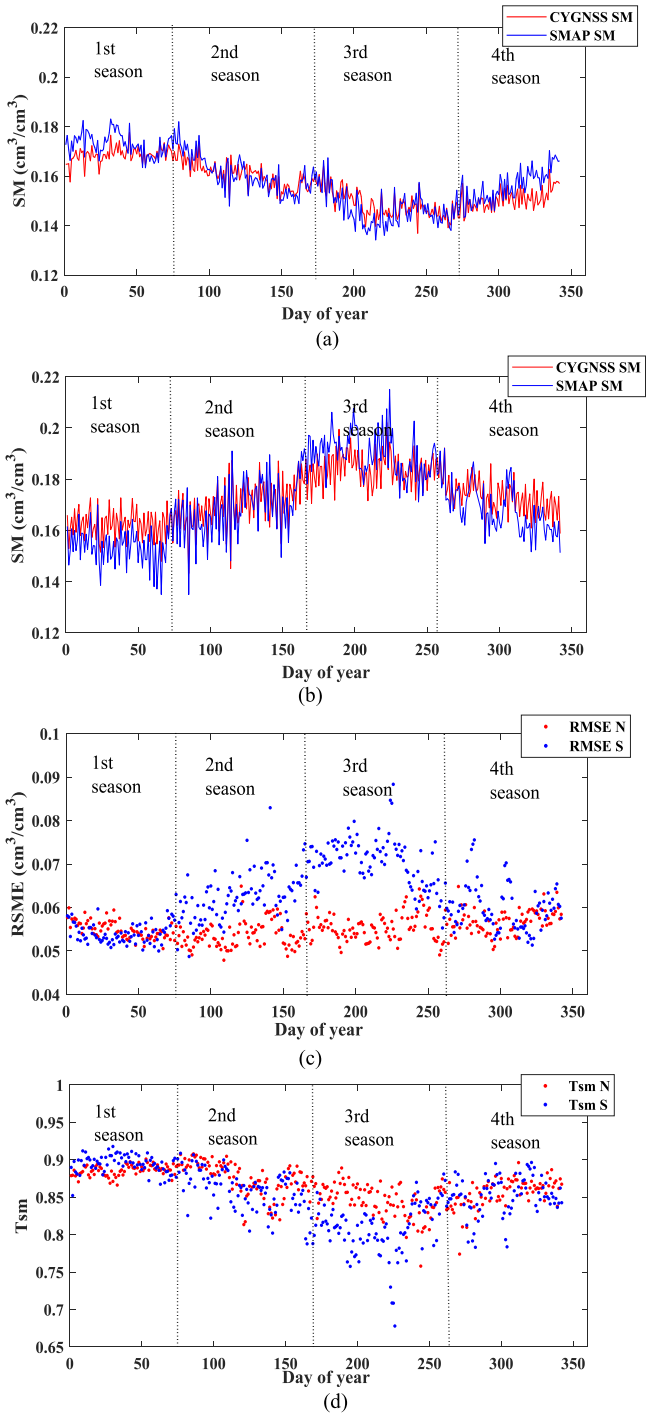

Fig. 8. Time series of seasonal predictions throughout the year (2019.7–2020.6). (a) Daily averaged SM in the northern hemisphere. (b) Daily averaged SM in the southern hemisphere. (c) RMSE in the southern and northern hemispheres. (d) Temporal correlation of T_{sm} in the southern and northern hemispheres.

data (0.056 cm^3/cm^3). Apart from that, different regions are dominated by different land types, which impacts the performance of SM estimation. In particular, higher median RMSE of 0.064 cm^3/cm^3 and ubRMSE of 0.057 cm^3/cm^3 were recorded for the Guangxi Region, which performed more poorly than

TABLE VIII
OVERALL PERFORMANCE OF THE PROPOSED SM AND SMAP PRODUCTS AGAINST *IN SITU* MEASUREMENTS

	Median ubRMSE (cm^3/cm^3)		Median RMSE (cm^3/cm^3)	
In situ vs	CYGNSS	SMAP	CYGNSS	SMAP
ALL point (n=301)	0.049	0.046	0.059	0.056
Guizhou Region	0.057	0.050	0.064	0.059
East-central Region	0.047	0.047	0.057	0.056
Southeastern coastal Region	0.040	0.039	0.053	0.048
Northwest Region	0.048	0.037	0.054	0.052
Other points	0.035	0.037	0.057	0.044

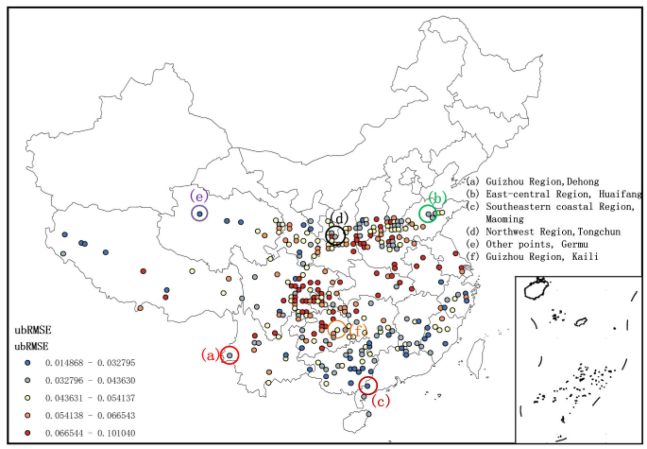

Fig. 9. Daily averaged ubRMSE between CYGNSS-based SM and *in situ* observations in Mainland China.

other regions. Its dominant land type is forest, which tends to have mountainous areas. Thus, the reflected signal was significantly affected by dense vegetation and the high terrain, which also has been evidenced in the previous section.

A map of the distribution of all employed *in situ* stations along with their respective ubRMSE values is shown in Fig. 9. We calculated the ubRMSE between daily averaged CYGNSS retrievals and *in situ* measurements. As mentioned, the ubRMSE values vary depending on the site and the surrounding environment. Generally, we find a good spatial correspondence between CYGNSS-based SM and *in situ* observations, indicating that the SM derived from the proposed CYGNSS-based approach is in good agreement with *in situ* measurement and can be used to produce the expected SM estimates.

Examples of SM time series derived from the CYGNSS-based models, SMAP and *in situ* time-series at each site of 2018 are shown for comparison and to further analyze the temporal variations of the estimated SM [Fig. 10(a–f)]. Six stations are selected as indicated in Fig. 9, i.e., Dehong (evergreen broadleaf forest), Huaifang (croplands), Maoming (natural vegetation mosaic), Tongchuan (woody savannas), Geermu (grasslands), and Kaili (mixed forest), corresponding to each region representing different land types.

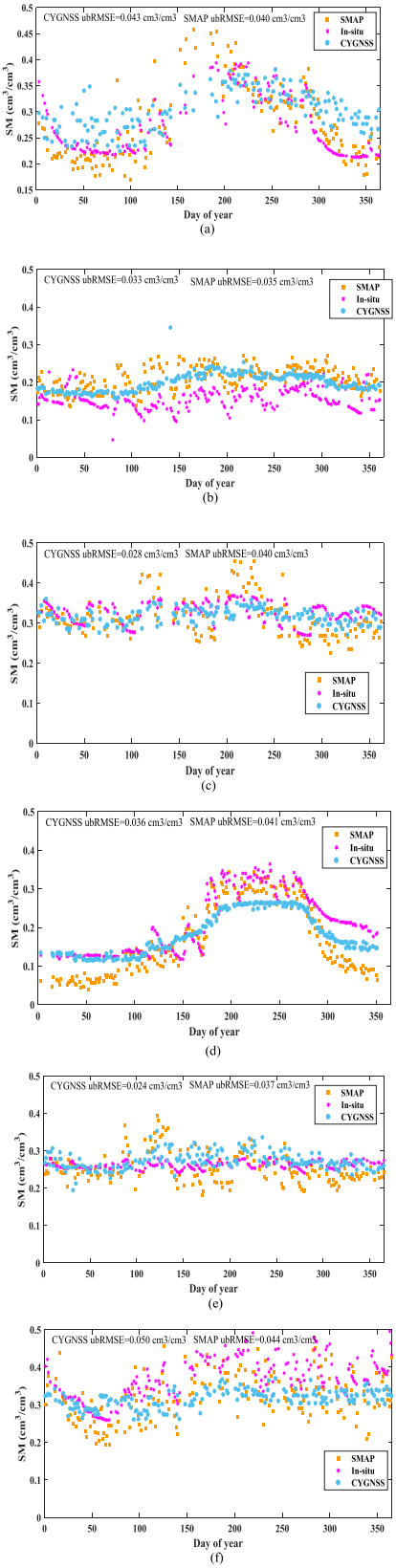

Fig. 10. Comparison for time series of SM derived from the CYGNSS-based model, SMAP, and *in situ* measurements from stations. (a) Guizhou region, Dehong. (b) East-central region, Huaifang. (c) Southeastern coastal region, Maoming. (d) Northwest Region, Tongchuan. (e) Other points, Geermu. (f) Guizhou Region, Kaili.

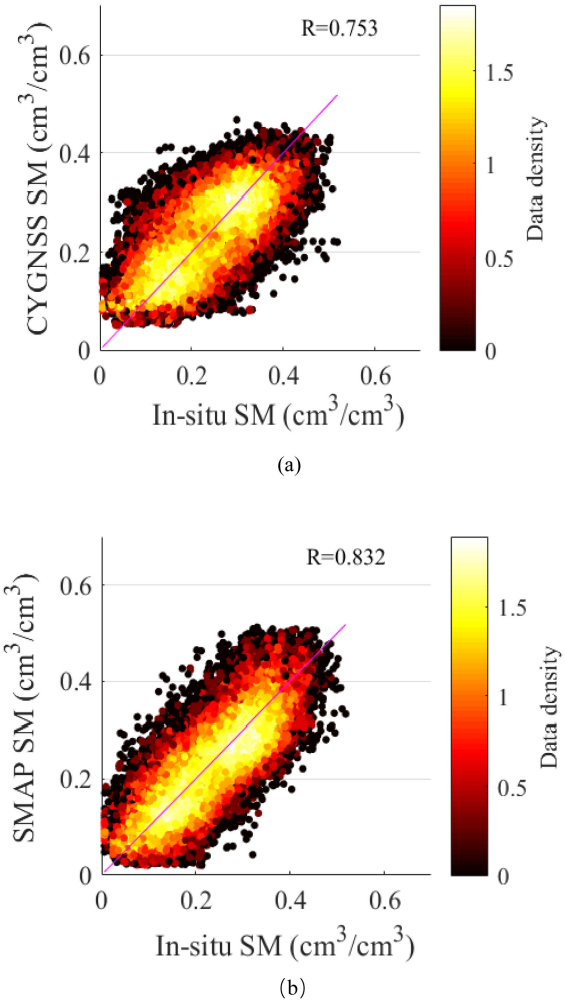

Fig. 11. Density plot in the log-scale of SM derived from (a) the CYGNSS-based model, (b) SMAP, and *in situ* measurements.

The SM retrievals from the CYGNSS-based model at stations agreed well with *in situ* measurements. CYGNSS-based SM from some stations shows comparable [Fig. 10(a,b)] or even better [Fig. 10(c–e)] SM estimation performance against *in situ* measurements compared to the SMAP product. The fluctuations of SM estimates show that the CYGNSS-based model has the ability to retrieve both low and high SM.

Meanwhile, some stations [e.g., Fig. 10(f)] show that the SM retrieved from CYGNSS are worse than SMAP and deviates from the *in situ* values. Commonly, since the SMAP data were employed as the reference SM in the training stage of the SM model, we can expect the CYGNSS-based SM to perform similarly or better than SMAP. In the meantime, the SMAP bias will be passed on to the CYGNSS SM retrieval. The deviation of the CYGNSS-based SM compared to that of SMAP could be mitigated by adding many more samples in the training step and can be investigated in the future. In short, the CYGNSS-based SM estimation shows some levels of variability from site to site but is generally close to the *in situ* measurements with low ubRMSE.

The density scatter plots in the log-scale of the CYGNSS-based SM [Fig. 11(a)] and SMAP [Fig. 11(b)] products against *in situ* measurements are shown. Overall, the comparison of the CYGNSS-based SM and SMAP against *in situ* measurements yields R of 0.753 and 0.823, respectively. SMAP achieves a slightly higher correlation compared to the CYGNSS-based SM. Both density plots highlight an overall fairly good agreement with the *in situ* measurements.

V. Conclusion

This article proposed an enhanced ML approach for estimating the global SM from CYGNSS with the least ancillary data. The novel preclassification strategy possesses a high integration feature that aggregates data from the same land type, helps to minimize the influence of different terrain, and increases the possibility to identify the rules from the data of each category. The CYGNSS data were used for learning the nonlinear relationship between the input vectors (obtained reflectivity, vegetation, and roughness) and the reference SM (SMAP). The approach was implemented as following: the overall data were first resampled according to the IGBP land type and, then, the data from each category were trained and tested separately to build land-specific sub-models.

The overall result was compared with that from the traditional ML regression approach. The preclassification strategy showed an enhanced prediction ability. Several submodels were constructed and compared through their RMSE, which can make full use of the data mining feature of ML. Compared with different typical ML methods, a clear drop of RMSE was observed when the preclassification strategy was employed. The XGBoost model performed best with RMSE of 0.052 cm^3/cm^3. Moreover, three types of input vectors with different deviations of reflectivity were investigated with a 10-fold CV process. The optimized reflectivity derived from CYGNSS BRCS was adopted with the proposed XGBoost to demonstrate the daily and seasonal SM estimation from the spatial and temporal aspects.

The results indicate that different land types have a significant impact on SM estimation [33], which also explains the effectiveness of the preclassification strategy in SM estimation. A strong correlation between CYGNSS reflectivity and SMAP was shown. Among the three predictors, the CYGNSS reflectivity has a larger positive weight, which is consistent with the fact that the enhancement of SM increases the soil permittivity and thus the Fresnel coefficient. The coefficient of VO is also quite large and positive, because the vegetation mainly reduces the coherent reflection [31] and, thus, the variable compensates for this effect. The roughness has a much smaller weight, as expected, but still produces some improvement in the empirical regression performances.

Additionally, in most areas, daily and seasonal SM predictions showed that the VO is negatively correlated to the spatial correlation $R_{\text{reflectivity}}$ and positively correlated to RMSE. Meanwhile, the values of SM are positively correlated with RMSE and also affect the contribution of the predictors. In some extremely dry places (e.g., barren or sparsely vegetated areas), variations in roughness or vegetation are very sensitive and the roughness may become a very important factor affecting signal receptions. On the other hand, the almost unchanged values of low SM are not good for ML modeling and thus also affect spatial correlation R_{sm}. Moreover, the high spatial correlation R_{sm} always appears with a high value of SD SMAP. Higher S.D. of VO and VO lead to increases in RMSE. The RMSEs in the southern hemisphere increase with the seasonal variation (increase with VO) and are higher than in the northern hemisphere. Furthermore, the satisfactory outcome of R_{sm} value of 0.86 and the RMSE of 0.056 cm^3/cm^3 were achieved at a global scale for daily and seasonal SM prediction.

Furthermore, validation of SM with an independent *in situ* source results in a median ubRMSE of 0.049 cm^3/cm^3 and R of 0.753, demonstrating that it could be generalized for regional SM estimation. Meanwhile, the promising results generated by the proposed CYGNSS-based model show comparable accuracy and a higher R with the UCAR/CU SM product [29]. A similar performance of CYGNSS and SMAP is expected since SMAP is the reference data for training the proposed SM model [29]. The proposed SM model obtained comparable results and showed the huge potential of CYGNSS as a complementary data source to SMAP, providing SM retrievals with high revisit times.

The proposed novel preclassification strategy based on the traditional ML regression model can greatly reduce the number of ancillary data and the complexity of modeling while minimizing the influence of different land types, and improve the SM estimation accuracy in a simple and practical way, especially for the big global scale data. The presented approach has been shown to be effective for different ML algorithms, and the estimated CYGNSS SM achieved a satisfactory performance in daily and seasonal predictions. Last but not least, the proposed preclassification strategy can be applied to other training and testing problems and could benefit such as hydrology and agriculture where accurate SM estimates play an important role.

References

[1] N. Rodríguez-Fernández *et al.*, "Soil moisture remote sensing across scales," *Remote Sens.*, vol. 11, no. 2, pp. 190–194, Jan. 2019.

[2] W. Ban, K. Yu, and X. Zhang, "GEO-satellite-based reflectometry for soil moisture estimation: Signal modeling and algorithm development," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1829–1838, Dec. 2017.

[3] A. Egido *et al.*, "Airborne GNSS-R soil moisture and above ground biomass observations," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 5, pp. 1522–1532, May 2014.

[4] C. Lu *et al.*, "Real-time retrieval of precipitable water vapor from Galileo observations by using the MGEX network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4743–4753, Jul. 2020.

[5] D. Masters, P. Axelrad, and S. Katzberg, "Initial results of land-reflected GPS bistatic radar measurements in SMEX02," *Remote Sens. Environ.*, vol. 92, no. 4, pp. 507–520, Sep. 2002.

[6] V. U. Zavorotny, K. M. Larson, J. J. Braun, E. E. Small, E. D. Gutmann, and A. L. Bilich, "A physical model of GPS multipath caused by land reflections: Toward bare soil moisture retrievals," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 3, no. 1, pp. 100–110, Mar. 2009.

[7] Y. Jia, P. Savi, D. Canone, and R. Notarpietro, "Estimation of surface characteristics using GNSS LH-Reflected signals: Land versus water," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 10, pp. 4752–4758, Oct. 2016.

[8] P. C. Dubois, J.V. Zyl, and T. Engman, "Measuring soil moisture with imaging radars." *IEEE Trans. Geosci. remote Sens.*, vol. 33, no. 4, pp. 915–926, Jul. 1995.

[9] D. Entekhabi *et al.*, "The soil moisture active passive (SMAP) mission." in *Proc. IEEE*, vol. 98, no. 5, pp. 704–716, May 2010.
[10] Y. H. Kerr *et al.*, "Soil moisture retrieval from space: The soil moisture and ocean salinity (SMOS) mission." *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 8, pp. 1729–1735, Aug. 2001.
[11] S. Paloscia *et al.*, "Soil moisture mapping using Sentinel-1 images: Algorithm and preliminary validation." *Remote Sens. Environ.*, vol. 134, pp. 234–248, Jul. 2013.
[12] M. Aubert *et al.*, "Analysis of TerraSAR-X data sensitivity to bare soil moisture, roughness, composition and soil crust." *Remote Sens. Environ.,* vol. 115, no. 8, pp. 1801–1810, Aug. 2011.
[13] J. Darrozes, N. Roussel, and M. Zribi, "The reflected global navigation satellite system (GNSS-R): From theory to practice," in *Microwave Remote Sensing of Land Surface*, Amsterdam, The Netherlands: Elsevier, Jul. 2016, pp. 303–355.
[14] M. C. Dobson and F. T. Ulaby, "Active microwave soil moisture research," *IEEE Trans. Geosci. Remote Sens.*, vol. GE-24, no. 1, pp. 23–36, Jan. 1986.
[15] K. M. Larson *et al.*, "GPS multipath and its relation to near-surface soil moisture content," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 3, no. 1, pp. 91–99, Mar. 2010.
[16] S. G. Jin, X. D. Qian, and H. Kutoglu, "Snow depth variations estimated from GPS-reflectometry: A case study in Alaska from L2P SNR data," *Remote Sens*, vol. 8, no. 1, pp. 63, Jan. 2016.
[17] X. Li *et al.*, "Real-time capturing of seismic waveforms using high-rate BDS, GPS and GLONASS observations: The 2017 mW 6.5 Jiuzhaigou earthquake in China," *GPS Solutions*, vol. 23, no.no. 1, pp. 17, Jan. 2019.
[18] S. T. Lowe, *et al.*, "A delay/Doppler-mapping receiver system for GPS-reflection remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 5, pp. 1150–1163, May 2002.
[19] J. F. Marchan-Hernandez *et al.*, "Correction of the sea state impact in the L-band brightness temperature by means of delay-Doppler maps of global navigation satellite signals reflected over the sea surface," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 10, pp. 2914–2923, Oct. 2007.
[20] G. Ruffini *et al.*, "Oceanpal: An instrument for remote sensing of the ocean and other water surfaces using GNSS reflections," *Elsevier Oceanogr. Ser.*, vol. 63, Dec. 2003, pp. 146–153 .
[21] S. Gleason, "Detecting bistatically reflected GPS signals from low earth orbit over land surfaces," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Denver, CO, USA, Jul–Aug. 2006, pp. 3086–3089.
[22] W. Li *et al.*, "First spaceborne phase altimetry over sea ice using TechDemoSat-1 GNSS-R signals," *Geophysical Res. Lett.,* vol. 44, no. 16, pp. 8369–8376, Aug. 2017.
[23] G. Foti *et al.*, "Spaceborne GNSS reflectometry for ocean winds: First results from the U.K. TechDemoSat-1 mission," *Geophysical Res. Lett.*, vol. 42, no. 13, pp. 5435–5441, Jul. 2015.
[24] C. Chew *et al.*, "Demonstrating soil moisture remote sensing with observations from the U.K. TechDemoSat-1 satellite mission," *Geophys. Res. Lett.*, vol. 43, pp. 3317–3324, Apr. 2016.
[25] A. Camps *et al.*, "Sensitivity of TDS-1 GNSS-R reflectivity to soil moisture: Global and regional differences and impact of different spatial scales," *Remote Sens.*, vol. 10, no. 11, p. 1856, Nov. 2018.
[26] C. S. Ruf *et al.*, "New ocean winds satellite mission to probe hurricanes and tropical convection." *Bull. Amer. Meteorological Soc.*, vol. 97, no. 3, pp. 385–395, Mar. 2016.
[27] H. Kim and L. Venkat, "Use of cyclone global navigation satellite system (CYGNSS) observations for estimation of soil moisture." *Geophysical Res. Lett.,* vol. 45, no. 16, pp. 8272–8282, Aug. 2018.
[28] C. C. Chew and E. E. Small, "Soil moisture sensing using spaceborne GNSS reflections: Comparison of CYGNSS reflectivity to SMAP soil moisture," *Geophys. Res. Lett.*, vol. 45, no. 9, pp. 4049–4057, May 2018.
[29] C. C.Chew and E. E. Small, "Description of the UCAR/CU soil moisture product," *Remote Sens.*, vol. 12, no. 10. pp. 1558, May 2020.
[30] M. M. Al-Khaldi *et al.*, "Time-series retrieval of soil moisture using CYGNSS." *IEEE Trans. Geosci. Remote Sens.*, ol. 57, no. 7, pp. 4322–4331, Jul. 2019.
[31] M. P. Clarizia, N. Pierdicca, F. Costantini, and N. Floury, "Analysis of CyGNSS data for soil moisture retrieval." *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2227–2235, Jul. 2019.
[32] A. Calabia, I. Molina, and S. Jin, "Soil moisture content from GNSS reflectometry using dielectric permittivity from fresnel reflection coefficients." *Remote Sens.*, vol. 12, no. 1, pp. 122, Jan. 2020.
[33] Q. Yan, W. Huang, S. Jin, and Y. Jia, "Pan-tropical soil moisture mapping based on a three-layer model from CYGNSS GNSS-R data." *Remote Sens. Environ.,* vol. 247, *Jan. 2020*, Art. no. 111944.
[34] O. Eroglu *et al.*, "High spatio-temporal resolution CYGNSS soil moisture estimates using artificial neural networks." *Remote Sens.*, vol. 11, no. 19, pp. 2272, Sep. 2019.
[35] V. Senyurek, F. Lei, D. Boyd, M. Kurum, A. C. Gurbuz, and R. Moorhead, "Machine learning-based CYGNSS soil moisture estimates over ISMN sites in CONUS." *Remote Sens.*, vol. 12, Apr. 2020, Art. no. 1168.
[36] V. Senyurek *et al.*, "Evaluations of a machine learning-based CYGNSS soil moisture estimates against SMAP observations." *Remote Sens.*, vol. 12, no. 21, Oct. 2020, Art. no. 3503.
[37] T. Yang *et al.*, "Comprehensive evaluation of using TechDemoSat-1 and CYGNSS data to estimate soil moisture over Mainland China." *Remote Sens.*, vol. 12, no. 11, May 2020, Art. no. 1699.
[38] C. S. Ruf, S. Gleason, and D. S. McKague, "Assessment of CYGNSS wind speed retrieval uncertainty." *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 1, pp. 87–97, Jan. 2019.
[39] C. Handbook, *Cyclone Global Navigation Satellite System: Deriving Surface Wind Speeds in Tropical Cyclones*, Ann Arbor, MI, USA: University of Michigan, 2016.
[40] H. Carreno-Luengo, G. Luzi, and M. Crosetto, "Sensitivity of CyGNSS bistatic reflectivity and SMAP microwave radiometry brightness temperature to geophysical parameters over land surfaces." *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 1, pp. 107–122, Aug. 2018.
[41] S. Chan and S. Dunbar, "Level 3 passive soil moisture product specification document," Beta Release, JPL D-72551, Aug. 2015.
[42] N. Rodriguez-Alvarez, E. Podest, K. Jensen, and K. C. McDonald, "Classifying inundation in a tropical wetlands complex with GNSS-R." *Remote Sens.*, vol. 11, no. 9, pp. 1053, May 2019.
[43] I. Ali *et al.*, "Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data," *Remote Sens*, vol. 12, no. 7, pp. 16398–16421, Nov. 2015.
[44] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogramm. Remote Sens.*, vol. 114, pp. 24–31, Apr. 2016.
[45] L. Breiman, Random forests. *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
[46] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 1995.
[47] G. Moser and S. Serpico, "Automatic parameter optimization for support vector regression for land and sea surface temperature estimation from remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 909–921, Mar. 2009.
[48] Q. Yuan *et al.*, "Deep learning in environmental remote sensing: Achievements and challenges." *Remote Sens. Environ.*, vol. 241, May 2020, Art. no. 111716.
[49] N. J. Rodríguez-Fernández *et al.*, "Soil moisture retrieval using neural networks: Application to SMOS," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 5991–6007, Nov. 2015.
[50] Q. Yan, S. Gong, S. Jin, W. Huang, and C. Zhang, "Near real-time soil moisture in china retrieved from CyGNSS reflectivity," *IEEE Geosci. Remote Sens. Lett.*, to be published, doi: 10.1109/LGRS.2020.3039519.

Yan Jia (Member, IEEE) received the double M.S. degree in telecommunications engineering and computer application technology from Politecnico di Torino, Turin, Italy, and Henan Polytechnic University, in 2013. She received the Ph.D. degree in electronics engineering from Politecnico di Torino, in 2017.

She is with the Nanjing University of Posts and Telecommunications, Nanjing, China. In 2013, she was with the Department of Electronics and Telecommunications, Politecnico di Torino. In 2014, she was with the SMAT project, mainly focusing on the retrieval of soil moisture and vegetation biomass content by GNSS-R. Her research interests include microwave remote sensing, soil moisture retrieval, and global navigation satellite system reflectometry (GNSS-R) applications to land remote sensing and antenna design.

Shuanggen Jin (Senior Member, IEEE) was born in Anhui, China, in September 1974. He received the B.Sc. degree in Geodesy from Wuhan University, Wuhan, China, in 1999, and the Ph.D. degree in geodesy from the University of Chinese Academy of Sciences, Beijing, China, in 2003.

He is currently a Professor and Dean with the Nanjing University of Information Science and Technology, China, and also a Professor with Shanghai Astronomical Observatory, CAS, Shanghai, China. His research areas include satellite navigation, remote sensing and space/planetary geodesy. He has authored/coauthored over 500 papers in peer-reviewed journals and proceedings, 10 patents/software copyrights, and 10 books/monographs with more than 6000 citations and H-index > 40.

Prof. Jin has been President of International Association of Planetary Sciences (IAPS) (2015–2019), President of the International Association of CPGPS (2016–2017), Chair of IUGG Union Commission on Planetary Sciences (UCPS) (2015–2023), Vice-President of the IAG Commission (2015–2019), Editor-in-Chief of International Journal of Geosciences, Editor of *Geoscience Letters*, Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and *Journal of Navigation*, Editorial Board Member of *Remote Sensing*, *GPS Solutions*, and *Journal of Geodynamics*. He has received one first-class and four second-class prizes of provincial awards, 100-Talent Program of CAS, Leading Talent of Shanghai, IAG Fellow, IUGG Fellow, Member of Russian Academy of Natural Sciences, Member of European Academy of Sciences, and Member of Academia Europaea.

Haolin Chen received the B. Eng. degree in geomatics engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2020.

His research interests include land and ocean remote sensing using global navigation satellite system reflectometry (GNSS-R).

Qingyun Yan (Member, IEEE) was born in Haimen, China. He received the B. Eng. degree in electronic science and engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2014, and the M.Eng. and Ph.D. degrees in electrical engineering from the Memorial University of Newfoundland, St. John's, NL, Canada, in 2015 and 2020.

He is now with the School of Remote Sensing & Geomatics Engineering, Nanjing University of Information Science & Technology, China. His research interests include tsunami, sea ice, and land remote sensing using global navigation satellite system-reflectometry.

Dr. Yan has received the 2019 IEEE GRSS Letters Prize Paper Award from the IEEE Geoscience and Remote Sensing Society.

Patrizia Savi (Senior Member, IEEE) received the Laurea degree in electronic engineering from the Politecnico di Torino, Turin, Italy.

In 1986, she was a Consultant in Alenia (Caselle Torinese, Italy). From 1987 to 1998, she was a Researcher with the Italian National Research Council (CNR). In 1998, she joined the Electronic Department, Politecnico di Torino, as an Associate Professor. She currently teaches a course on electromagnetic field theory. Her research interests include dielectric radomes, frequency-selective surfaces, waveguide discontinuities and microwave filters, high-altitude platform (HAP) propagation channels, and global navigation satellite system reflectometry (GNSS-R) for soil moisture retrieval.

Dr. Savi is currently a member of SIEM (Società Italiana di Elettromagnetismo).

Yan Jin received the B.S. degree in information and computation science and the M.S. degree in applied mathematics, from Chang'an University, Xi'an, China, in 2011 and 2014, respectively, and the Ph.D. degree in cartography and geographical information system from the State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, University of Chinese Academy of Sciences, Beijing, China, in 2018.

She is currently a Lecturer with the School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing, China. Her research interests focus on scale transformation, data fusion, geostatistics, and remote sensing applications.

Yuan Yuan (Member, IEEE) received the B.S. degree in geography from Nanjing University, Nanjing, China, in 2011, and the Ph.D. degree in signal and information processing from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, China, in 2016.

She is currently with the Department of Surveying and Geoinformatics, Nanjing University of Posts and Telecommunications, China. Her research interests include remote sensing time series analysis and machine learning.