Near Real-Time Soil Moisture in China Retrieved From CyGNSS Reflectivity

Qingyun Yan, *Member, IEEE*, Shaoqi Gong, Shuanggen Jin, *Member, IEEE*, Weimin Huang, *Senior Member, IEEE*, and Cunjie Zhang

***Abstract*— This work presents a novel scheme to retrieve soil moisture (SM) from the Cyclone Global Navigation Satellite System (CyGNSS) data, which is accomplished by using a bagged regression trees (BRT) algorithm with the inputs being the CyGNSS-derived products, the corresponding geolocation, and associated climate type. This algorithm is validated with the *in situ* hourly SM data acquired by China's automatic SM observation stations throughout the year 2018. High consistency between the retrieved SM results and the measured SM is achieved, with a correlation coefficient of 0.86 and a root-mean-square error of 0.05 cm^3/cm^3. The results obtained in this work indicate that the proposed BRT-based method can effectively estimate SM from CyGNSS data in different scenarios of various station locations and climate types in a near real-time manner.**

***Index Terms*— Bagged regression trees (BRT), climate type, cyclone global navigation satellite system (CyGNSS), global navigation satellite system-reflectometry (GNSS-R), soil moisture (SM).**

I. Introduction

SOIL moisture (SM) is critical for the hydrological, geophysical, and agricultural processes due to its significance in the Earth's water cycle [1]. Thus, the knowledge of SM is vital for hydrologists, ecologists, agriculturalists, and climatologists to improve hydrological and environmental models/predictions. Remote sensing techniques have been widely adopted as an excellent tool for obtaining SM data in a cost-effective and efficient way.

Recently, Global Navigation Satellite System (GNSS)-Reflectometry (GNSS-R) has emerged as a promising remote sensing approach, which is able to provide all-day and all-weather surveillance. It has successfully demonstrated its capability in sea surface wind speed estimation [2], altimetry [3], sea ice sensing [4]–[6], tsunami detection [7], and wetland classification [8]. Due to the launch of the Cyclone GNSS (CyGNSS) mission in 2016, massive data with fine spatial and temporal resolutions are available for the public. In particular, the study of SM retrieval from CyGNSS data has been a topic of interest [9]–[14]. The fluctuations in the CyGNSS signal-to-noise ratio (SNR) were found to be correlated with the SM Active Passive (SMAP) SM results in [9]. An SM estimation method based on the relative SNR was presented in [10]. Clarizia *et al.* [11] proposed a reflectivity-vegetation-roughness (R-V-R) algorithm and achieved SM retrieval using a trilinear regression function. SM estimation through time-series analysis was performed in [12]. It is worth noting that these studies were evaluated with SMAP SM products of coarse spatial resolutions. To demonstrate the capacity of CyGNSS for retrieving SM at high spatiotemporal scales, an artificial neural network (ANN)-based approach was designed in [13] and the results were assessed with field data of fine resolution, and its extended work was presented in [14] and validated over larger and more diverse data sets (from over 100 International SM Network sites in the Continental United State). However, these two studies heavily rely on a bunch of ancillary data. In this work, we propose an effective model that incorporates the CyGNSS-derived surface reflectivity (Γ), bistatic radar cross section (BRCS or σ), and coherence flag as well as the corresponding geolocation, and its climate type for estimating hourly SM. Furthermore, we validate this model with the *in situ* SM data obtained by China's automatic SM observation stations (with more than 1700 sites). The assessment is proceeded on an hourly basis demonstrating the capacity of CyGNSS for near real-time SM retrieval.

This letter is organized as follows. Section II introduces the employed CyGNSS and reference SM data. Section III describes the proposed bagged regression trees (BRT)-based SM estimation scheme. The experimental evaluation and associated discussions are presented in Section IV. Section V gives a summary and possible future work of this study.

Manuscript received June 9, 2020; revised September 14, 2020 and October 31, 2020; accepted November 17, 2020. Date of publication December 1, 2020; date of current version January 4, 2022. The work of Qingyun Yan was supported in part by the National Natural Science Foundation of China under Grant 42001362 and in part by the Research Startup Fund of NUIST under Grant 2020R078. The work of Shuanggen Jin was supported in part by the Strategic Priority Research Program Project of the Chinese Academy of Sciences under Grant XDA23040100, in part by the Shanghai Leading Talent Project under Grant E056061, in part by the Jiangsu Province Distinguished Professor Project under Grant R2018T20, and in part by the Startup Foundation for Introducing Talent of NUIST under Grant 2018R037. The work of Weimin Huang was supported in part by the Natural Sciences and Engineering Research Council of Canada Discovery under Grant NSERC RGPIN-2017-04508 and Grant RGPAS-2017-507962 and in part by the Canadian Space Agency CubeSat under Grant 17CCPNFL11. The work of Cunjie Zhang was supported by the National Research and Development Program of China under Grant 2020YFA0608203. *(Corresponding author: Shaoqi Gong.)*

Qingyun Yan, Shaoqi Gong, and Shuanggen Jin are with the School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: shaoqigong@163.com).

Weimin Huang is with the Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1B 3X5, Canada.

Cunjie Zhang is with the National Climate Center, China Meteorological Administration, Beijing 100081, China.

Digital Object Identifier 10.1109/LGRS.2020.3039519

II. Data Description and Study Region

In this section, the acquisition of CyGNSS remote sensing data is first described. Next, the processing of the reference *in situ* SM data along with the study region is introduced.

A. CyGNSS Remote Sensing Data

The CyGNSS constellation consists of eight microsatellites, and each of them is able to provide four GNSS-R measurements from different locations simultaneously. Since the data are processed every second, 32 separate measurements can be obtained per second. In addition, the achievable revisit time is within several hours. Therefore, CyGNSS is capable of monitoring SM with extensive spatial coverage and high temporal resolution within latitudes between ±38°. The data employed in this work span the year 2018.

The CyGNSS metadata include the BRCS and SNR at each specular point (SP) as well as their associated information about the measuring geometry and navigation message, such as the incidence angle, coordinates of SP, distances from SP to the transmitter and receiver, and so on. In this work, data collected over land with an SNR over 0 dB at SP are retained. It is worth mentioning that the CyGNSS BRCS is stored in a signal box with 17 delay × 11 Doppler bins. The peak power position is varying due to the change of terrain elevation. To ensure that the error in the CyGNSS SP location estimation is within a reasonable range, only the BRCS data with a peak position between the 4th and 15th bins in the delay axis are persevered. A similar process has been done in, e.g., [13].

B. Reference Data

In situ SM data collected by China's automatic SM observation stations are used as the reference data. This observation network is composed of more than 2600 sites, and about 1900 of them are spread over the area with CyGNSS's coverage (see Fig. 1). Each site provides hourly SM measurement from 0 to 100 cm depth below the soil surface with an interval of 10 cm. The penetration depth of the GNSS-R signals in soils can vary from several centimeters up to about 20 cm, depending on SM and soil type [15]. Therefore, this work uses 10-cm SM data that are regarded as the "optimum" value as it covers down to 0.1 cm^3/cm^3 of SM. The hourly data in each day are utilized and regarded as the ground truth in this study. In addition, the geolocation information (including latitude, longitude, and altitude) of the site is also provided and employed as input after the data collocation that is described in Section IV-A.

Considering the coverage of CyGNSS and the distribution of SM observation stations in China, the study region goes from 18° to 38° N and 75.9° to 132.5° E. Due to the wide terrain within China, there are several different climate types, and each of them has a distinct characteristic in the temporal and spatial variability of, e.g., temperature and precipitation, which eventually affects SM. Thus, the impact of climate type on SM retrieval is evaluated in this work, and such data are adapted from [16]. The climate types in the study region include the mountain plateau climate (MPC), tropical monsoon climate (TrMC), temperate monsoon climate (TeMC), subtropical monsoon climate (SMC), and temperate continental climate (TCC). Their spatial distribution is shown in Fig. 1. Topography at the sites mainly includes plain, plateau, and basin.

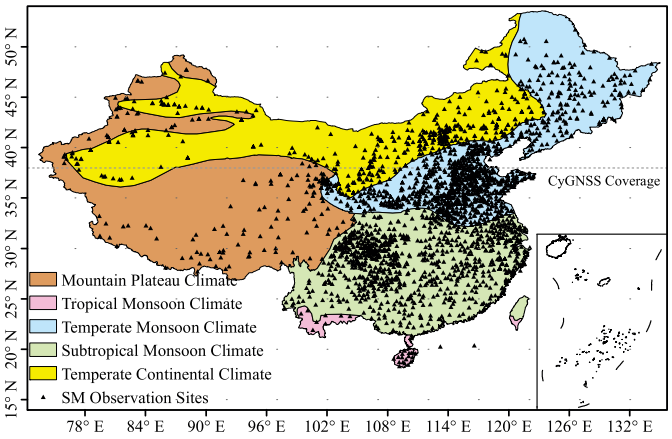

Fig. 1. Distributions of automatic SM observation stations and the climate types in China.

III. Estimation Method

In this section, a detailed procedure for estimating SM from CyGNSS data is described, including deriving surface reflectivity Γ along with coherence flag and implementing the BRT-based SM retrieval.

A. Derivation of Reflectivity From CyGNSS

It is widely accepted that the signals at microwave frequencies are sensitive to the dielectric constant of soil that is a function of SM [17], which indicates the connection between the Fresnel reflection coefficient ($\Re$) and SM. For a flat and smooth region covered by vegetation, the surface reflectivity Γ can be modeled as [18], [19]

$$\Gamma(\theta) = \Re(\theta)^2 \gamma^2 \exp(-4k^2 s^2 \cos^2(\theta)) \tag{1}$$

where θ is the incidence angle, the transmissivity γ accounts for the attenuation of signal propagation by vegetation, and the exponential term represents the surface roughness effects with k being the signal wavenumber and s being the surface root-mean-squared height. In this work, the effect of surface roughness is considered using the coherence flag that describes power spreading in CyGNSS data (see more details in [20]). Although the impact of vegetation cover can be strong and vary significantly from site to site [21]. However, in this work, the retrieval is based on each site by inputting its specified geolocation. At each site, the impact of vegetation is insignificant since it only alters the retrieval performance statistics by about 1% [12], and for this reason, the impact of vegetation cover is neglected.

By following the assumption of coherent reflections over smooth land for (1) [9], [11], [13], the surface reflectivity Γ can be readily derived from CyGNSS BRCS σ, through [22]

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

where R_t and R_r are the distances from the transmitter and receiver to SP, respectively. As mentioned in Section II-A, these parameters are accessible from the CyGNSS data. It is suggested in [23] that both the surface reflectivity and the normalized BRCS (NBRCS) should be used for land GNSS-R applications considering the duality of possible physical reflection/scattering mechanisms. Here, the peak value of BRCS is taken to represent the CyGNSS NBRCS (and for Γ) as done in [23]. In addition to the CyGNSS-derived Γ, BRCS, and coherence flag, their corresponding geolocation (latitude, longitude, and altitude) and associated climate type are also adopted as input in the present work.

B. BRT-Based SM Retrieval Model

The BRTs are deployed here to model the relationship between the devised input and the SM data. It is worth noting that the BRT algorithm has been successfully applied to downscaling SMAP SM products in [24].

Regression trees (RTs) [25] are able to recursively partition the input space, which contains the CyGNSS Γ, σ, and coherence flag along with the ancillary geolocation and climatology data, and consequently map each partition to the desired output, i.e., SM in this study. However, a single RT tends to overfit the data. Bagging (or bootstrap aggregation) [26] is able to improve both the stability and the predictive power of an RT. This technique resamples the original training data to form several new data sets of the same size as the original training set. Each bootstrapped sample is fitted with a new RT, and the averaged output of all generated RTs is the final result.

In summary, the input of the BRT-model consists of the CyGNSS Γ, σ, and coherence flag as well as the associated site locations (latitude, longitude, and altitude) and climate type, whereas the targeted output is SM. Each RT is created from a replicate of the training data set that is generated using the bootstrapping method. Finally, the input is regressed using the trained trees and the averaged output of these is the desired SM. The detailed description of implementing BRT can be found in [24]. The analyses and model development in this study are performed using the regression learner toolbox of MATLAB R2019b software.

IV. EXPERIMENTS

A. Data Collocation Scheme

In this work, the CyGNSS data and ground truth were collocated on an hourly basis. In terms of the spatial match-up, different CyGNSS measurements were averaged where its SP was within a certain distance from a site. However, there is no such standard distance for collocating the field data and CyGNSS results. In this study, to better explore the optimal match-up strategy for CyGNSS-based SM sensing, the performance of the model was tested using different collocation distances (from 1 to 15 km, with a step size of 1 km). The performance statistics in terms of root-mean-square error (RMSE) and correlation coefficient (R) was found to be similar when the collocation distance was above 2 km. Nonetheless, it should be noted that the number of available samples rose with increasing collocation distances. Here, the collocation distance was set to 7 km that produced the lowest RMSE. To mitigate the impact of inland water body, which tends to result in high CyGNSS Γ and errors in SM estimation, we excluded CyGNSS Γ greater than 0.1 [27]. In summary, the same-hour CyGNSS data (Γ, σ, and coherence flag) whose SPs were within 7 km from a certain site were averaged. Such values along with the site location (latitude, longitude, and altitude) and the corresponding climate type were regarded as input, whereas the associated hourly SM value was the target.

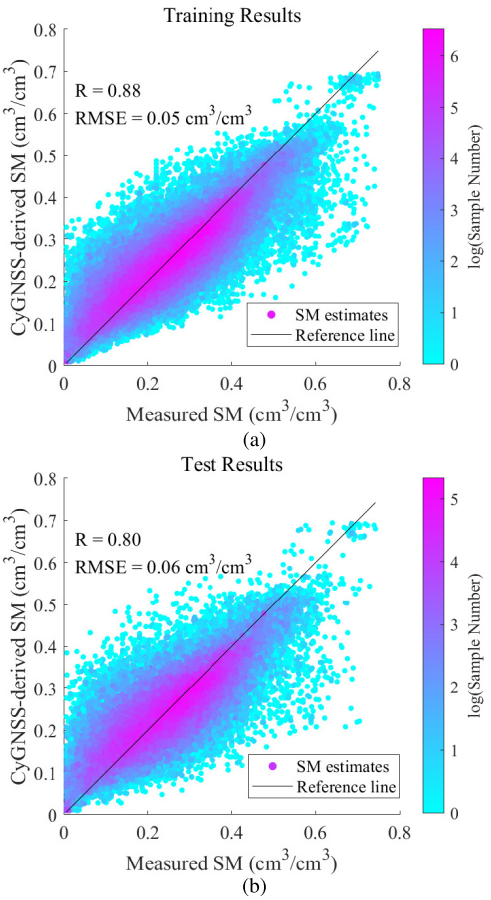

Fig. 2. Density plot showing the alignment between the CyGNSS-derived SM and the ground truth. (a) Training (75% of total data). (b) Test (25% of total data) sets. The colorbar indicates the sample number (in log).

B. Results and Discussion

By using the abovementioned collocation scheme, we obtained 305 529 samples from 1733 different stations for the year 2018. These data were randomly divided into two separate groups that contain 75% and 25% of the total data for the model training and test, respectively. Through validating the prediction produced by the proposed model against the ground truth (see Fig. 2), we obtained an overall R of about 0.86 and an RMSE of 0.05 cm^3/cm^3, and the performances of training and tests are shown in Table I. The clear consistency between the reference and retrieved SM products indicates the effectiveness of the proposed model, and its generalizability is proved by the negligible drop in accuracy for the test set.

TABLE I
PERFORMANCE STATISTICS FOR SM RETRIEVAL

Category	R	RMSE (cm^3/cm^3)			
Training	0.88	0.05	Mean SM	Prec.	Temp.
Test	0.80	0.06	(cm^3/cm^3)	(mm)	(°C)
TCC	0.84	0.05	0.17	23.85	6.86
MPC	0.86	0.06	0.18	35.63	7.95
TeMC	0.84	0.05	0.22	47.65	12.26
SMC	0.83	0.06	0.29	98.70	16.62
TrMC	0.90	0.05	0.27	137.67	24.10
Overall	0.86	0.05			

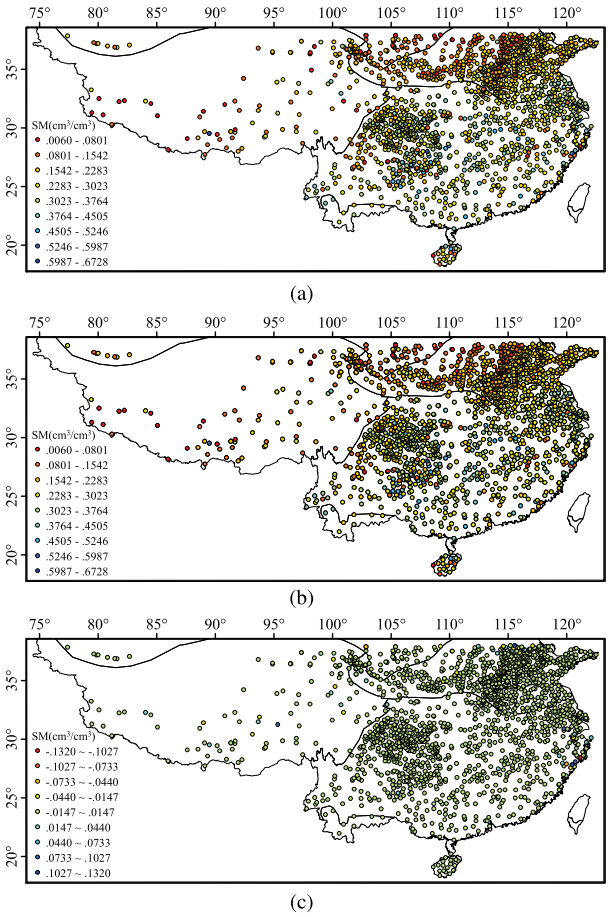

Fig. 3. Annually averaged SM for each site: (a) ground truth, (b) prediction, and (c) their deviation.

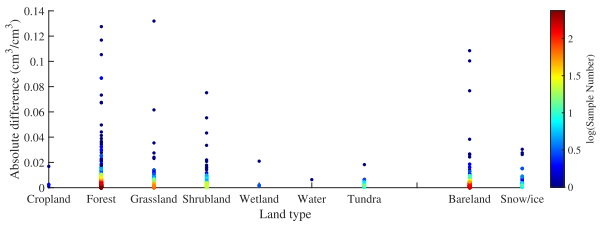

Fig. 4. Deviation between the annual mean reference and estimated SMs in terms of land use.

In addition, we examined the proposed model in terms of spatial variability. The SM results assessed in this work were grouped and averaged according to site locations, which are presented in Fig. 3. It is clear that these mean SM values derived from the ground-truth and retrieval results agree well with each other and their discrepancy (with a mean value of 5×10^{-5} cm^3/cm^3 and a standard deviation of 0.01 cm^3/cm^3) is generally insignificant. Still, a very small amount of high discrepancies occurred (with an absolute difference value above 0.08 cm^3/cm^3) and were mostly resulted from forests (see Fig. 4), by referring to the land use data (from http://data.ess.tsinghua.edu.cn/fromglc10_2017v01.html).

The average number of collocated points per hour is generally low (about one or two), but it can reach an order of tens if daily collocation is used. For the latter case, the impact of the number of averages has been analyzed in [27] and, thus, is not evaluated here.

The impact of climatologic ancillary data is investigated here. It is commonly known that different climate types are characterized by varying temperatures and precipitations that have great impacts on SM. The representative annual mean temperatures and precipitations (based on 30-year averaged data from http://en.weather.com.cn/) for the five examined climate types are presented in Table I, and it is clear that both temperatures and precipitations increase in the sequence of TCC, MPC, TeMC, SMC, and TrMC. The correlation between the average precipitation and SM can be noticed, except that the mean SM of SMC is higher than that of TrMC. This may be due to the higher temperature of TrMC that results in more evaporation of SM than SMC. Furthermore, the accuracy measures for different climate types are all plausible (see Table I), which demonstrates the robustness of the proposed method in terms of different climate types. Through testing various combinations of ancillary data, the minimum amount of ancillary data required for satisfactory retrieval is three. The exclusion of climate type causes a drop of performance, specifically, an increase of 0.006 cm^3/cm^3 in RMSE and a decrease of 0.05 when the geolocations are not provided.

V. CONCLUSION

In the present work, we developed a BRT-based model for retrieving SM from the CyGNSS data with auxiliary geolocation and climatology information. The retrieved SM results were assessed with field data from China's automatic SM observation stations during the year 2018. Different data collocation strategies were performed, and the one with the best accuracy was selected. Satisfactory agreement between the prediction and ground truth showed the efficiency of this proposed model that was demonstrated by a correlation coefficient of 0.86 and an RMSE of 0.05 cm^3/cm^3. The hourly synchronized validation indicated the potential of CyGNSS-based near real-time SM monitoring.

In the future, this model will be tested with more *in situ* data from stations over the globe. In addition, the resulting products can be compared with other satellite-based SM results (such as SMAP data) and employed to improve their spatial and temporal resolutions.

ACKNOWLEDGMENT

The authors would like to thank the Cyclone Global Navigation Satellite System (CyGNSS) Team for making the data available at https://www.esrl.noaa.gov/psd/.

REFERENCES

[1] S. I. Seneviratne *et al.*, "Investigating soil moisture-climate interactions in a changing climate: A review," *Earth-Sci. Rev.*, vol. 99, nos. 3–4, pp. 125–161, May 2010.

[2] G. Foti *et al.*, "Spaceborne GNSS reflectometry for ocean winds: First results from the UK TechDemoSat-1 mission," *Geophys. Res. Lett.*, vol. 42, no. 13, pp. 5435–5441, Jul. 2015.

[3] E. Cardellach *et al.*, "Consolidating the precision of interferometric GNSS-R ocean altimetry using airborne experimental data," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4992–5004, Aug. 2014.

[4] Q. Yan and W. Huang, "Spaceborne GNSS-R sea ice detection using delay-Doppler maps: First results from the U.K. TechDemoSat-1 mission," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 10, pp. 4795–4801, Oct. 2016.

[5] N. Rodriguez-Alvarez, B. Holt, S. Jaruwatanadilok, E. Podest, and K. C. Cavanaugh, "An arctic sea ice multi-step classification based on GNSS-R data from the TDS-1 mission," *Remote Sens. Environ.*, vol. 230, Sep. 2019, Art. no. 111202.

[6] Q. Yan and W. Huang, "Detecting sea ice from TechDemoSat-1 data using support vector machines with feature selection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 5, pp. 1409–1416, May 2019.

[7] Q. Yan and W. Huang, "Tsunami detection and parameter estimation from GNSS-R delay-Doppler map," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 10, pp. 4650–4659, Oct. 2016.

[8] 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, p. 1053, May 2019.

[9] 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.

[10] H. Kim and V. Lakshmi, "Use of cyclone global navigation satellite system (CyGNSS) observations for estimation of soil moisture," *Geophys. Res. Lett.*, vol. 45, no. 16, pp. 8272–8282, Aug. 2018.

[11] 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.

[12] 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.

[13] 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, p. 2272, Sep. 2019.

[14] 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, no. 7, p. 1168, Apr. 2020.

[15] F. Li, X. Peng, X. Chen, M. Liu, and L. Xu, "Analysis of key issues on GNSS-R soil moisture retrieval based on different antenna patterns," *Sensors*, vol. 18, no. 8, p. 2498, Aug. 2018.

[16] National Meteorological Administration, *Climatological Atlas for the People's Republic of China*. Beijing, China: China Meteorological Press, 2002.

[17] M. Dobson, F. Ulaby, M. Hallikainen, and M. El-rayes, "Microwave dielectric behavior of wet soil-part II: Dielectric mixing models," *IEEE Trans. Geosci. Remote Sens.*, vol. GE-23, no. 1, pp. 35–46, Jan. 1985.

[18] B. J. Choudhury, T. J. Schmugge, A. Chang, and R. W. Newton, "Effect of surface roughness on the microwave emission from soils," *J. Geophys. Res.*, vol. 89, no. C9, pp. 5699–5706, 1979.

[19] H. Carreno-Luengo, G. Luzi, and M. Crosetto, "Above-ground biomass retrieval over tropical forests: A novel GNSS-R approach with CyGNSS," *Remote Sens.*, vol. 12, no. 9, p. 1368, Apr. 2020.

[20] M. M. Al-Khaldi, J. T. Johnson, S. Gleason, E. Loria, A. J. O'Brien, and Y. Yi, "An algorithm for detecting coherence in cyclone global navigation satellite system mission Level-1 delay-Doppler maps," *IEEE Trans. Geosci. Remote Sens.*, early access, Aug. 11, 2020, doi: 10.1109/TGRS.2020.3009784.

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

[22] H. Carreno-Luengo, A. Camps, J. Querol, and G. Forte, "First results of a GNSS-R experiment from a stratospheric balloon over boreal forests," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2652–2663, May 2016.

[23] S. Gleason, A. O'Brien, A. Russel, M. M. Al-Khaldi, and J. T. Johnson, "Geolocation, calibration and surface resolution of CYGNSS GNSS-R land observations," *Remote Sens.*, vol. 12, no. 8, p. 1317, Apr. 2020.

[24] S. Chakrabarti, J. Judge, T. Bongiovanni, A. Rangarajan, and S. Ranka, "Spatial scaling using temporal correlations and ensemble learning to obtain high-resolution soil moisture," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1238–1250, Mar. 2018.

[25] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Evanston, IL, USA: Routledge, Oct. 1984.

[26] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[27] 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, p. 111947, Sep. 2020.