An Effective Land Type Labeling Approach for Independently Exploiting High-Resolution Soil Moisture Products Based on CYGNSS Data

Yan Jia, *Member, IEEE*, Shuanggen Jin, *Senior Member, IEEE*, Qingyun Yan, *Member, IEEE*, Patrizia Savi, *Senior Member, IEEE*, Rongchun Zhang, and Wenmei Li, *Member, IEEE*

***Abstract*—Recently, soil moisture (SM) has been estimated using Cyclone Global Navigation Satellite System (CYGNSS) data. Machine learning (ML) algorithms for CYGNSS SM estimation can minimize unpredictable influences and help improve the accuracy of SM retrieval. However, ML-based CYGNSS SM estimation requires ancillary data from other sources, and thus, the uncertainty, internal errors, and even dependence on external parameters of this process may complicate and limit SM estimation. In this article, a simple land type (LT) digitization strategy that incorporates the idea of classification is proposed with feature optimization to achieve an effective and independent SM retrieval without any other auxiliary data. The input features are chosen from the CYGNSS data themselves, and the corresponding labels (digitized stable LTs) are used in the training stage of the SM estimation model. During the fine-tuning stage, several input features (such as the dielectric constant and incident angle) are compared and selected after optimization to achieve better results. Moreover, the CYGNSS data are gridded at 9 × 9 km to validate the enhanced soil moisture active passive mission SM products at a resolution of 9 km. Only three input variables are adopted for the SM learning model, which are directly derived from the CYGNSS data for independently estimating SM at a high spatial resolution. Powerful performance is achieved by extreme gradient boosting based on a LT digitalization strategy, with root-mean-square error (RMSE) and unbiased RMSE (ubRMSE) values of 0.063 cm^3/cm^3 and a correlation coefficient (R) of 0.71 for the entire dataset. The performances of different ML learning models for various LTs are presented. The mean ubRMSE and RMSE are 0.041 cm^3/cm^3 and 0.057 cm^3/cm^3, respectively. The results demonstrate the effectiveness of the proposed LT digitization strategy for retrieving SM from CYGNSS data with various ML methods and the capability of SM estimation using the CYGNSS product as a new independent source.**

***Index Terms*—Cyclone-GNSS (CYGNSS), global navigation satellite system-reflectometry (GNSS-R), machine learning (ML), soil moisture (SM), soil moisture active passive (SMAP).**

I. Introduction

SOIL moisture (SM) directly controls surface water and energy balance and, thus, is vital to the actual needs of climate, agriculture, and drought monitoring [1], [2]. Microwave remote sensing possesses 24-h, all-weather, and large-scale monitoring capabilities for high-precision SM retrieval [3], [4]. At present, many passive microwave satellites and sensors were used to observe surface SM (<5 cm), such as the National Aeronautics and Space Administration's (NASA's) Advanced Microwave Scanning Radiometer-Earth Observing System [5], the soil moisture active passive (SMAP) mission [6] and the Soil Moisture and Ocean Salinity (SMOS) mission of the European Space Agency [7]. The use of microwave sensors can obtain high-precision SM products; for example, the error of 36-km SMAP SM products was approximately 0.04 m^3/m^3 [8]. However, its long revisit period of 2–3 days restricts its application with higher time resolutions (1d).

Cyclone Global Navigation Satellite System (CYGNSS), based on the measurement technology of the signals reflected by the GNSS, was launched by NASA on Dec. 15, 2016. The high-precision and excellent data provided by the CYGNSS constellation offers a very favorable opportunity for realizing long-term dynamic SM monitoring with high spatial and temporal resolutions [9]–[17]. In general, CYGNSS-based land remote sensing is facilitated by the observed surface reflectivity Γ [9] or the bistatic radar cross-section (BRCS) [11], which both lead to successful land remote sensing applications. This work mainly focuses on using CYGNSS reflectivity. In the literature, Chew and Small [12] reported that CYGNSS reflectivity changes were correlated with changes in SMAP SM. By using a linear regression method, SM estimates can be obtained with an unbiased root-mean-square error (ubRMSE) of 0.045 cm^3/cm^3. Clarizia *et al.* [14] proposed a significant reflectivity-vegetation-roughness algorithm to retrieve SM with

Manuscript received November 2, 2021; revised February 15, 2022 and April 26, 2022; accepted May 13, 2022. Date of publication May 20, 2022; date of current version June 2, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42001375 and Grant 42001362, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20180765 and Grant BK20191384, in part by the Nanjing Technology Innovation Foundation for Selected Overseas Scientists under Grant RK032YZZ18003, 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, Rongchun Zhang, and Wenmei Li are with the Department of Surveying and Geoinformatics, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: jiayan@njupt.edu.cn; rongchunzhang@njupt.edu.cn; liwm@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.2022.3176031

TABLE I
APPLICATIONS OF CYGNSS SM ESTIMATION METHODS WITH A RESOLUTION OF 9 KM

Source	Time span	Spatial coverage	Reference SM	Validation SM	Adopted algorithms	No. of ancillary data	Overall performance (cm^3/cm^3)	Spatial resolutions
Kim and Lakshmi (2018)	1 year	Regional	SMAP	SMAP	Linear regression	**1**	R=0.68/0.77	9x9 km
Eroglu et al. (2019)	2 year	Regional	ISMN sites	ISMN sites	ANN	**5**	0.054 (ubRMSE), R=0.90	9x9 km
Seuyurek et al. (2020)	3 year	Regional	ISMN sites	ISMN sites	RF/ANN/SVM	**5–7**	0.052/0.061/0.065 (RMSE), R=0.64-0.89 (RF)	9x9 km
Senyurek et al. (2020)	3 year	Quasi-global	ISMN sites	SMAP and ISMN sites	Random forest	**5–7**	0.066 (RMSD), R=0.66, 0.044 (mean ubRMSD) (CYGNSS vs. SMAP)	9x9 km
Proposed method	1 year	Quasi-global	SMAP	SMAP	XGBoost with an LT digitization strategy	**1**	0.063(RMSE), R=0.71, 0.041 (mean ubRMSE), 0.057 (mean RMSE)	9x9 km

RMSE of 0.07 cm^3/cm^3, where a linear regression model was developed for daily SM estimates with variables including the CYGNSS reflectivity, roughness coefficient, and vegetation opacity in SMAP. Yan *et al.* [16] adopted a similar method but utilized the statistical properties of CYGNSS reflectivity to analyze the effect of surface roughness, and then the SM was determined with a correlation coefficient (R) of 0.80 and an RMSE of 0.07 cm^3/cm^3. In addition, machine learning (ML) algorithms have been rapidly applied to CYGNSS-based SM estimation. A backpropagation-artificial neural network (ANN) algorithm was adopted by Yang *et al.* [17] to evaluate the SM estimation performance of two spaceborne GNSS-R satellite missions (TechDemoSat and CYGNSS). The SM was obtained with an R of 0.79 and an ubRMSE of 0.062 cm^3/cm^3. However, the use of a few (six) ancillary variables was inevitable, and all these previous studies utilized SMAP as reference and validation data with a resolution of 36 km.

In terms of high-resolution CYGNSS-based SM retrieval, a geosystems research group at Mississippi State University took the International SM Network sites as references and reported some results by employing ML algorithms [18]–[20]. Five to seven ancillary datasets were used, which were obtained from external sources. For example, the 16-day composite normalized difference vegetation index (NDVI) derived from the moderate-resolution imaging spectroradiometer (MODIS) data of MYD13A1 was utilized for characterizing vegetation conditions, and the surface elevation information was provided by a 1-km digital elevation model (GTOPO30) product from the United States Geological Survey Earth Resources Observation and Science archive [19], [20]. Similarly, the NDVI and elevation data were all spatially averaged from their own high resolution to a coarser resolution corresponding to the CYGNSS data. Additionally, NDVI data usually suffer from clouds because they are generated by optical instruments. A sliding window average over 16 days was applied in a previous study [18]. This spatial/temporal averaging operation and the use of imperfect data sources with insufficient observation days or a running process of one day become potential and inevitable risks, which may increase the computational cost and even deteriorate the estimation model. Thus, the dependence on ancillary data may bring big issues, and it is essential to diminish the subset of relevant features to achieve independent retrieval.

Furthermore, it is unfair to directly compare the obtained RMSEs to those in the literature since cases vary in terms of data samples, time spans, spatial coverage, assumptions regarding gridding, validation datasets, employed ancillary data, and spatial resolutions. These factors all impact the performance of the CYGNSS-based SM estimation. Here, the related 9-km high-resolution CYGNSS estimations are summarized in Table I.

In this article, we propose a novel land type (LT) digitalization strategy to balance the accuracy and efficiency of independent CYGNSS-based SM estimations. We adopt digitalized (labeling) LTs as the main physically-based features to incorporate the surface conditions (e.g., topography, vegetation, and soil properties, among others) into the SM learning model. Quasi-global SMAP data are used as reference data and to crossly validate the learning model. In this approach, enormous amounts of ancillary data can be discarded. Background terrain knowledge is accumulated in the learning model through one variable. This extracted feature incorporates the idea of classification, thus making the SM estimation model smart and promoting stand-alone retrieval. Moreover, various ML methods and variants are compared with different LTs to obtain optimized performance in high-resolution CYGNSS SM estimation.

The key innovations and aims of the article are threefold as follows.

1) A simple and effective independent SM retrieval scheme from CYGNSS data is proposed to cope with the involvement of complicated auxiliary data.
2) The introduced LT digitization/labeling strategy solves the insufficient geographical representation when performing stone-alone CYGNSS retrieval and greatly avoids the complex calculations when other data sources are considered.
3) Comprehensive experiments on different ML algorithms with feature optimization are conducted to validate the effectiveness of the proposed method and high-resolution CYGNSS SM estimation is achieved with different LTs at the global scale.

The rest of this article is organized as follows. Section II depicts the dataset used in this article. Section III describes the proposed CYGNSS SM estimation architecture and LT digitization scheme. Section IV reports the experimental results. Finally, Section V concludes this article.

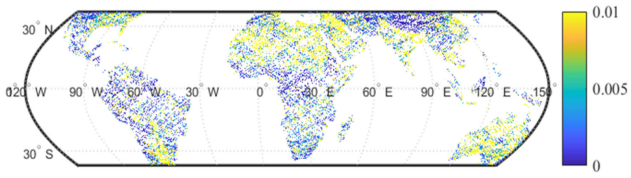


Fig. 1. Example BRCS of CYGNSS reflectivity on Jan. 1, 2018.

II. Data and Quality Control

A. CYGNSS Data

The CYGNSS constellation contains eight CubeSats, which can provide reflected signal data covering the whole pantropical zone (38°S–38°N) and has the characteristics of high spatial and temporal resolution compared to SMAP. We adopted the data for the entire year of 2018 and the time period for which the reference and validation data were available. The employed CYGNSS Level 1 (L1) data[1] include a time-delay Doppler map (DDM), the radar cross-section (BRCS or σ) shown in Fig. 1, and other measurement and geographic coordinate information, such as the incident angle θ, signal-to-noise ratio, the longitude and latitude (Lat/Lon) of the specular reflection point (SP), and the distance between the SP and the transmitter and receiver (*Rt* and *Rr*) [9].

B. SMAP Data

The SMAP enhanced radiometer Level 3 SM data (global daily 9-km EASE-Grid SM, Version 4) were used as a reference dataset (which can be downloaded freely at[2]) for a comparison with the SM estimated by the CYGNSS. Furthermore, a 16-bit binary string of 1 s and 0 s called a SMAP retrieval quality flag (RQF) was taken as a significant quality control indicator [24]. The first position, "recommended quality," indicated whether the SM retrieval possessed the recommended quality (first position = 0). In this article, the extracted data were filtered according to the first position of the "recommended quality" flag to ensure the quality of the data used in the modeling calculations [20]. Although this operation decreased the overall amount of data, the high-resolution global datasets at 9 km could still provide valid training datasets for the learning model. Thus, the data points that had 1 s in the first positions of their RQFs were deleted and excluded from subsequent calculations. This article used the data for the whole year of 2018 that were marked with "retrieval recommended". To facilitate subsequent verification and comparison procedures, the CYGNSS data were projected to the EASE-Grid employed by the SMAP data with 9×9 km cells. As such, the resolution of SM estimation was considered to be 9 km. The global coverage of SMAP was achieved approximately every three days and was more than CYGNSS for two days. An example of SMAP SM without QF is shown in Fig. 2.

[1]Free [Online]. Available: https://podaac.jpl.nasa.gov

[2][Online]. Available: https://nsidc.org/data/SPL3SMP_E/versions/4

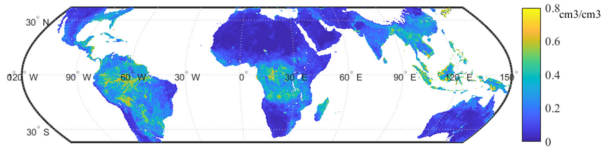


Fig. 2. SM example from Jan. 1–3, 2018.

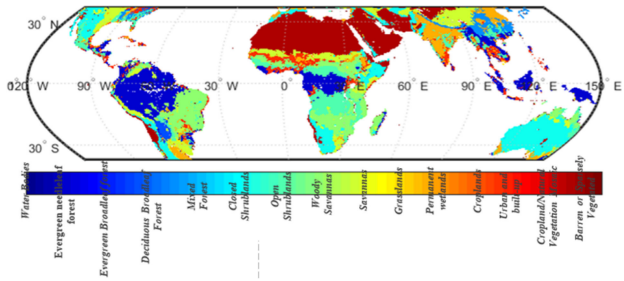


Fig. 3. IGBP LTs in 2018.

C. LT Information

The LT information was relatively stable and could be obtained from various data sources, making it more reliable than other ancillary data types that were adopted by previous studies. This information could be acquired from sources such as the MODIS Aqua Surface Reflectance Daily Global 500 m dataset [10], [11], [17]–[20], the Global Land Cover Map for 2009 (GLCover 2009) dataset [17], and the SMAP mission [16], [25]. In this article, we employed the LT information obtained from SMAP radiometer Level 3 global daily 36-km[3] EASE-Grid SM data. The global LT information is shown in Fig. 3, which corresponds to the International Geosphere-Biosphere Programme (IGBP) ecosystem surface classifications. Although the resolution of the LT information (36 km) was not consistent with the 9-km CYGNSS data and reference SM data, this LT information was easier to manipulate and match with the CYGNSS and reference SM datasets since they all used the EASE-Grid projections and could, thus, obtain the expected results. High-resolution LT information can be upgraded to match the training data in future work; this may produce even better results.

D. Quality Control

The collected data were filtered using several criteria as follows.

1) CYGNSS reflectivity values below 0 and above 0.1 were excluded. Some extremely high reflectivity values were found, and they appeared consecutively within the same ground track over the EASE grid [16].
2) We excluded the data obtained at elevation angles smaller than 30°, which could effectively remove very weak signals (which may have resulted in the inclusion of noisy DDMs and errors in the SM estimations) coming from the side lobe of the circular polarization antenna [11], [17], [21].

[3][Online]. Available: https://nsidc.org/data/SPL3SMP

TABLE II
NUMBER OF SAMPLES USED IN THE LEARNING MODEL AFTER DATA QUALITY CONTROL

LT (IGBP)	Number of samples
All LTs	31 200 471
Evergreen needleleaf forest	10 243
Evergreen broadleaf forest	15 681
Deciduous broadleaf forest	7776
Mixed forest	47 468
Closed shrublands	34 777
Open shrublands	8 113 635
Woody savannas	2 005 196
Savannas	2 519 624
Grasslands	4 085 592
Permanent wetlands	9806
Croplands	3 836 044
Urban and built-up	68
Cropland/Natural vegetation mosaics	1 039 599
Barren or sparsely vegetated	9 461 226
Water bodies	13 736

3) The negative antenna gain was removed (corresponding to the uncertainties reported in the measured antenna gain patterns) to ensure that only high-quality data obtained from the left-hand circularly polarized (LHCP) data were used [17], [18], [19], [22].
4) Observations with DDM peak values outside of the range of 5 to 11 delay bins were removed from the dataset to avoid the inclusion of high-altitude measurements [19], [23].
5) The SMAP "retrieval recommended" quality flag was used to filter the SMAP data to ensure the quality of the SM estimations [20]. The total numbers of samples obtained after conducting the quality control procedure for the training and testing datasets are displayed in Table II.

III. SELF-SM ESTIMATION SCHEME

The input features were designed and intentionally selected from the CYGNSS data to achieve stand-alone SM estimations. Hence, several input variables were tested and compared for feature optimization. Except for the digitized LT, the CYGNSS reflectivity, incident angle, dielectric constant, and TES are selected to demonstrate the capability of the SM estimation model.

A. CYGNSS Reflectivity and Dielectric Constant

Assuming that the signal over land is predominantly determined by the coherent reflection from the surface, eventually reduced by roughness and attenuated by vegetation, the reflectivity $\Gamma_{lr}(\theta)$ over a vegetated terrain can be regarded as a function of SM, vegetation, and the surface roughness effect. Thus, the relationship between reflectivity and the reflection coefficient $R_{lr}(\theta)$ for most soils can be expressed by the following formula:

$$\Gamma_{lr}(\theta) = R_{lr}(\theta)^2 \gamma^2 \exp -\mathrm{hcos}^2(\theta) \tag{1}$$

where γ is transmissivity, which accounts for vegetation canopy attenuation; θ is the local incidence angle; and the h-parameter is assumed to be linearly related to the root-mean-square height surface roughness.

The reflectivity is commonly used as the primary variable in SM estimation models. Following the assumption of coherent reflection [12]–[20], [22], [23], [25], the reflectivity $\Gamma_{lr}(\theta)$ can be computed from CYGNSS BRCS σ, which has been verified as optimal for SM estimation [18], [23]

$$\Gamma_{lr}(\theta) = \frac{\sigma(R_t + R_r)^2}{4\Pi(R_t R_r)^2} \tag{2}$$

where R_t and R_r are the distances from the transmitter and receiver to the SP, respectively, and σ, R_t, and R_r are obtainable from CYGNSS data.

In practice, after the surface reflectivity Γ_{lr} is obtained, the Fresnel reflection coefficient R_{lr} can be approximately simplified based on the square root of Γ_{lr} [15], [27] with respect to the smooth surface assumption. Thus, R_{lr} can be calculated from the CYGNSS surface reflectivity Γ_{lr}. With the known R_{lr}, the soil dielectric constant can be obtained by substituting (4) and (5) into (3). The Fresnel reflection (R_{lr}) of the soil surface is a function of the permittivity ε_{r} and the angle of incidence θ

$$R_{lr}(\theta) = \frac{1}{2}\left(R_{\mathrm{vv}}(\theta) - R_{\mathrm{hh}}(\theta)\right) \tag{3}$$

where $R_{\mathrm{vv}}(\theta)$ and $R_{\mathrm{hh}}(\theta)$ are the Fresnel coefficients for, respectively, horizontal and vertical polarization

$$R_{\mathrm{hh}}(\theta) = \frac{\cos(\theta) - \sqrt{\epsilon_{\mathrm{r}} - \sin^2(\theta)}}{\cos(\theta) + \sqrt{\epsilon_{\mathrm{r}} - \sin^2(\theta)}} \tag{4}$$

$$R_{\mathrm{vv}}(\theta) = \frac{\epsilon_{\mathrm{r}}\cos(\theta) - \sqrt{\epsilon_{\mathrm{r}} - \sin^2(\theta)}}{\epsilon_{\mathrm{r}}\cos(\theta) + \sqrt{\epsilon_{\mathrm{r}} - \sin^2(\theta)}} \tag{5}$$

where ε_{r} is the complex permittivity of the soil.

It should be noted that the imaginary part of permittivity can be neglected [27] for most soils (dry and wet). With this assumption, the real part of permittivity (dielectric constant) can be obtained. In this article, the obtained dielectric constant is tested as one of the inputs of the training model.

B. CYGNSS TES

A delay waveform is the returned power profile as a function of the delay only, with the frequency set to a constant value (normally the value at the specular point). A direct signal will exhibit a sharp triangle shape as a result of the GPS correlation process. Consequently, various delay waveform shapes are formed due to the effects of soil humidity and surface roughness (which contributes to the total signal power as a noncoherent influence). The trailing edge slope (TES) is considered the slope of the reflectivity delay waveform (see Fig. 4). The TES slope can be computed from the reflectivity delay waveform values for delay bins m and $m+3$, where m corresponds to the position

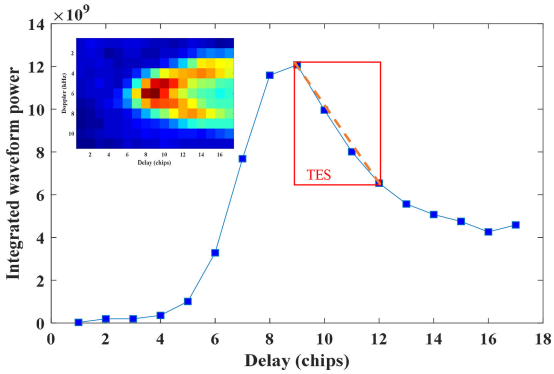


Fig. 4. Example BRCS DDM and the corresponding delay waveform from which the TES is derived.

of the peak of the waveform. TES is a shape-based observable variable and more incoherent mixing through the scattering surface makes TES smaller [23]. The contribution of TES was examined and found significant in improving the SM regression model [18]–[20]; thus, it is included in the input features.

C. LT Digitization/Labeling Strategy

By considering the linear and nonlinear relations among input features, the SM estimation regression model can be improved. However, using too many ancillary data or models that are too complex may lead to overfitting. A large feature set will increase the computational cost and number of cross-correlated features, which might lead to marginal improvements or even reductions in the final performance. Moreover, the ancillary data from other data sources may be characterized by uncertainty and internal errors, and the dependence on external parameters may complicate and limit CYGNSS SM estimation.

In this article, LT information was employed to integrate the complete characterizations of various land surfaces. We digitized LT information for modeling; thus, labels were created as inputs for the SM learning model (see Fig. 5) to enhance SM estimation. The CYGNSS data were identified with labels according to their geographic coordinate information. With this LT digitization/labeling approach, global bio/geophysical dynamics can ultimately be retained and represented in an intelligent form. The rules for SM variations tend to be consistent when the data belong to the same LT. Thus, this digitized LT can be regarded as an extracted feature and incorporates the idea of classification. The learning model is easy to identify the rules from the data and is expected to show better estimation accuracy. Moreover, the proposed SM estimation model does not rely on other auxiliary data sources. The LT changed little, and data were easily obtained.

D. Feature Optimization

The CYGNSS TES was selected to be a significant SM estimation input in addition to reflectivity. Apart from that, in this article, the input variables were limited to being selected from the CYGNSS data to avoid the need for other data sources.

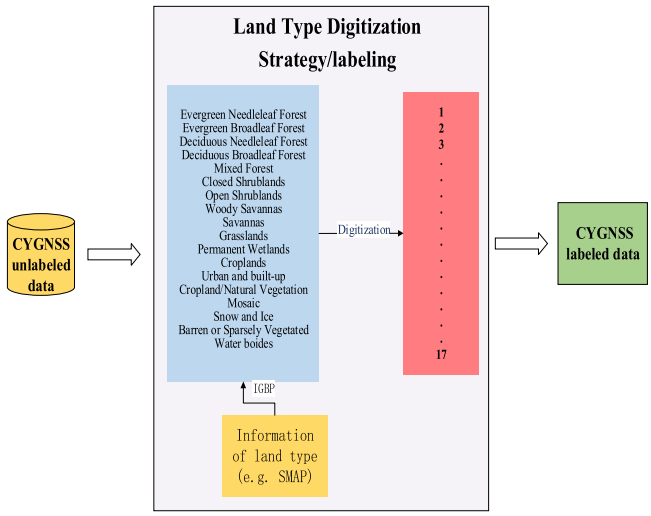


Fig. 5. Proposed LT digitization strategy that integrates land surface information.

Moreover, the number of input variables was constrained (four or less) to balance accuracy and efficiency. Input variables such as the {TES (T), reflectivity (B), dielectric constant (D), LTs (L), and incident angle (A)} were considered. As we have mentioned before, the inclusion of too many ancillary input data has drawbacks. In addition, using many correlated features easily causes the model to ignore the essence and mechanism of SM retrieval and makes it a simple numerical calculation. Therefore, feature selection and optimization procedures are particularly important. They must take the intrinsic principle and rules of the SM retrieval problem into account while also achieving higher prediction accuracy and better model expression. Considering both the principles and practical applications of SM retrieval, the input variants were set with the following combination: $\{D+T+L+A, B+T+L+A, B+T+A, B+T+L\}$. In the fine-tuning stage, the input features are compared, and the variables that display the best performance are selected as the optimal variables for CYGNSS SM estimation.

E. CYGNSS-Based Digitized LT SM Estimation Scheme

The flowchart in Fig. 6 shows the training and validation processes designed to independently estimate CYGNSS SM using the LT digitization/labeling approach.

The downloaded CYGNSS dataset was preprocessed to obtain reflectivity and called the "unlabeled data". The data were identified by the LT information and converted to digital information to be labeled. In this way, the land surface background was mostly retained and accumulated in the learning mode. Then, the labels and other selected variables (reflectivity, dielectric constant, incident angle, and TES) were combined and used as inputs to train the ML-based SM learning model. Here, a 10-fold cross validation (CV) procedure was adopted to remove the codependence of the training and testing datasets, and the overall performance matrix was obtained.

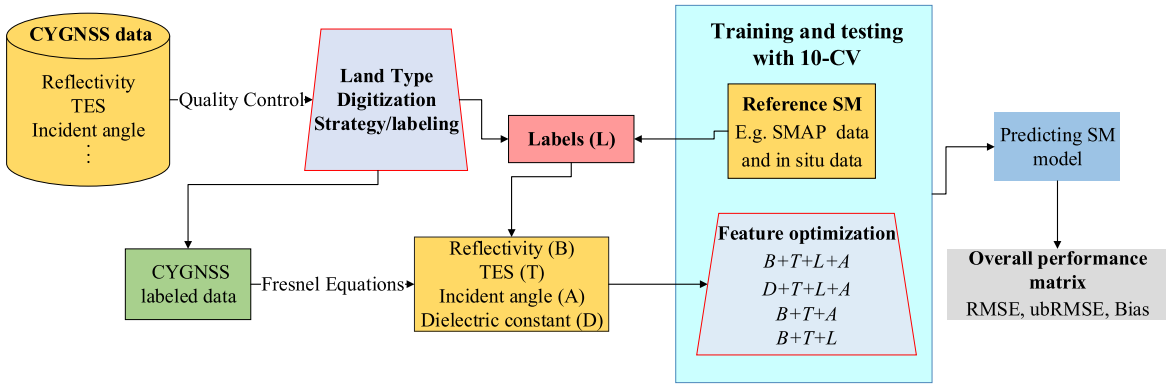


Fig. 6. Flowchart showing the independent SM estimation process using ML regression with the proposed LT digitization strategy.

IV. Experiments and Results

A. Evaluation Criteria and Methods

In this article, the CYGNSS data were resampled to a 9 km EASE grid to correspond with the SMAP data. In particular, if there are multiple CYGNSS SPs that fall within a 9 km grid, all were considered in the experiments [13], [18]. In such a case, a constant SM value was assumed for all the CYGNSS observations in the same 9 km grid. This was considered feasible because the geophysical parameters (such as LT and TES) corresponding to each CYGNSS observation differ due to spatial variations, which in turn could explain variations in the CYGNSS observations, even for uniform SM values.

Three ML methods including random forests (RF), extreme gradient boosting (XGBoost), and ANNs, were selected and compared to assess the effectiveness of the proposed scheme. Among the traditional ML methods, RFs are popular and widely employed, and they are powerful tools for ML regression [19], [20], [27], [28], [29]. XGBoost exhibits superior performance and provides several advantages, such as a fast speed, easy parameterization scheme, and high robustness [30], [31]; thus, it is chosen here. Advanced ANNs are also commonly used ML techniques and have performed well in previous ML-based CYGNSS SM studies [17], [18], [25], and for these reasons, ANNs are employed as well. These approaches are advantageous in capturing the features of training data, and their designs, in this article, are described as follows.

1) XGBoost: Optimal hyperparameters are used to prevent overfitting and underfitting in ML learning models. The optimal parameters include *n_estimators* (from 200 to 5000 with a 100-step interval), *min_child_weight*, *max_depth* (from 1 to 10 with a 1-step interval), and the learning rate (set to 0.1).
2) RF: The number of trees was varied, and the optimal number was selected for the RF learning model. The optimal number of trees was chosen from the set of {50, 100, 200, 300, 400, 500}.
3) ANN: A three-layer ANN-1D network was used in the experiment [17]–[19], and it included one input layer, two hidden layers (32 neuron nodes), and one output layer. The nonlinear activation function selected was a rectified linear unit (ReLU) function; ReLU functions have been proven effective for the SM retrieval by a previous study [18].

RF and XGBoost were implemented using the Python Scikit-Learn library, and the ANN was implemented using the Python Keras library. Experiments were performed on a workstation with an Intel Xeon Gold 16-Core 5218 CPU (2.30 GHz) with 64 GB of RAM.

The performance of different evaluated algorithms was compared based on the following metrics.

1) RMSE: a measure of accuracy that indicates the differences between values or samples predicted by a model or an estimator and observed values.
2) ubRMSE: this metric is the traditional RMSE with bias removed.
3) Bias: this metric depicts the deviations of the estimate with respect to the true values.

It must be noted that, in general, the performance matrix was calculated and shown for an entire dataset. Except for the distribution maps, the performance matrix was calculated for each grid pixel, so we also show the mean values here to compare them with those obtained in other studies; these values are displayed in Table I.

B. Model Configuration

The performance of different variants is evaluated for independent SM retrieval by using XGBoost, which is the fastest and most effective algorithm. These input variables are selected from the CYGNSS data. Additionally, the number of input variables is constrained (four or less) to balance the accuracy and efficiency of the model. The variables, including the dielectric constant (D), BRCS (B), TES (T), incident angle (A), and LT labels (L), and the corresponding performance matrix for feature optimization are described as follows (see Table III).

The SM estimation results are shown in Table III. For almost all the input variables, the prediction performance improved as the number of variables increased, and the (B+T+L+A) model notably outperformed the others. Specifically, the proposed LT digitization strategy leads to an accuracy increase of 0.02 cm^3/cm^3 for RMSE and ubRMSE at the global scale. The bias is quite small, and the RMSE and ubRMSE are almost the same (the difference is in the fifth decimal) when using all

TABLE III
COMPARISON OF XGBOOST PERFORMANCE FOR DIFFERENT VARIANTS AND DIFFERENT LTS

LT	XGBoost	RMSE (cm^3/cm^3)	ubRMSE (cm^3/cm^3)	Bias (cm^3/cm^3)
All types	$D+T+L+A$	0.0633	0.0633	0.0001
	$B+T+L+A$	0.0630	0.0630	0.0001
	$B+T+L$	0.0633	0.0633	0.0001
	$B+T+A$	0.0863	0.0863	0.0001
Evergreen needleleaf forest	$D+T+L+A$	0.0479	0.0437	0.0195
	$B+T+L+A$	0.0472	0.0431	0.0192
	$B+T+L$	0.0475	0.0434	0.0193
	$B+T+A$	0.0474	0.0471	0.0056
Evergreen broadleaf forest	$D+T+L+A$	0.0869	0.0791	0.0359
	$B+T+L+A$	0.0865	0.0788	0.0358
	$B+T+L$	0.0868	0.0791	0.0358
	$B+T+A$	0.1373	0.0878	0.1056
Deciduous broadleaf forest	$D+T+L+A$	0.0695	0.0658	0.0226
	$B+T+L+A$	0.0699	0.0661	0.0225
	$B+T+L$	0.0702	0.0664	0.0225
	$B+T+A$	0.0882	0.0708	0.0527
Mixed forest	$D+T+L+A$	0.0690	0.0663	0.0190
	$B+T+L+A$	0.0688	0.0660	0.0192
	$B+T+L$	0.0691	0.0663	0.0192
	$B+T+A$	0.1202	0.0827	0.0873
Closed shrublands	$D+T+L+A$	0.0581	0.0518	0.0263
	$B+T+L+A$	0.0578	0.0514	0.0265
	$B+T+L$	0.0581	0.0517	0.0265
	$B+T+A$	0.0528	0.0517	0.0109
Open shrublands	$D+T+L+A$	0.0455	0.0455	0.0001
	$B+T+L+A$	0.0452	0.0452	0.0001
	$B+T+L$	0.0455	0.0455	0.0001
	$B+T+A$	0.0553	0.0476	0.0280
Woody savannas	$D+T+L+A$	0.0875	0.0875	0.0001
	$B+T+L+A$	0.0872	0.0872	0.0001
	$B+T+L$	0.0875	0.0875	0.0001
	$B+T+A$	0.1446	0.0941	0.1099
Savannas	$D+T+L+A$	0.0758	0.0758	0.0001
	$B+T+L+A$	0.0753	0.0753	0.0001
	$B+T+L$	0.0756	0.0756	0.0001
	$B+T+A$	0.0906	0.0821	0.0384
Grasslands	$D+T+L+A$	0.0858	0.0858	0.0003
	$B+T+L+A$	0.0855	0.0855	0.0003
	$B+T+L$	0.0858	0.0858	0.0003
	$B+T+A$	0.0956	0.0903	0.0314
Permanent wetlands	$D+T+L+A$	0.0909	0.0909	0.0001
	$B+T+L+A$	0.0906	0.0906	0.0001
	$B+T+L$	0.0909	0.0909	0.0001
	$B+T+A$	0.1801	0.0929	0.1542
Croplands	$D+T+L+A$	0.0802	0.0802	0.0001
	$B+T+L+A$	0.0797	0.0797	0.0001
	$B+T+L$	0.0800	0.0800	0.0001
	$B+T+A$	0.1063	0.0866	0.0618
Urban and built-up	$D+T+L+A$	0.0597	0.0595	0.0058
	$B+T+L+A$	0.0588	0.0587	0.0033
	$B+T+L$	0.0591	0.0590	0.0033
	$B+T+A$	0.0848	0.0558	0.0639
Cropland/Natural vegetation mosaic	$D+T+L+A$	0.0974	0.0974	0.0003
	$B+T+L+A$	0.0971	0.0971	0.0003
	$B+T+L$	0.0974	0.0974	0.0003
	$B+T+A$	0.1234	0.0987	0.0741
Barren or sparsely vegetated	$D+T+L+A$	0.0374	0.0374	0.0001
	$B+T+L+A$	0.0373	0.0373	0.0001
	$B+T+L$	0.0376	0.0376	0.0001
	$B+T+A$	0.0704	0.0413	0.0570
Water bodies	$D+T+L+A$	0.0954	0.0945	0.0134
	$B+T+L+A$	0.0954	0.0944	0.0139
	$B+T+L$	0.0957	0.0947	0.0140
	$B+T+A$	0.1187	0.1009	0.0626

data to obtain the performance matrix. We note that the RMSE and ubRMSE in the tables were calculated for an entire dataset, but the distribution map (see Fig. 7) displays the performance for each pixel. Consequently, the figures indicate that the mean ubRMSE is lower than the mean RMSE in almost all areas, contradicting the results in the tables.

The most considerable accuracy improvement, 0.09 cm^3/cm^3 for RMSE in permanent wetland areas with the LT digitization strategy, indicates that the proposed strategy does transfer surface features; i.e., the labels of LTs in the IGBP system effectively represent the land surface and are involved in ML learning, thereby enhancing the performance of the model in GNSS-R SM estimation, in particular, when the SM changes little, and the surface conditions are relatively simple. However, there was almost no improvement in $B+T+L$ for the LT closed shrublands in terms of ubRMSE and RMSE. The reason for this result could be insufficient samples for the land type digitization strategy or variable surface conditions, leading to inaccurate labeling representation. Additionally, in the LT labeling strategy, the accuracy improvement in terms of RMSE is greater than that for ubRMSE, and the RMSE results agree with the ubRMSE results for most land categories. Since ubRMSE eliminates systematic deviation, the dynamic distribution of predicted SM obtained using the LT labeling strategy is remarkably improved and closer to the reference SM. Moreover, the LT classes (e.g., forest) with higher RMSEs often show higher biases and vice versa. This was expected because any deviation and noise in the CYGNSS observations will significantly affect the ML regression. The RMSE and ubRMSE in water bodies are worse than the overall performance since their presence within the footprint can strongly affect the signal.

In addition, outstanding accuracy is achieved for barren and sparsely vegetated LTs, with an RMSE of 0.037 cm^3/cm^3; the number of samples (see nine million in Table II) is higher for these land use types than for all others, indicating that the proposed method is advantageous for handling big data scenarios and suitable for high-resolution global SM estimation. This phenomenon is also observed for open shrubland, which includes over eight million samples; notably, high accuracy of 0.045 cm^3/cm^3 is achieved for RMSE and ubRMSE.

It should be noted that the performance of the variant model differs based on LT. The variant ($B+T+L+A$) model outperforms models with other variable combinations in terms of the RMSE and ubRMSE; thus, it was selected for global SM estimation. The impact factors for different LTs could be determined from the standard deviation (SD) of SM, vegetation indices, and the number of samples; therefore, these variables could be investigated in future studies of several smaller regions.

C. Method Comparison

The performance of the proposed scheme and other ML methods was compared. According to the results in the previous subsection, the variant ($B+T+L+A$) was employed as the default input configuration. The results of the performance comparison for different ML methods are displayed in Table IV.

A significant finding is that the ANN yields worse results than the traditional ML methods (XGBoost and RF). XGBoost outperformed the RF in most cases, and RF performed comparably to or slightly better than XGBoost for certain LTs, such as evergreen needleleaf forest and evergreen broadleaf forest; these two categories were characterized by a small number of samples,

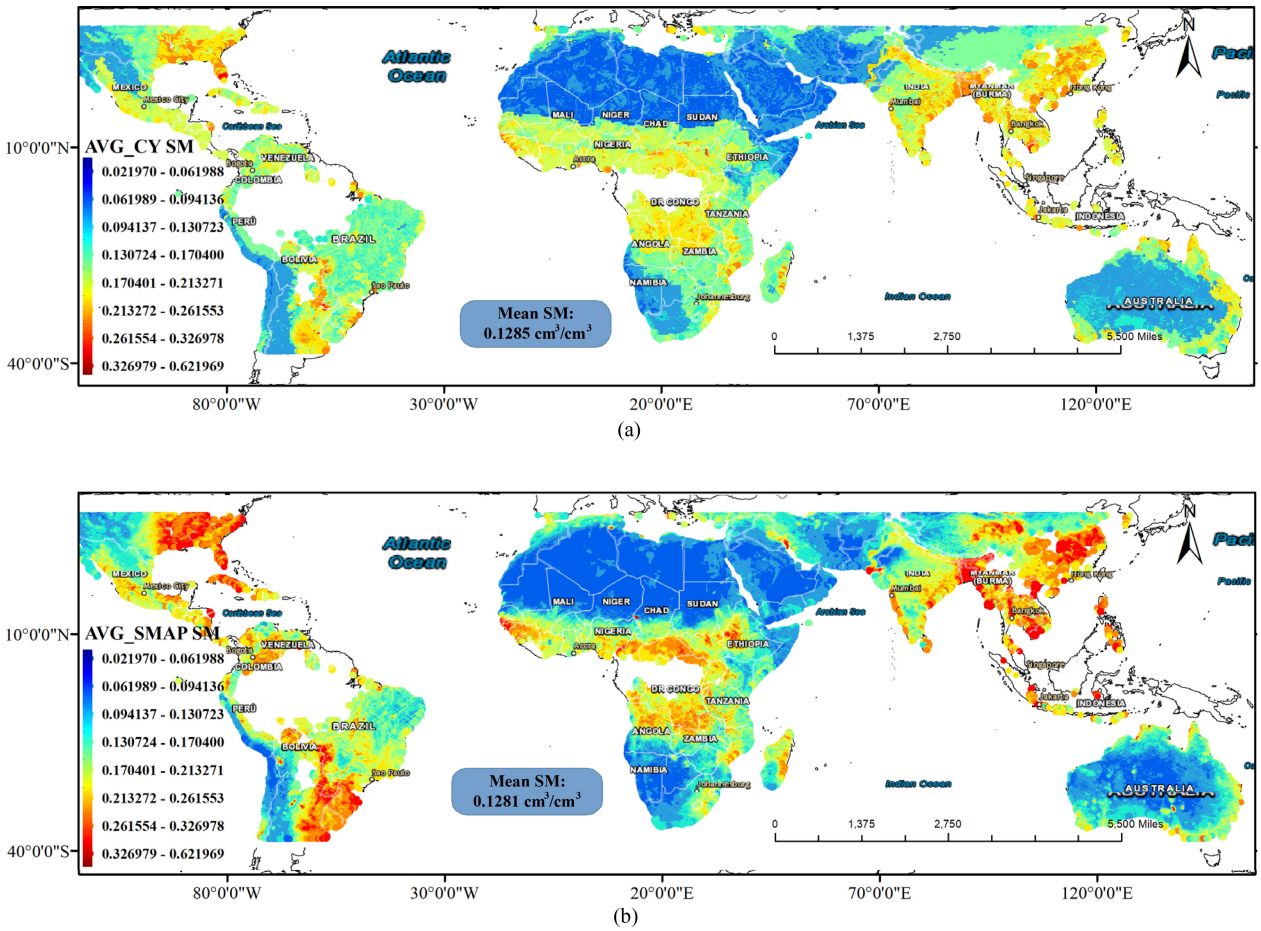


Fig. 7. Predicted daily. (a) Average SM using XGBoost with the LT strategy. (b) Average SMAP SM.

suggesting that the XGBoost algorithm is advantageous over the RF model when handling a large amount of data.

In addition, both XGBoost and RF performed considerably well overall. This finding suggests that for this dataset, sophisticated deep learning models have no substantial advantages over traditional ML methods. This similarity is likely due to the serious overfitting problem that occurs when training deep neural networks with an insufficient number of samples. In general, XGBoost with the LT labeling strategy remarkably outperforms the other competitors based on all indicators. Specifically, compared with the RF method, XGBoost increases the accuracy by 0.002 cm^3/cm^3 for the RMSE and ubRMSE based on the entire dataset. Moreover, it should be noted that different ML methods are suitable for different datasets; that is, the ANN performed well for certain LT categories, as did the RF. This conclusion was previously noted [19]; notably, it was stated that multiple ML models can be built for training in different geological areas to obtain enhanced SM predictions for different terrain types. Furthermore, the ANN achieves an expected accuracy of 0.040 cm^3/cm^3 for the barren or sparsely vegetated LT, with over nine million samples. This finding indicates that the ANN method is good at handling big data tasks and that the proposed XGBoost method yields the best accuracy of 0.037 cm^3/cm^3 compared to the second-best method (RF).

To better visualize the SM estimation results for the optimized method, the CYGNSS and SMAP daily averaged SM distributions for each cell are shown in Fig. 7(a) and (b). Additionally, the RMSE, ubRMSE, and SD results are shown in Fig. 8(a), (b) and (c), respectively. The mean value of the SM based on CYGNSS is 0.1285 cm^3/cm^3, which agrees with the reference SMAP SM of 0.1281 cm^3/cm^3. Another expected result is that the mean ubRMSE is 0.0412 cm^3/cm^3, and the mean RMSE is 0.0578 cm^3/cm^3. This demonstrates that the SM obtained with the XGBoost LT strategy is consistent with the reference SMAP data. It is noted that the legend was classified with a natural breaks model in which classes were based on natural groupings inherent in the data. Breakpoints were identified by picking the class breaks those best grouped similar values and maximized the differences among classes. The features were divided into classes with boundary sets where there were relatively large jumps in values. Thus, Fig. 7(a) and (b) indicates that CYGNSS tends to underestimate values when SM levels are higher than approximately 0.32 cm^3/cm^3, as has been previously reported [14].

In addition, in most areas, high SM values (see Fig. 7) are associated with high RMSE, ubRMSE, and SD values (see Fig. 8). Moreover, the use of ubRMSE enhanced performance since the systematic error was removed, as shown in Fig. 8(b).

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT METHODS WITH THE LT-DIGITIZATION STRATEGY FOR SM ESTIMATION USING 10-FOLD

LT	Methods (B+T+L)	RMSE (cm^3/cm^3)	ubRMSE (cm^3/cm^3)	Bias (cm^3/cm^3)
All types	XGBoost - LT	0.0630	0.0630	0.0001
	RF - LT	0.0651	0.0651	0.0004
	ANN - LT	0.0690	0.0690	0.0008
Evergreen needleleaf forest	XGBoost - LT	0.0472	0.0431	0.0192
	RF - LT	0.0447	0.0447	0.0015
	ANN - LT	0.0469	0.0469	0.0006
Evergreen broadleaf forest	XGBoost - LT	0.0865	0.0788	0.0358
	RF - LT	0.0772	0.0772	0.0002
	ANN - LT	0.0857	0.0857	0.0007
Deciduous broadleaf forest	XGBoost - LT	0.0699	0.0661	0.0225
	RF - LT	0.0642	0.0641	0.0029
	ANN - LT	0.0721	0.0721	0.0026
Mixed forest	XGBoost - LT	0.0688	0.0660	0.0192
	RF - LT	0.0637	0.0637	0.0007
	ANN - LT	0.0700	0.0700	0.0004
Closed shrublands	XGBoost - LT	0.0578	0.0514	0.0265
	RF - LT	0.0470	0.0470	0.0008
	ANN - LT	0.0507	0.0507	0.0001
Open shrublands	XGBoost - LT	0.0452	0.0452	0.0001
	RF - LT	0.0452	0.0452	0.0004
	ANN - LT	0.0495	0.0495	0.0009
Woody savannas	XGBoost - LT	0.0872	0.0872	0.0001
	RF - LT	0.0873	0.0873	0.0001
	ANN - LT	0.0950	0.0950	0.0005
Savannas	XGBoost - LT	0.0753	0.0753	0.0001
	RF - LT	0.0754	0.0754	0.0010
	ANN - LT	0.0824	0.0824	0.0011
Grasslands	XGBoost - LT	0.0855	0.0855	0.0003
	RF - LT	0.0854	0.0854	0.0007
	ANN - LT	0.0946	0.0946	0.0006
Permanent wetlands	XGBoost - LT	0.0906	0.0906	0.0001
	RF - LT	0.0932	0.0930	0.0058
	ANN - LT	0.1005	0.1004	0.0034
Croplands	XGBoost - LT	0.0797	0.0797	0.0001
	RF - LT	0.0798	0.0798	0.0021
	ANN - LT	0.0873	0.0873	0.0010
Urban and built-up	XGBoost - LT	0.0588	0.0587	0.0033
	RF - LT	0.0639	0.0632	0.0093
	ANN - LT	0.0768	0.0744	0.0193
Cropland/Natural vegetation mosaic	XGBoost - LT	0.0971	0.0971	0.0003
	RF - LT	0.0970	0.0970	0.0005
	ANN - LT	0.1057	0.1057	0.0013
Barren or sparsely vegetated	XGBoost - LT	0.0373	0.0373	0.0001
	RF - LT	0.0373	0.0373	0.0002
	ANN - LT	0.0405	0.0405	0.0007
Water bodies	XGBoost - LT	0.0954	0.0944	0.0139
	RF - LT	0.0941	0.0941	0.0012
	ANN - LT	0.1020	0.1020	0.0010

Fig. 9 shows examples of density plots (XGBoost model, RF model, and ANN model) on a log scale comparing the CYGNSS-based SM and the reference SM. The performance of the SM test dataset (three million samples) for each model with the LT digitization strategy is shown. The density plot demonstrates an overall fairly good consistency between the CYGNSS-based SM and reference SM from SMAP, especially when samples are abundantly available.

Notably, the XGBoost model [see Fig. 9(a)] exhibits the highest R of 0.71. Meanwhile, we noticed the data cloud in the density plot looks like many bold horizontal lines stacked along the 1:1 line. This data cloud demonstrates the principles of the digitized LTs/labeling strategy that possesses the significant effects of classification, which is working for redistributing the samples into groups (lines). In this way, the learning model can more easily search and build the rules between the inputs and outputs, which can provide more accurate results. This is the reason that the digitized LT strategy works well in global SM estimation. A similar feature can be seen in Fig. 9(b). Such pattern in Fig. 9(c) is not very obvious. The R value in Fig. 9(c) is also lower than that in Fig. 9(a) and (b). From the comparison of the three figures, it can also be seen that the labeling method can effectively improve the correlation of prediction.

Moreover, when the data density is low, a tendency to deviate from the line is displayed. This result demonstrates that the CYGNSS product tends to underestimate SM values to some degree, as has been observed in previous figures (see Fig. 7) and was previously reported in [14]. We notice that surfaces with high SM usually have dense vegetation growth and high moisture contents, resulting in an increase in a variety of incoherent components and a decrease incoherent components. In such cases, the change in reflectivity cannot be used to completely and correctly express the change in SM, and there is a positive correlation between SM and the coherent components of the signal. As a result, the learning model cannot correctly extract the characteristics of surfaces with high SM, so the SM predicted by the model is low.

D. Effect of the Proposed Scheme on Other Models

To further verify whether the proposed scheme is effective, we trained each variable combination without the LT strategy (see Table V). The variant without labels is B+T+A, and the variant with the LT strategy is B+T+L.

The proposed digitalization LT strategy may be widely effective for use with different neural networks. In this section, we compare its performance when combined with RF and ANN models. The outstanding performance of RFs has been previously discussed, and ANNs have been used in many previous studies. Both learning models were trained on the same dataset, and the exact labeling configurations were adopted. The 10-fold CV method was again adopted.

We observe that for all methods, the SM estimation accuracy of the digitization LT strategy is improved. In particular, for the RF, the digitization LT strategy results in an increase of 0.021 cm^3/cm^3 for the RMSE and ubRMSE for the entire dataset. For the ANN, the digitization LT strategy performed better, with an accuracy improvement of 0.024 cm^3/cm^3 for the RMSE and ubRMSE. The results demonstrate that the proposed LT strategy also works for SM estimation with neural networks and indicate that the risk of overfitting when an ANN is used can be largely reduced by applying a labeling model.

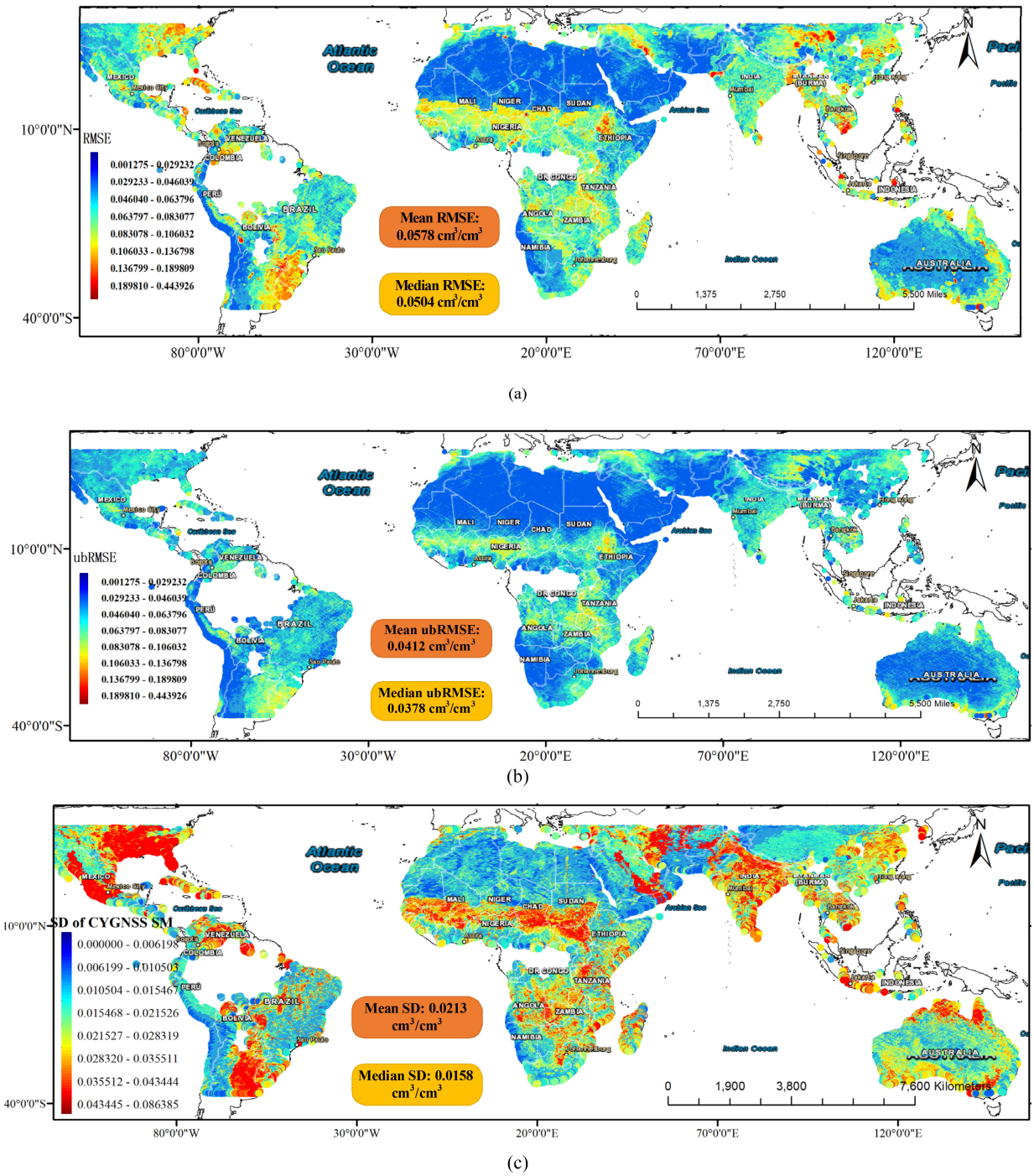


Fig. 8. Predicted daily. (a) RMSE distribution for SM estimation using XGBoost with the LT strategy. (b) ubRMSE distribution for SM estimation using XGBoost with the LT strategy. (c) SD distribution for SM estimation using XGBoost with the LT strategy.

In addition, the performance gain of the RF is larger than that of the ANN for certain LT categories, indicating that the proposed digitization LT strategy scheme appears to be potentially more effective for traditional ML networks than for ANNs. This result may be attributed to the ability of RFs to eliminate the effects of unbalanced data and missing data. The labeling strategy, which incorporates the idea of classification, enhances the performance of the RF method more than that observed for the ANN, even though ANNs are more advantageous for handling large datasets.

E. Effects of Land Types

We also investigated the accuracy of the SM model and considered incorrect input LT labels to perform analyses that demonstrated the effect of compressing vegetation and surface

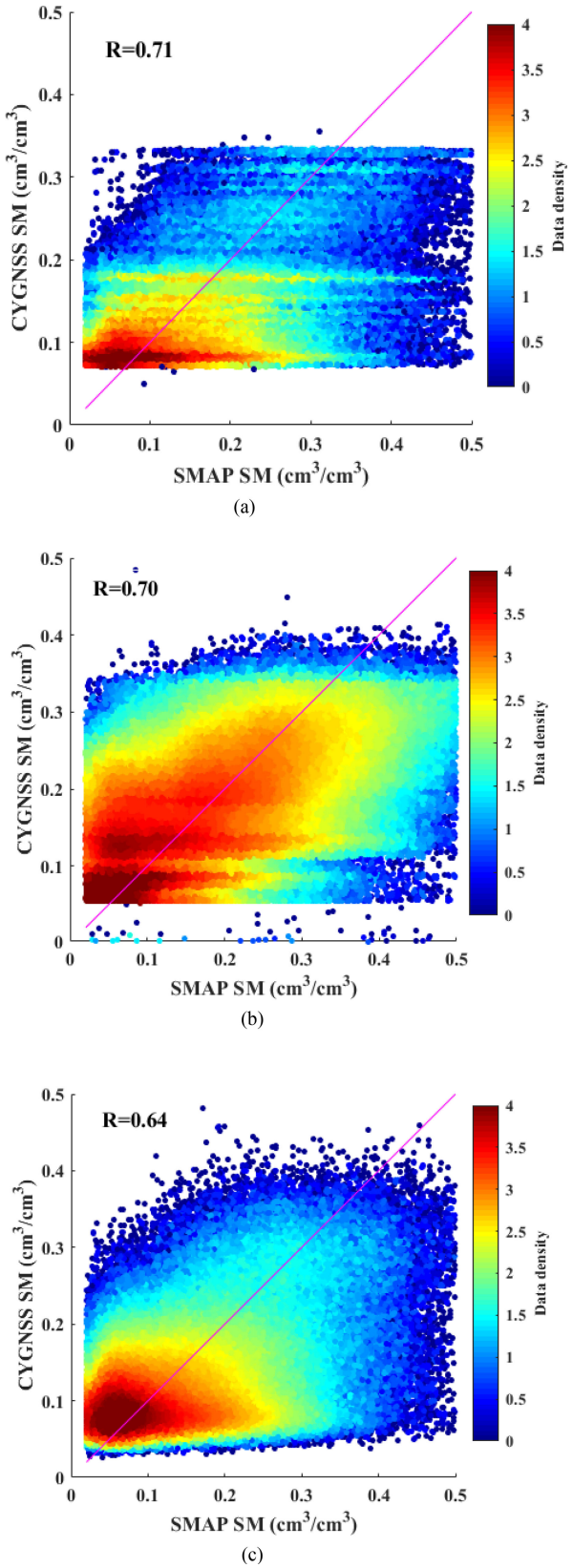


Fig. 9. Density plot with 10-fold CV. (a) XGBoost for SM estimation using the LT strategy. (b) RF for SM estimation using the LT strategy. (c) ANN for SM estimation using the LT strategy.

TABLE V
PERFORMANCE MATRIX OF DIFFERENT METHODS WITH/WITHOUT LT STRATEGY FOR SM ESTIMATION

LT	Methods (*B+T*)	RMSE (cm^3/cm^3)	ubRMSE (cm^3/cm^3)	Bias (cm^3/cm^3)
All types	RF+ *LT*	0.0654	0.0654	0.0004
	RF+*A*	0.0863	0.0863	0.0001
	ANN+*LT*	0.0693	0.0693	0.0008
	ANN+ *A*	0.0932	0.0932	0.0019
Evergreen needleleaf forest	RF+ *LT*	0.0450	0.0450	0.0015
	RF+*A*	0.0471	0.0467	0.0061
	ANN+*LT*	0.0472	0.0472	0.0006
	ANN+ *A*	0.0569	0.0567	0.0045
Evergreen broadleaf forest	RF+ *LT*	0.0775	0.0775	0.0002
	RF+*A*	0.1372	0.0877	0.1055
	ANN+*LT*	0.0860	0.0860	0.0007
	ANN+ *A*	0.1404	0.0947	0.1037
Deciduous broadleaf forest	RF+ *LT*	0.0645	0.0644	0.0029
	RF+*A*	0.0882	0.0708	0.0527
	ANN+*LT*	0.0724	0.0724	0.0026
	ANN+ *A*	0.0938	0.0786	0.0512
Mixed forest	RF+ *LT*	0.0640	0.0640	0.0007
	RF+*A*	0.1204	0.0829	0.0873
	ANN+*LT*	0.0703	0.0703	0.0004
	ANN+ *A*	0.1241	0.0903	0.0852
Closed shrublands	RF+ *LT*	0.0473	0.0473	0.0008
	RF+*A*	0.0524	0.0511	0.0114
	ANN+*LT*	0.0510	0.0510	0.0001
	ANN+ *A*	0.0617	0.0610	0.0091
Open shrublands	RF+ *LT*	0.0455	0.0455	0.0004
	RF+*A*	0.0553	0.0476	0.0281
	ANN+*LT*	0.0498	0.0498	0.0009
	ANN+ *A*	0.0656	0.0584	0.0299
Woody savannas	RF+ *LT*	0.0876	0.0876	0.0001
	RF+*A*	0.1446	0.0941	0.1098
	ANN+*LT*	0.0953	0.0953	0.0005
	ANN+ *A*	0.1469	0.0996	0.1080
Savannas	RF+ *LT*	0.0757	0.0757	0.0010
	RF+*A*	0.0906	0.0822	0.0382
	ANN+*LT*	0.0827	0.0827	0.0011
	ANN+ *A*	0.0958	0.0886	0.0364
Grasslands	RF+ *LT*	0.0857	0.0857	0.0007
	RF+*A*	0.0958	0.0903	0.0318
	ANN+*LT*	0.0949	0.0949	0.0006
	ANN+ *A*	0.1012	0.0967	0.0297
Permanent wetlands	RF+ *LT*	0.0935	0.0933	0.0058
	RF+*A*	0.1792	0.0931	0.1531
	ANN+*LT*	0.1008	0.1007	0.0034
	ANN+ *A*	0.1837	0.1034	0.1518
Croplands	RF+ *LT*	0.0801	0.0801	0.0021
	RF+*A*	0.1064	0.0867	0.0618
	ANN+*LT*	0.0876	0.0876	0.0010
	ANN+ *A*	0.1120	0.0946	0.0599
Urban and built-up	RF+ *LT*	0.0642	0.0635	0.0093
	RF+*A*	0.0851	0.0548	0.0651
	ANN+*LT*	0.0771	0.0747	0.0194
	ANN+ *A*	0.0959	0.0694	0.0662
Cropland/Natural vegetation mosaic	RF+ *LT*	0.0973	0.0973	0.0005
	RF+*A*	0.1233	0.0986	0.0740
	ANN+*LT*	0.1060	0.1060	0.0013
	ANN+ *A*	0.1272	0.1048	0.0722
Barren or sparsely vegetated	RF-LT	0.0376	0.0376	0.0002
	RF	0.0702	0.0410	0.0569
	ANN-LT	0.0408	0.0408	0.0007
	ANN	0.0804	0.0547	0.0590
Water bodies	RF-LT	0.0944	0.0944	0.0012
	RF	0.1186	0.1009	0.0622
	ANN-LT	0.1023	0.1023	0.0010
	ANN	0.1227	0.1069	0.0603

TABLE VI
PERFORMANCE MATRIX FOR THE EVALUATION OF INPUT LTS

LT, XGBoost (*B+T+L+A*)	RMSE (cm^3/cm^3)		ubRMSE (cm^3/cm^3)	
	Without changes	Changing LT labels	Without changes	Changing LT labels
All types	**0.0630**	**0.0642**	**0.0630**	**0.0642**
Evergreen needleleaf forest	0.0472	0.0472	0.0431	0.0431
Evergreen broadleaf forest	0.0865	0.0865	0.0788	0.0788
Deciduous broadleaf forest	0.0699	0.0699	0.0661	0.0661
Mixed forest	0.0688	0.0688	0.0660	0.0660
Closed shrublands	0.0578	0.0578	0.0514	0.0514
Open shrublands	**0.0452**	**0.0451**	**0.0452**	**0.0451**
Woody savannas	0.0872	0.0872	0.0872	0.0872
Savannas	0.0753	0.0753	0.0753	0.0753
Grasslands	0.0855	0.0855	0.0855	0.0855
Permanent wetlands	0.0906	0.0906	0.0906	0.0906
Croplands	0.0797	0.0797	0.0797	0.0797
Urban and built-up	0.0588	0.0588	0.0587	0.0587
Cropland/Natural vegetation mosaic	0.0971	0.0971	0.0971	0.0971
Barren or sparsely vegetated	**0.0373**	**0.0583**	**0.0373**	**0.0444**
Water bodies	0.0954	0.0954	0.0944	0.0944

TABLE VII
COMPUTATION SPEEDS OF DIFFERENT METHODS WITH THE LT STRATEGY

Methods (*B+T+L+A*)	Min/Sample (10-fold CV)
XGBoost_LT	420
RF_LT	1320
ANN_LT	1020

roughness descriptions into a single land classification variable. In other words, we examined the SM estimation accuracy when the LT labels of the testing samples and the established model did not match. Here, we selected two LT (open shrublands and barren or sparsely vegetated) datasets to exchange LT labels. These two LT datasets have a larger amount of data and perform well for SM estimation. The labels for these two types were intentionally exchanged to "16" and "7" for testing, which were supposed to be "7" and "16". Labels with other LTs remained constant and were regarded as the testing samples to obtain the corresponding RMSE, and the results are shown in Table VI.

All of the data, including the testing LT "open shrublands" and "barren or sparsely vegetated" samples, were tested. The performance for the entire dataset decreased, i.e., the accuracy decreased from 0.0630 cm^3/cm^3 to 0.0642 cm^3/cm^3 for the RMSE and ubRMSE after exchanging the labels in the testing datasets. This shows that for these two LTs, the RMSE (from 0.0373 cm^3/cm^3 to 0.0583 cm^3/cm^3) and ubRMSE (from 0.0373 cm^3/cm^3 to 0.0444 cm^3/cm^3) worsened for "barren or sparsely vegetated". The performance of "open shrublands" did not change much. This finding demonstrates the importance and significance of the LT datasets and provides compelling evidence for proposing the digitized LT concept. The contributions and weight of the digitized LT numbers could be another subject to be investigated in future work.

F. Computational Efficiency

In practice, high-resolution global SM estimation often involves millions of samples. Therefore, the computational efficiency of ML and neural network models should be taken into consideration. In this section, we compare the computational speeds of several models. The results are given in Table VII. According to the results, XGBoost is the fastest model, followed by the ANN model, and RF is the slowest. Specifically, XGBoost is approximately 68% faster than the RF model. Additionally, XGBoost, which has displayed competitive performance in previous experiments [30], [31], is 58.8 faster than the ANN. Moreover, XGBoost outperformed the other methods in SM estimation. Therefore, XGBoost is most suitable for high-resolution global SM estimation because it possesses high accuracy and efficiency at once.

V. CONCLUSION

In this article, an SM data product from two data sources (CYGNSS and SMAP products) is achieved daily with a spatiotemporal resolution of 9×9 km. A simple but powerful LT digitization strategy clearly shows the improvement of SM estimation, which incorporates the idea of classification and can avoid the use of complex data from multiple sources and learn robust and complete geographic surface information. The number of input variants is set based on a combination of three variables derived from the CYGNSS product and processed with a fine-tuning step to achieve optimal performance. We also introduce the XGBoost ML algorithm for SM estimation. Evaluations of different LTs reveal that the proposed LT digitization/labeling strategy is generally effective for different ML algorithms, including RF, ANN, and XGBoost models. The comparison of the results with/without LT labeling supports our design and indicates that the digitization/labeling of LTs can improve the representation of surface conditions and, thus, enhance the performance and generalization ability of models in global SM estimation. In addition, the results demonstrate that XGBoost outperforms the RF and ANN models in global SM estimation. The proposed approach can be easily implemented and adjusted for small- and large-scale surfaces with very low costs since no other data sources are required beyond the CYGNSS data. Once trained on reference samples (SM reference data), the model does not require SM information from other sources for SM estimation. This approach allows the learning model to be trained with any SM source data, and SM predictions are performed independently of complex ancillary data for stand-alone SM estimation. In the future, better results could be obtained by increasing the number of samples for certain LTs, and the incorporation of matching spatial information should be investigated. Although this work focuses on using CYGNSS reflectivity, the proposed LT labeling scheme could be extended to other CYGNSS observables in future work, e.g., BRCS/NBRCS.

The highlights are as follows.

1) A practical and effective strategy that directly extracts and uses complete LT information (LT digitization) in an ML framework.
2) A comparative study with different LTs and variable combinations is conducted as a multiple-feature optimization process for CYGNSS ML-based SM estimation.
3) The model can achieve good performance while requiring the least ancillary data among several state-of-the-art models.
4) Our SM product provides faster revisit times and is verified with multiple ML methods on a global scale with a 9×9 km resolution.
5) The approach is applicable to other ML-based regression tasks.

ACKNOWLEDGMENT

The authors would like to thank to the CYGNSS and SMAP Team for providing the utilized data available.

REFERENCES

[1] 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, 2002.
[2] W. Li, E. Cardellach, F. Fabra, A. Rius, S. Ribó, and M. Martín-Neira, "First spaceborne phase altimetry over sea ice using Techdemosat-1 GNSS-R signals," *Geophysical Res. Lett.,* vol. 44, no. 16, pp. 8369–8376, 2017.
[3] Q. Yan and W. Huang, "Sea ice thickness measurement using spaceborne GNSS-R: First results with Techdemosat-1 data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 577–587, Jan. 2020.
[4] 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, Mar. 2018.
[5] E. G. Njoku, T. J. Jackson, V. Lakshmi, T. K. Chan, and S. V. Nghiem, "Soil moisture retrieval from AMSR-E," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 2, pp. 215–229, Feb. 2003.
[6] D. Entekhabi *et al.*, "The soil moisture active passive (SMAP) mission," *Proc. IEEE*, vol. 98, no. 5, pp. 704–716, May 2001.
[7] Y. H. Kerr *et al.*, "The SMOS soil moisture retrieval algorithm," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 5, pp. 1384–1403, May 2012.
[8] S. Chan *et al.*, "Assessment of the SMAP passive soil moisture product," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4994–5007, Aug. 2016.
[9] S. Gleason, A. O'Brien, A. Russel, M. M. Al-Khaldi, and J. T. Johnson, "Geolocation, calibration and surface resolution of spaceborne GNSS-R land observations," *Remote Sens.*, vol. 12, no. 8, 2020, Art. no. 1317.
[10] 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, 2018.
[11] M. M. Al-Khaldi, J. T. Johnson, A. J. O'Brien, A. Balenzano, and F. Mattia, "Time-series retrieval of soil moisture using CYGNSS," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4322–4331, Jul. 2019.
[12] 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, 2018.
[13] C. Chew and E. E. Small, "Description of the UCAR/CU soil moisture product," *Remote Sens.*, vol. 12, 2020, Art. no. 1558.
[14] 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.
[15] 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, 2020, Art. no. 122.
[16] 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, 2020, Art. no. 111944.
[17] T. Yang, W. Wan, Z. Sun, B. Liu, and X. Chen, "Comprehensive evaluation of using Techdemosat-1 and CYGNSS data to estimate soil moisture over Mainland China," *Remote Sens.*, vol. 12, no. 11, 2020, Art. no. 1699.
[18] O. Eroglu, M. Kurum, D. Boyd, and A. C. Gurbuz, "High spatio-temporal resolution CYGNSS soil moisture estimates using artificial neural networks," *Remote Sens.*, vol. 11, no. 19, pp. 2272–2303, 2019.
[19] 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, 2020, Art. no. 1168.
[20] V. Senyurek, F. Lei, D. Boyd, A. C. Gurbuz, and R. Moorhead, "Evaluations of a machine learning-based CYGNSS soil moisture estimates against SMAP observations," *Remote Sens.,* vol. 12, no. 21, 2020, Art. no. 3503.
[21] CYGNSS Handbook, *Cyclone Global Navigation Satellite System: Deriving Surface Wind Speeds in Tropical Cyclones*. Ann Arbor, MI, USA: Univ. Michigan, 2016.
[22] 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, 2016.
[23] 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, 2019, Art. no. 1053.
[24] S. Chan and S. Dunbar, "Level 3 passive soil moisture product specification document," Tech. Rep. JPL D-72551, 2019.
[25] Y. Jia *et al.*, "Temporal-spatial soil moisture estimation from CYGNSS using machine learning regression with a pre-classification approach," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4879–4893, Apr. 2021, doi: 10.1109/JSTARS.2021.3076470.
[26] K. Jensen, K. M. Donald, E. Podest, N. Rodriguez-Alvarez, V. Horna, and N. Steiner, "Assessing L-band GNSS-reflectometry and imaging radar for detecting sub-canopy inundation dynamics in a tropical wetlands complex," *Remote Sens.,* vol. 10, no. 9, 2018, Art. no. 1431.
[27] 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.
[28] Q. Yuan *et al.*, "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, 2020, Art. no. 111716.
[29] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[30] Y. Jia, S. Jin, P. Savi, Y. Gao, and W. Li, "GNSS-R soil moisture retrieval based on a XGboost machine learning aided method: Performance and validation," *Remote Sens.*, vol. 11, no. 14, 2019, Art. no. 1655.
[31] C. Wei and C. Hsu, "Extreme gradient boosting model for rain retrieval using radar reflectivity from various elevation angles," *Remote Sens.*, vol. 12, no. 14, 2020, Art. no. 2203.

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

She is currently working with the Nanjing University of Posts and Telecommunications. In 2013, she was with the Department of Electronics and Telecommunications, Politecnico di Torino, Torino, Italy, where she performed research on the GNSS system construction and GNSS antenna analysis. 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, 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 from Wuhan University, Wuhan, China, in 1999, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2003, both in geodesy.

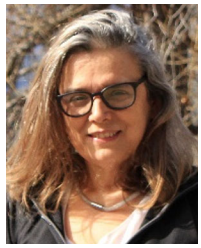
He is Vice-President and Professor with Henan Polytechnic University, Jiaozuo, China and also Professor with Shanghai Astronomical Observatory, CAS, Shanghai, China. He has authored and coauthored more than 500 papers in peer-reviewed journals and proceedings, ten patents/software copyrights, and ten books/monographs with more than 8000 citations and H-index>50. His main research interests include satellite navigation, remote sensing, and space/planetary exploration.

Prof. Jin has been President of International Association of Planetary Sciences (2015–2019), President of the International Association of CPGPS (2016–2017), Chair of IUGG Union Commission on Planetary Sciences (2015–2023), 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 100-Talent Program of CAS, Leading Talent of Shanghai, IAG Fellow, IUGG Fellow, Fellow of Electromagnetics Academy, World Class Professor of Ministry of Education and Cultures, Indonesia, Chief Scientist of National Key R&D Program, China, Member of Russian Academy of Natural Sciences, Member of European Academy of Sciences, Member of Turkish Academy of Sciences, and Member of Academia Europaea.

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, respectively.

He is currently with the School of Remote Sensing and Geomatics Engineering from the Nanjing University of Information Science and Technology. His research interests include tsunami, sea ice, and land remote sensing using Global Navigation Satellite System-Reflectometry.

Dr. Yan was a recipient of 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 1985.

In 1986, she was a consultant in Alenia (Caselle Torinese, Italy) where she conducted research on the analysis and design of dielectric radomes. From 1987 to 1998, she was a Researcher with the Italian National Research Council. In 1998, she was with the Electronic Department, Politecnico di Torino, as an Associate Professor. She currently teaches a course on electromagnetic field theory. Recently, she is focused on the analysis and characterization at microwave frequency of novel materials (polymers and cements with carbon nanotubes, graphene or biochar as fillers) for various applications. Her areas of research interests include dielectric radomes, frequency-selective surfaces, waveguide discontinuities and microwave filters, high-altitude platform propagation channels, and Global Navigation Satellite System Reflectometry for soil moisture retrieval.

Prof. Savi is currently member of Società Italiana di Elettromagnetismo.

Rongchun Zhang received the B.S. degree in geomatics engineering, the M.S. degree in photogrammetry and remote sensing, and the Ph.D. degree in geodesy and survey engineering from Hohai University, Nanjing, China, in 2008, 2012, and 2017, respectively.

She is a Lecturer with the Nanjing University of Posts and Telecommunications, Nanjing, ChinaShe has authored or coauthored more than 20 research papers. Her current research interests include photogrammetry, multisource data fusion technology, and computer vision.

Wenmei Li (Member, IEEE) received the M.S. and Ph.D. degrees from the Nanjing University and Chinese Academy of Forestry, Nanjing, China, in 2010 and 2013, respectively.

She is currently an Associate Professor with the School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing, China. She is working for her Postdoctoral Studies (2018-) with Nanjing University of Posts and Telecommunications. Her research interests include deep learning, optimization, image reconstruct, and their application in land remote sensing.