




Soil Moisture Retrieval Using BuFeng-1 A/B Based on Land Surface Clustering Algorithm

Zhizhou Guo , Baojian Liu , Wei Wan , Feng Lu, Xinliang Niu, Rui Ji, Cheng Jing, Weiqiang Li ,
Xiuwan Chen, Jun Yang, and Zhaoguang Bai

Abstract— A new land surface clustering algorithm is developed to retrieve soil moisture (SM) using the Global Navigation Satellite System reflectometry (GNSS-R) technique. Data from the BuFeng-1 (BF-1) twin satellites A/B, a pilot mission for the Chinese GNSS-R constellation, is used for SM retrieval. The core concept of the algorithm is to cluster global land areas into different types according to the land properties and calculate the SM type by type, based on the linear relationship between equivalent specular reflectivity and SM. The global comparison between the results and SM product from the Soil Moisture Active Passive mission shows the correlation coefficient (R) is 0.82, and unbiased root mean square error (ubRMSE) is $0.070 \text{ cm}^3 \cdot \text{cm}^{-3}$. The results also show good agreement compared with *in situ* SM measurements with the mean ubRMSE of $0.036 \text{ cm}^3 \cdot \text{cm}^{-3}$. This study proves that the global SM can be retrieved successfully from the BF-1 mission with the land surface clustering algorithm. By taking full advantage of the similarity of land surface physical properties in different regions, the algorithm provides a practical approach for global SM retrieval using spaceborne GNSS-R data.

Index Terms—BuFeng-1 (BF-1), global navigation satellite system reflectometry (GNSS-R), land surface clustering, soil moisture (SM).

INTRODUCTION

SOIL moisture (SM) is of great significance in surface evapotranspiration [1], terrestrial water migration [2], and the

Manuscript received October 31, 2021; revised April 20, 2022; accepted May 18, 2022. Date of publication May 31, 2022; date of current version June 17, 2022. This work was supported in part by the National Natural Science Foundation of China (NSFC) Project under Grant 41971377, in part by the BUFENG-1 Application Extension Program of the China Spacesat Company, Ltd., in part by the ESA-MOST China Dragon5 Programme (ID.58070), and in part by the Natural Science Foundation of Fujian Province under Grant 2019J01853. (Zhizhou Guo and Baojian Liu contributed equally to this work.) (Corresponding authors: Wei Wan; Feng Lu.)

Zhizhou Guo, Baojian Liu, Wei Wan, Rui Ji, and Xiuwan Chen are with the Institute of Remote Sensing and GIS, School of Earth and Space Sciences, Peking University, Beijing 100871, China (e-mail: zzguo@pku.edu.cn; liubaojian@pku.edu.cn; w.wan@pku.edu.cn; toplane@pku.edu.cn; xwchen@pku.edu.cn).

Feng Lu is with the Key Laboratory of Radiometric Calibration and Validation for Environmental Satellites, National Satellite Meteorological Center (National Center for Space Weather), China Meteorological Administration, Beijing 100081, China, and also with the Innovation Center for FengYun Meteorological Satellite (FYSIC), Beijing 100081, China (e-mail: lufeng@cma.gov.cn).

Xinliang Niu, Cheng Jing, and Jun Yang are with the China Academy of Space Technology Xi'an Branch, CAST-XIAN, Xi'an 710100, China (e-mail: xlniu1983@hotmail.com; jingcheng@radi.ac.cn; junyang@cma.gov.cn).

Weiqiang Li is with the Earth Observation Research Group, Institute of Space Sciences (ICE, CSIC), 08193 Barcelona, Spain (e-mail: weiqiang@ice.csic.es).

Zhaoguang Bai is with the DFH Satellite Company Ltd., Beijing 100094, China (e-mail: 13910027870@139.com).

Digital Object Identifier 10.1109/JSTARS.2022.3179325

carbon cycle [3], [4]. In practical applications, SM data are used in numerical weather prediction [5], agriculture monitoring [6], and stream-flow forecasting [7]. At present, optical [8] and microwave remote sensing sensors [9] are used for large-scale SM retrieval. However, optical remote sensing has a high spatial resolution but with a relatively long revisit time, and it is easily affected by clouds and mist. Passive microwave remote sensing at L -band ($\sim 1.4 \text{ GHz}$) has become the primary approach of SM observation worldwide due to its all-weather and all-time capability. For example, the Soil Moisture Active Passive (SMAP) can provide global SM data products with a minimum interval of 2/3 days [10].

Global Navigation Satellite System-Reflectometry (GNSS-R) is a new remote sensing technique to observe the earth's surface by receiving the forward scattering signal of the GNSS satellites. Spaceborne GNSS-R utilizes payloads on board of low-orbit satellite to obtain the surface scattered signal broadcasted by the GNSS. GNSS-R is a cost-effective technique because it takes advantage of the GNSS as the transmitters. Moreover, because the signal broadcasted by the GNSS satellite is L -band, it preserves the same advantages that microwave sensors, such as SMAP and Soil Moisture and Ocean Salinity have. Spaceborne GNSS-R missions, e.g., the Technology Demonstration Satellite-1 of the U.K., the Cyclone Global Navigation Satellite System (CyGNSS) of NASA, and the BuFeng-1 (BF-1) twin satellites A/B of China, have been used to observe not only the ocean surface [11]–[16] but also the land surface [17]–[22]. SM retrieval is one of the most popular and practical land surface applications [23].

Various SM retrieval methods using spaceborne GNSS-R data have been proposed in previous studies. Most of the algorithms were established on the linear relationships between the GNSS-R equivalent specular reflectivity (ESR) and other land surface SM products. For example, some studies directly linear fit the calculated ESR with SMAP SM in each surface grid [17], [24]. Some focused on quantifying the effect of land properties, e.g., vegetation and surface roughness in [25] and [26]. There have also been some studies that retrieved SM by machine learning [27]–[29]. However, the methods mentioned above have the following limitations: 1) the linear relationships were usually obtained in each dispersed surface grid or subgrid, and this solution requires long-time and high-density observations to build reliable relationships; and 2) In order to decrease the quantified effect on the ESR calculation, some frequently-changing

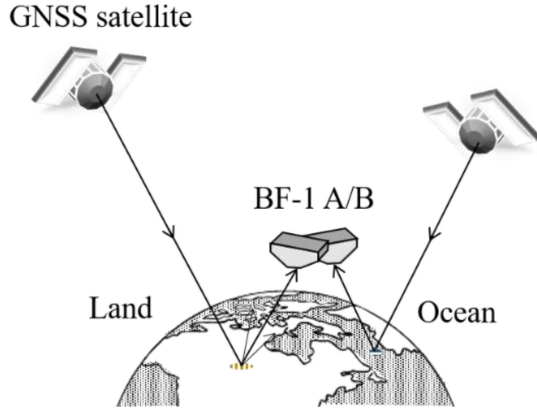


Fig. 1. Functional diagram of BF-1. GNSS-R receiver receives ocean/land scattered signals of GNSS (GPS/BDS) satellite, and the observation position is determined by locations of GNSS satellite and GNSS-R receivers.

reference data, e.g., vegetation opacity (VO), have to be constantly updated grid-by-grid in real-time SM retrieval.

To benefit from the high accuracy of the aforementioned linear fitting algorithm and the physical meaning of the land-property-based algorithm, we propose a novel land surface clustering method to retrieve SM. Based on the consensus that SM retrieval is highly affected by vegetation and surface roughness, the proposed algorithm uses critical parameters [i.e., the roughness coefficient (RC) and VO] for global land surface clustering. This solution provides a new way for SM retrieval by taking full advantage of the similarity of land surface physical properties in different regions. The algorithm is initially applied to five-month BF-1 data (i.e., From July 1 to November 31, 2019) to verify its effectiveness. Three specific regions (i.e., southern Australia, eastern United States, and India) and four *in situ* sites (i.e., Ithaca-13-E, Whitman-5-ENE, Geneva #1, and Tucson-11-W) are also used to further evaluate the detailed performance of the algorithm. Both the global and regional comparisons prove that the SM can be retrieved successfully from the BF-1 mission with this new algorithm.

II. BACKGROUND

A. BF-1 Mission

The BF-1 A/B twin satellites, launched in June 2019, constitute the first Chinese GNSS-R constellation [30]. The primary payloads were developed by the China Academy of Space Technology Xi'an Branch. Each platform has two nadir GNSS-R antennas and a GNSS-R receiver. BF-1 operates at the height of 579 km with an orbital inclination of 45° . The data coverage ranges from 53°S to 53°N . BF-1 receives signals of both GPS L1C/A and BeiDou B1I bands (The BeiDou data are not yet available for usage). As shown in Fig. 1, GNSS satellites constantly emit *L*-band signals under normal working conditions, and BF-1 observes these signals through the down-looking antennas after reflectivity on the earth's surface.

Delay-Doppler map (DDM) [31] is the main product to record GNSS reflected signals in BF-1 L1B data. BF-1's DDM is computed by 1-ms coherent integration and 1-s incoherent

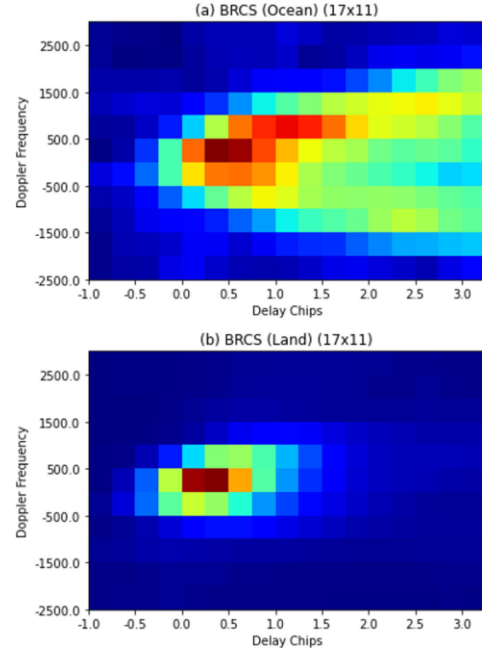


Fig. 2. BF-1 DDM over: (a) ocean; (b) land.

integration, and it contains 17×11 BRCS in the L1B product [30]. Fig. 2(a) and (b) shows the instances of the DDMs observed over ocean and land, respectively.

B. Principle of SM Retrieval

The main principle of SM retrieval using GNSS-R is to calculate the ESR and retrieve SM with the linear relationship between ESR and SM. The ESR can be calculated with several approaches based on the assumptions of both coherence and incoherence [17], [26], [29]. For now, there are still some disputations on the coherence of GNSS-R signal on land applications. In our research, based on the previous studies, SM measurement is normalized to coherent ESR.

Based on the assumption that the coherent reflection signal dominates the observed GNSS-R signal, the ESR affected by vegetation and roughness [17] can be described as follows:

$$\Gamma = \frac{\sigma(R_{st} + R_{sr})^2}{4\pi(R_{st}R_{sr})^2} \quad (1)$$

where Γ is the ESR; R_{st} is the distance between the specular reflection point and the transmitter; R_{sr} is the distance between the specular reflection point and the receiver; σ is the value of the DDM of the BF-1 product. The Γ is converted to $\Gamma(\text{dB})$ using (2), which is then used for the linear regression to calculate the resulting SM. The method used to determine the linear regression of this study will be described in the following Section II-B

$$\Gamma(\text{dB}) = 10 \log \sigma + 20 \log (R_{st} + R_{sr}) - 20 \log R_{st} - 20 \log R_{sr} - 10 \log 4\pi. \quad (2)$$

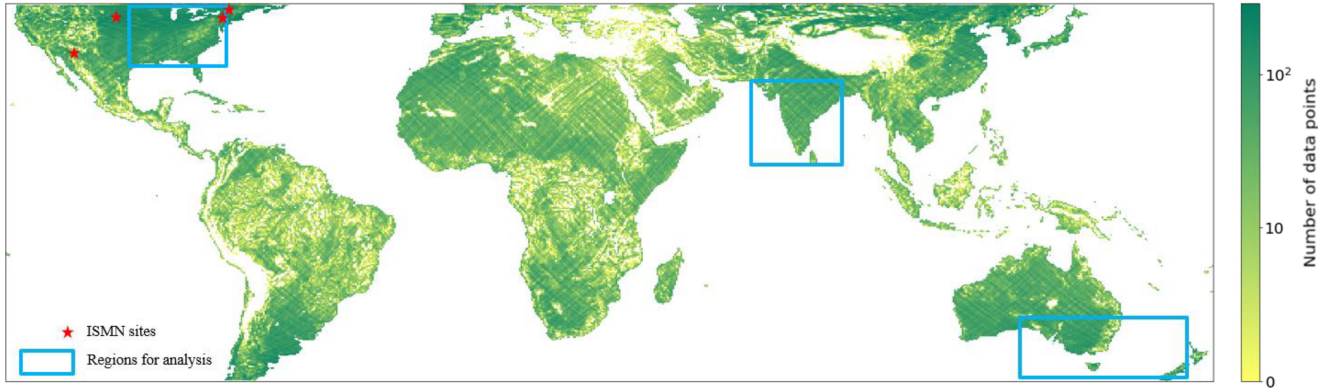


Fig. 3. Spatial distribution of the five-month BF-1 data point used in this study (shown in 36-km EASE-Grid). The ISMN *in situ* sites used for validation are marked as red stars; The three specific regions are used in Section IV-B to analyze the accuracy of the proposed algorithm is shown as the blue rectangles.

II. DATASET AND METHODOLOGY

A. Dataset

In this study, BF-1 L1B sample data from July to November 2019 are used. Similar to the filtering method proposed to process the land data of CyGNSS [17], we remove the BF-1 data with a signal-to-noise ratio (SNR) < 5 dB and the incident angle > 45°. Fig. 3 shows the number of valid BF-1 data points over the globe in 36-km Equal-Area Scalable Earth Grid (EASE-Grid). It is noted that, due to the design of BF-1 orbit, the density of data coverage at high latitudes is higher than that at low latitudes.

The SMAP data are used as reference data, and all the reference data come from SMAP L3 Radiometer Global Daily 36 km EASE-Grid SM, Version 7, which are freely available at <https://nsidc.org/data/SPL3SMP>. The five variables used in this study are listed as follows: [32]:

- 1) SM;
- 2) RC;
- 3) VO;
- 4) landcover class;
- 5) static water body fraction.

RC (from 0 to 3.0) is the same “h” RC for a given polarization channel, and VO (from 0 to 5.0) is the parameter normalized by the cosine of the incidence angle [33]. The land cover class is from ranking among the International Geosphere Biosphere Programme (IGBP) land cover classes using a statistical model. The static water body fraction is computed based on the number of water and land pixels reported on a 250-m grid. With the static water body fraction, the static water bodies in BF-1 land data are removed for retrieval.

The SM data of the International Soil Moisture Network (ISMN) are used as ground truth to verify the SM retrieval results. Locations of the sites used in this study are shown in Fig. 3, and the properties of these sites will be shown in the following Section IV-C.

B. Methodology

1) *Land Surface Clustering*: Previous studies have proven that the confounding factors of vegetation and surface roughness reduce the sensitivity of the L-band to SM [34]. In most real

situations, the original ESR is a combined value of reflections from the rough surface of soil and vegetation [26], [35]. Two expressions, i.e., $\exp[-\tau \sec(\theta)]$ and $\exp[-h \cos^2(\theta)]$, were commonly used to correct the effects of vegetation and surface roughness, respectively. Where τ is the VO, h is the surface roughness parameter, and θ is the incidence angle [36]. h can be expressed as $h = (\frac{2\pi}{\lambda}) 4\sigma^2$, which depends on the GNSS signal wavelength λ and the surface height standard deviation (STD) σ [37].

This study assumes that the vegetation and surface roughness effects for the same land type can be considered similar enough to be neglected. Based on this, we propose a land surface clustering algorithm using critical vegetation and surface roughness parameters. Averaged RC and VO for five months (From July 1 to November 31, 2019) are used as the key features in the land surface clustering for each pixel. The heatmap of two parameters is shown in Fig. 4(a). Except for the ocean pixels, there are several high data density areas in this figure. The *K-Means++* algorithm [38] is employed in the clustering. The clustering results are iteratively derived according to the number of clustering types inputted to the algorithm, and each pixel will be imparted a type ID according to its key features. For a specific type, the ratio of type area versus total global area is defined as a proxy of the area-percentage of this type. To compare the clustering results of different type numbers, Fig. 4(c)–Fig. 4(g) shows the top 17 large-area-percentage types out of total clustering numbers (20, 50, 100, 200, and 500) for readability. The IGBP types are also shown as a benchmark in Fig. 4(b) for comparison. These subfigures visually show the feasibility of the *K-Means++* algorithm.

To quantitatively determine a reasonable number of types, the (SSE) and average contour coefficient (ACC) are calculated to represent the quality of clustering (SSE closed to 0 and ACC closed to 1 are considered with high quality). Because of our program’s random selection of the cluster center, clustering results with a bit of difference, the clustering predictions are made 30 times, and the average ACC and SSE are finally used. As shown in Fig. 5, with the increase of the number of types, SSE decreases monotonously as expected. On the other hand, ACC fluctuates irregularly, which comes to the peak value when the number of types is 330. Therefore, to avoid the senseless types

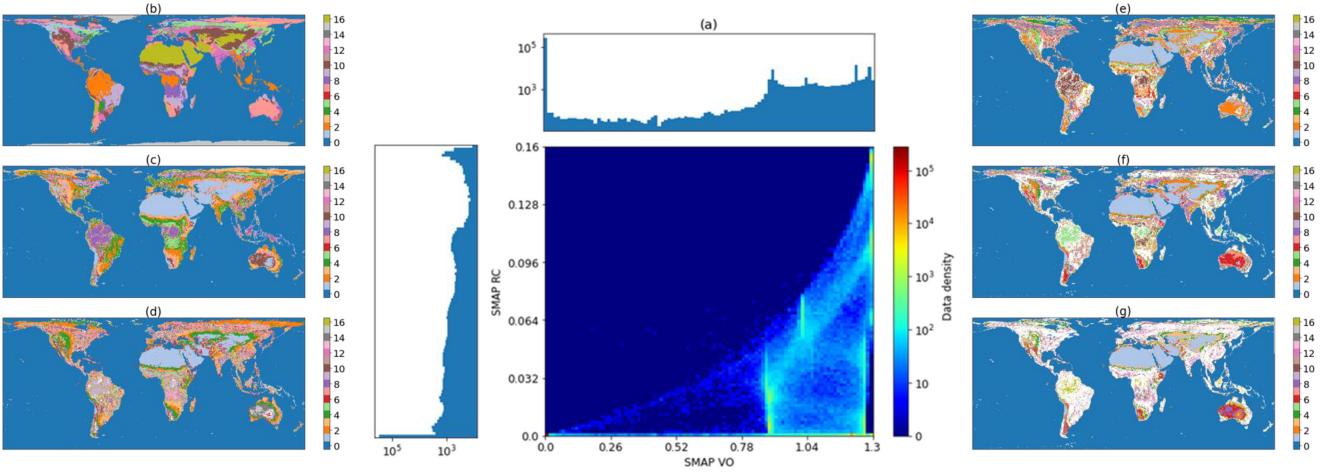


Fig. 4. Land clustering results for different type numbers. (a) Distribution of VO and RC (the clustering basis). (b) IGBP land type. (c) Top 17 types of area ratio in 20-types-clustering. (d) Top 17 types of area ratio in 50-types-clustering. (e) Top 17 types of area ratio in 100-types-clustering. (f) top 17 types of area ratio in 200-types-clustering. (g) Top 17 types of area ratio in 500-types-clustering.

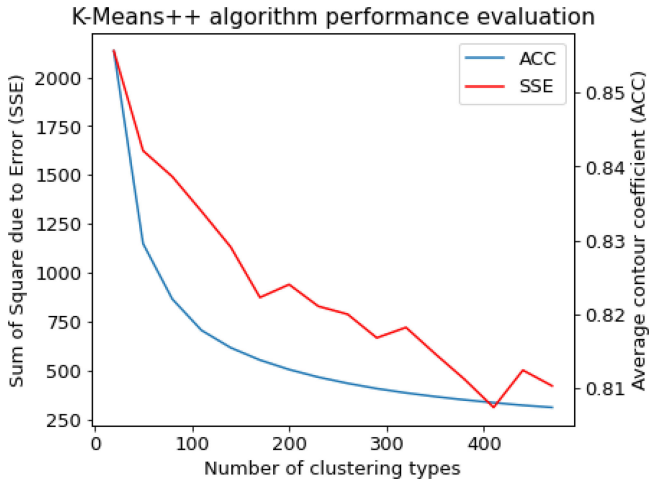


Fig. 5. K-Means++ clustering quality judged by the sum of square due to error (SSE) and ACC.

with insufficient sample size and ensure a low SSE value and a high ACC value for the clustering, the number of clustering types in this study is determined to be 200.

We use the IGBP landcover class to merge the *K*-means clustering results for a more straightforward analysis and display. When merging categories, the proportion of IGBP pixels in a particular clustering type in the total number of such pixels (P_{IGBP}) is taken as a reference, which is calculated as follows:

$$P_{IGBP} = N_{IGBP,i}/N_i \quad (3)$$

where $N_{IGBP,i}$ is the number of IGBP pixels in the clustering type i . N_i is the pixel number of the clustering type i .

For the types whose P_{IGBP} are greater than 70%, the clustering types will be linked to the IGBP types. The link will be used for display only and will not affect the retrieval; otherwise, the type is considered a new type without a specific IGBP type. The final result of the combination of clustering results in this experiment is shown in Fig. 6(a). To compare the result of clustering types

and IGBP types, we use the IGBP types as the reference target land types for the combination of the clustering types, and Fig. 6(b) shows the basis of this combination. The proportion of the pixel number of IGBP to clustering types is used to get clustering-IGBP types with the combination line of 70%. It should be noted that the surface clustering results obtained in this experiment are quite different from those of IGBP land type products. The clustered land types only consider the similarity of vegetation and surface roughness, which directly contributes to the SM retrieval algorithm. The effects of the two parameters on SM are depicted in Fig. 6(c) using SMAP data. Note that SM is very sensitive to vegetation and surface roughness, which further reflects the effectiveness of the clustering method proposed in this study.

2) *SM Retrieval With Linear Regression*: Based on the linear relationship between GNSS-R reflectivity and SM, the algorithm use ESR (Γ (dB)) of BF-1 calculated by (2) in this section. For each clustering type, we assume that vegetation and roughness have a similar effect on SM retrieval, which means the impact is a constant in the calculation. Thus, we use Γ (dB) only in the linear fitting. We calculate the slope and intercept of the best-fit linear least-squares regression between SM_{SMAP} and Γ in each clustering type.

As shown in Fig. 7, the BF-1 ESR and SMAP SM are correlated for each type, and a linear fitting equation is determined. Note that for each correlated match-up (type), there is one specific slope derived from the linear regression.

And the final SM_{BF-1} calculation can be described as the following equation:

$$S M_{BF-1} = k_i \cdot \Gamma_i + b_i \quad (4)$$

where i is the type ID of clustering type. k_i , b_i are the slope and intercept. Γ_i is the BF-1 ESR.

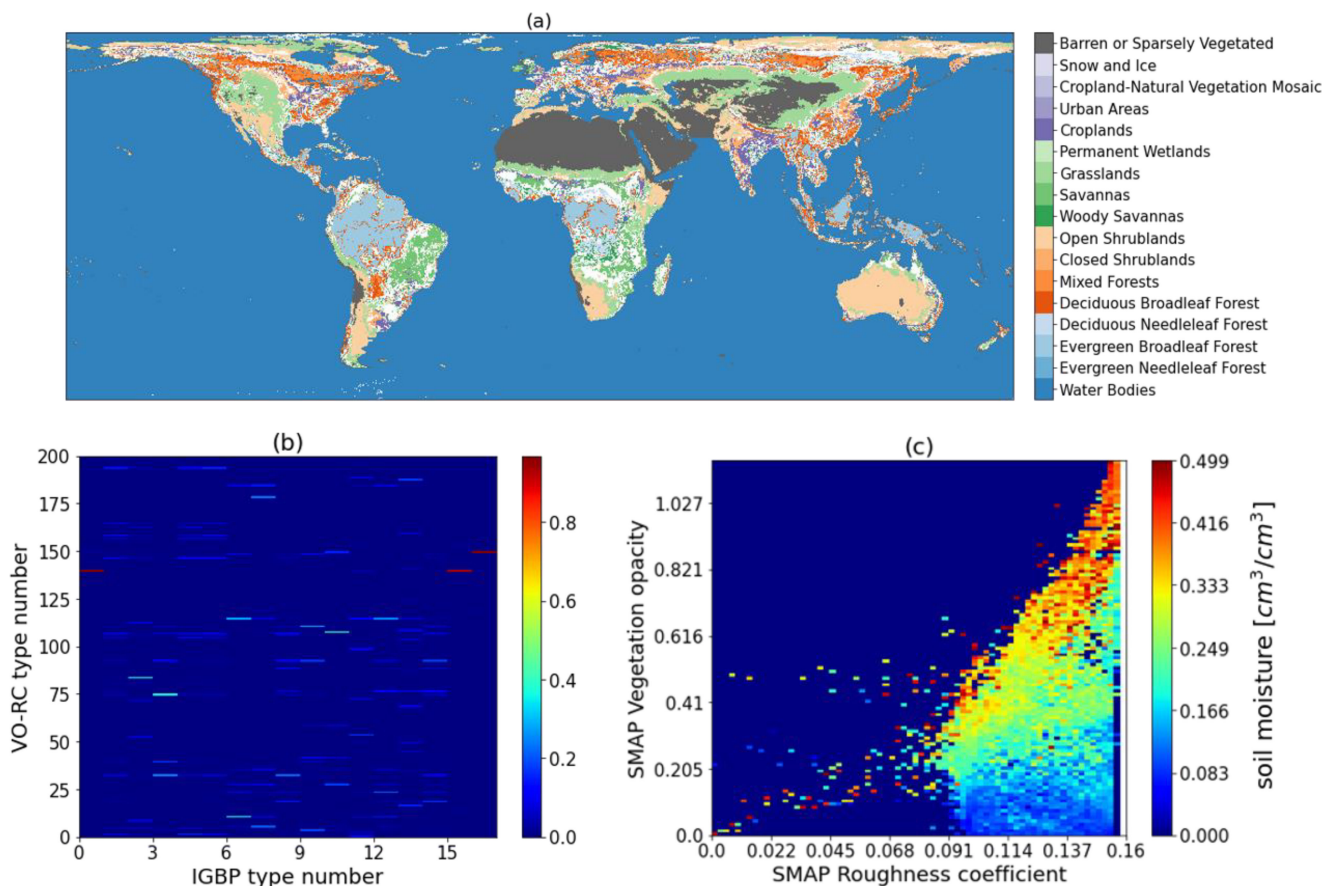


Fig. 6. Clustering results. (a) Distributions of the 17 clustering-IGBP clustering types. (b) Proportion of IGBP types pixels within clustering types pixels (ranging between 0 and 1). (c) Average SM in different VO and RC ranges obtained from SMAP SM data.

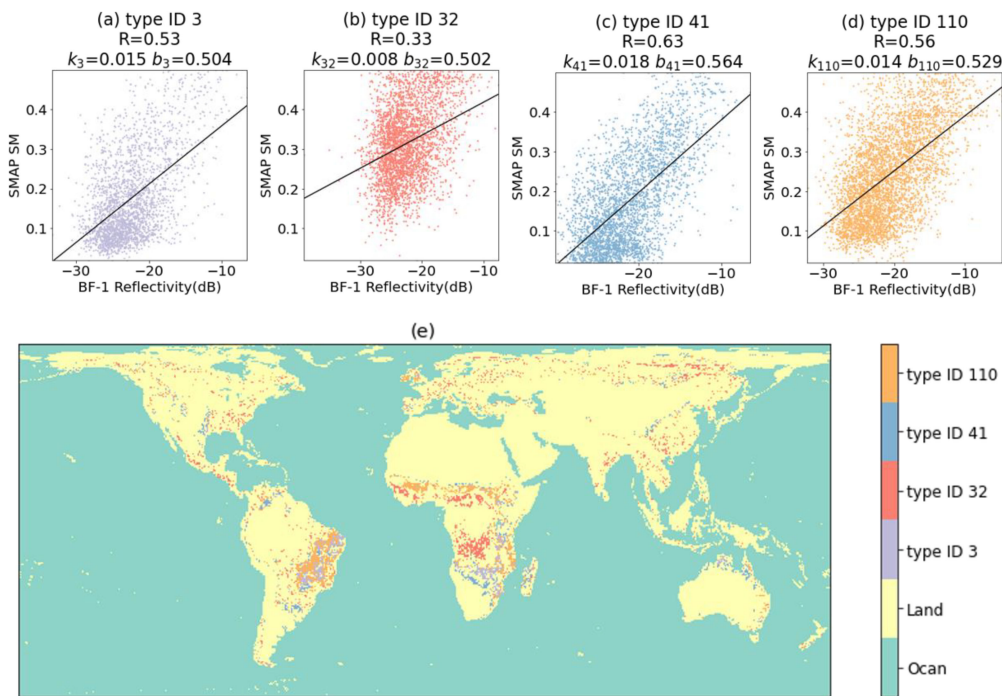


Fig. 7. Examples of the linear fitting of BF-1 ESR and SMAP SM for different clustering types. (a) Type ID = 3. (b) Type ID = 32. (c) Type ID = 41. (d) Type ID = 110. (e) Map of the types in (a)-(d).

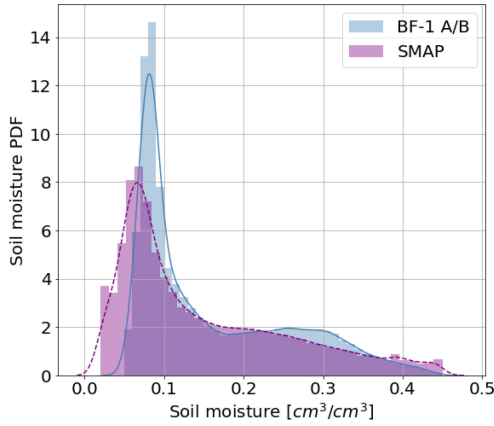


Fig. 8. PDFs of BF-1 SM and SMAP 36 km SM.

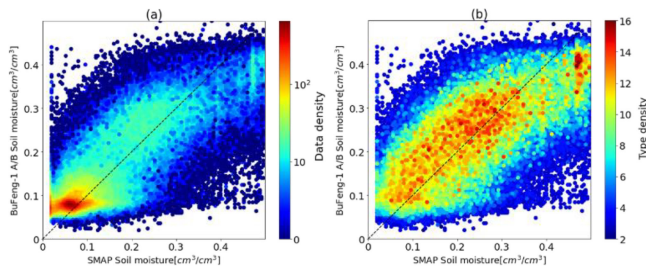


Fig. 9. Correlation of BF-1 SM and SMAP SM colored by (a) data density; (b) type density.

III. RESULT AND DISCUSSION

A. Overview of BF-1 Retrieved Global SM

The probability density functions (PDFs) of the BF-1 SM and SMAP SM are shown in Fig. 8. Both PDFs show that more than 40% of the SM data has a low SM value with 42.18% BF-1 SM and 45.30% SMAP SM less than $0.1 \text{ cm}^3 \cdot \text{cm}^{-3}$. The distribution range of BF-1 SM is less than that of SMAP SM, and the difference mainly distributes between $0\text{--}0.1 \text{ cm}^3 \cdot \text{cm}^{-3}$ and $0.2\text{--}0.4 \text{ cm}^3 \cdot \text{cm}^{-3}$.

Fig. 9 shows the correlation of the BF-1 SM and SMAP SM, where the correlation coefficient (R) is 0.82, and unbiased root mean square error (ubRMSE) is $0.070 \text{ cm}^3 \cdot \text{cm}^{-3}$. Fig. 9(a) and (b), respectively, shows the data density of SM and the type density of SM. Similar to data density, the type density is defined as the number of different clustering types around each data point, and it is calculated by statisticizing the number of different clustering types in each grid divided by SMAP SM and BF-1 SM. It clearly shows that most data are centrally distributed in low SM ($< 0.1 \text{ cm}^3 \cdot \text{cm}^{-3}$) in Fig. 9(a) with a low-type density in Fig. 9(b), which means these low SM data come from various types. Besides, most type variations appear in SMs range from $0.05 \text{ cm}^3 \cdot \text{cm}^{-3}$ to $0.35 \text{ cm}^3 \cdot \text{cm}^{-3}$ and distribute along the $y = x$ line. It proves the effectiveness of the land surface clustering method to solve the problem of the uneven data distribution, and SM can be retrieved by merging the data in similar pixels.

Fig. 10(a) and (b) shows the spatial distribution of BF-1 SM and SMAP SM. It shows a good similarity of distribution trend

of the two SM data. Fig. 10(c) and (d) shows the R and ubRMSE between BF-1 SM and SMAP SM with obvious regional distribution characteristics. For low SM regions ($< 0.1 \text{ cm}^3 \cdot \text{cm}^{-3}$), the BF-1 SM shows good ubRMSE ($0.042 \text{ cm}^3 \cdot \text{cm}^{-3}$) but poor R (0.15) with SMAP SM, such as the Sahara Desert, the middle east and north of China, and the Russian border region (most of these regions belongs to IGBP 16). However, some sparse shrub regions with similar low SM, such as central Australia, Southern Africa, and Argentina (most of these regions belongs to IGBP 7), have both good ubRMSE ($0.053 \text{ cm}^3 \cdot \text{cm}^{-3}$) and good R (0.56) with SMAP SM. For high SM regions ($> 0.4 \text{ cm}^3 \cdot \text{cm}^{-3}$) (most of which is tropical rainforest), such as the Amazon, Indonesian rainforest, and Congo rainforest, both BF-1 SM and SMAP SM are considered to be unreliable due to dense vegetation cover [6].

To sum up, the proposed algorithm performs differently in different regions on a global scale. First, the algorithm performs well in low SM areas of subtropical for the clustering of flat and sparsely vegetated deserts and open shrublands. However, in some highly covered moist areas in tropic and subtropic, the error of the land surface clustering algorithm gets larger significantly. In some dense vegetation pixels and their neighboring pixels under the influence, e.g., rainforests, thick grasslands, and croplands, the retrieved SM is considered unreliable. Second, for the significant impact of the open water on the ESR [17], the surface water bodies affect the SM retrieval under the observation scale of this study, particularly in the areas with a high water body fraction, e.g., Southern China, and the southwestern coast of South America. Third, the BF-1 data at very high altitude areas, e.g., the Rocky Mountains, the Andes Mountains, and the Tibetan Plateau, are not used in the algorithm because of the low SNR of those data.

B. BF-1 SM in Local Regions

In this section, three regions, i.e., southern Australia, eastern United States, and India, are chosen to show the detailed information of the BF-1 retrieved SM in Fig. 11.

BF-1 SM retrieval in Eastern United States [see Fig. 11(a1)–Fig. 11(a5)] has the best agreement with SMAP among the three regions, with $R = 0.75$ and $\text{ubRMSE} = 0.066 \text{ cm}^3 \cdot \text{cm}^{-3}$. SMs in this region mainly distribute between 0.05 and $0.16 \text{ cm}^3 \cdot \text{cm}^{-3}$, and quite a few SMs distribute between 0.16 and $0.35 \text{ cm}^3 \cdot \text{cm}^{-3}$. As shown in Fig. 11(a2) and (a3), the variations of SM from the western mountainous area to the eastern coast are clearly depicted.

As the high-latitude region with more BF-1 observations, the SM retrieval in southern Australia is shown in Fig. 11(b1)–Fig. 11(b5). As shown in the 2-D histogram in Fig. 11(b5), the majority of this region is with low SM ($0.05\text{--}0.08 \text{ cm}^3 \cdot \text{cm}^{-3}$), and the residential areas around the southeast coast can be recognized by higher SM values [see Fig. 11(b1) and (b2)]. Thus, the retrieval result in this region is good, with R of 0.70 and ubRMSE of $0.083 \text{ cm}^3 \cdot \text{cm}^{-3}$.

BF-1 SM in Indian Fig. 11(c1)–Fig. 11(c5) has the lowest consistency compared with the eastern United States and the southern Australia, with $R = 0.65$ and $\text{ubRMSE} = 0.091$

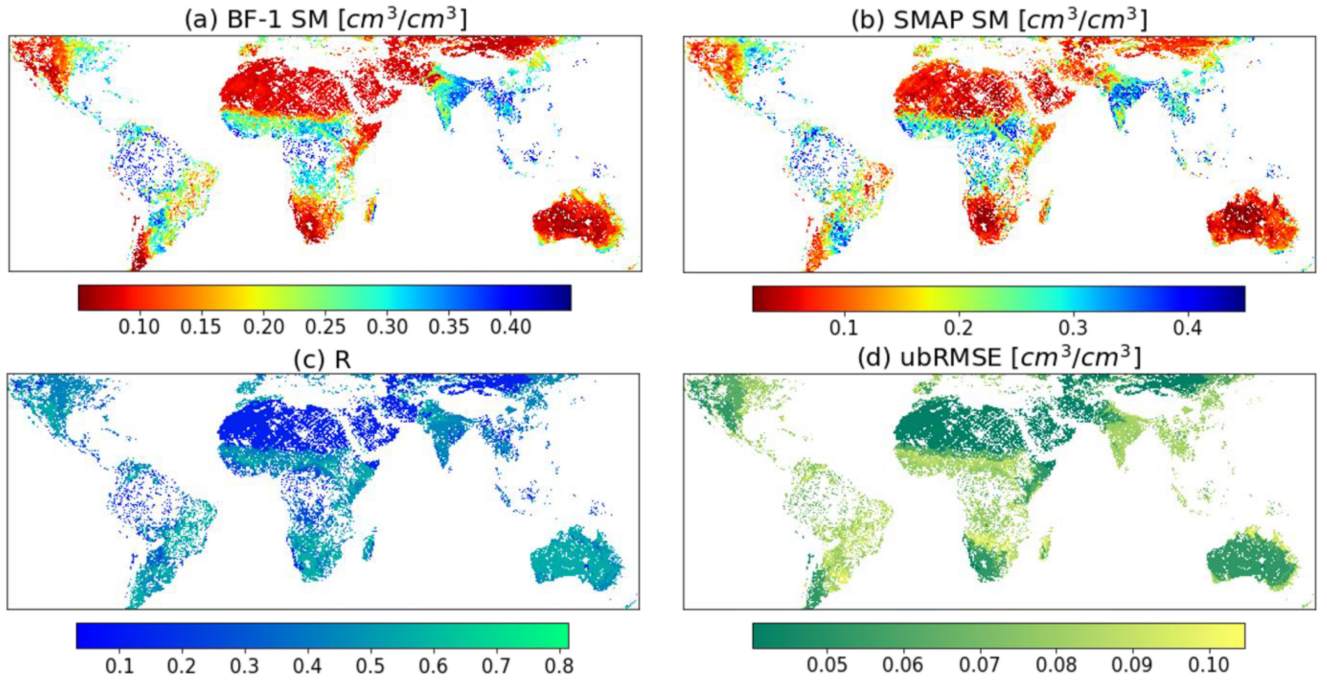


Fig. 10. Comparison of the BF-1 SM and SMAP SM. (a) BF-136 km global SM. (b) SMAP 36 km global SM. (c) R value between SMAP SM and BF-1 SM. (d) ubRMSE between SMAP SM and BF-1 SM.

TABLE I
COMPARISON OF THE ACCURACIES OVER THE THREE REGIONS

| Regions | Average Contour Coefficient (ACC) | Average R between SM_{SMAP} and $\text{Ref}_{\text{BF-1}}$ (R_{ESR}) | Average R between SM_{SMAP} and $\text{SM}_{\text{BF-1}}$ | Average ubRMSE between SM_{SMAP} and $\text{SM}_{\text{BF-1}}$ ($\text{cm}^3\text{cm}^{-3}$) |
|-----------------------|-----------------------------------|---|---|--|
| Southern Australia | 0.28 | 0.46 | 0.70 | 0.083 |
| Eastern United States | 0.35 | 0.45 | 0.75 | 0.066 |
| India | 0.41 | 0.35 | 0.65 | 0.091 |

$\text{cm}^3\cdot\text{cm}^{-3}$. As shown in Fig. 11(c2) and (c3), in the middle of this region, BF-1 underestimates quite a lot of high SMs, which is the largest error in the region. It can be seen in Fig. 11(c5) that SMs in this region mainly distribute between 0.39 and 0.5 $\text{cm}^3\cdot\text{cm}^{-3}$, and the error between SMAP SM and BF-1 SM is larger than errors in the other two regions.

A comparison of the accuracies over the three regions is shown in Table I. It is interesting to note that, based on the similar average R between SM_{SMAP} and $\text{ESR}_{\text{BF-1}}$, the accuracy of SM increases with the increase of ACC. However, India has the highest ACC (0.41) but the lowest R_{ESR} , with the lowest accuracy.

C. Validation Using the In Situ Data

Four ISMN sites, i.e., Ithaca-13-E, Whitman-5-ENE, Geneva#1, and Tucson-11-W are chosen to validate the accuracy of the BF-1 SM. Basic information of the sites is listed in Table II. During the observed period of BF-1, there should be

more than 100 ISMN sites for validation. However, since the BF-1 SM is sparse, there are actually only a few available sites. Even so, for a specific ISMN site, we have to set the search radius to be 10 km to get more than 10 BF/ISMN match-ups for comparison. Therefore, considering the range and significance of SM variations, the four sites in Table II are finally selected for validation.

As shown in Fig. 12(a) and Table II, for the Ithaca-13-E site, the average in situ SM is $0.276 \text{ cm}^3\cdot\text{cm}^{-3}$, and the ubRMSE between BF-1 SM and ISMN SM is $0.044 \text{ cm}^3\cdot\text{cm}^{-3}$. SM in this site changes over a wide range (STD is $0.057 \text{ cm}^3\cdot\text{cm}^{-3}$). In Fig. 12(b), for the Whitman-5-ENE site, the variation range of the *in situ* SM is $0.05\text{--}0.11 \text{ cm}^3\cdot\text{cm}^{-3}$ with the mean of $0.07 \text{ cm}^3\cdot\text{cm}^{-3}$ and STD of $0.013 \text{ cm}^3\cdot\text{cm}^{-3}$. The consistency between BF-1 SM and ISMN SM is good except that within DOY 190–221. The ubRMSE between BF-1 SM and ISMN SM in this site is $0.030 \text{ cm}^3\cdot\text{cm}^{-3}$. From Fig. 12(c), which is for the Geneva#1 site, except for the outlier in DOY 260, the accuracy is good within the narrow range of SM variation (i.e., $0.17\text{--}0.32 \text{ cm}^3\cdot\text{cm}^{-3}$ with the STD of $0.038 \text{ cm}^3\cdot\text{cm}^{-3}$). The ubRMSE between BF-1 SM and ISMN SM is $0.046 \text{ cm}^3\cdot\text{cm}^{-3}$. Fig. 12(d) shows the results of the Tucson-11-W site, where the variation frequency of SM is high, but the range is narrower ($0.02\text{--}0.14 \text{ cm}^3\cdot\text{cm}^{-3}$). There are some low accuracy SM data during the low variation frequency period like DOY 278–319. The ubRMSE between BF-1 SM and ISMN SM is $0.024 \text{ cm}^3\cdot\text{cm}^{-3}$ for this site. Besides, as it can be seen, the sites with lower VO have a better result, e.g., Whitman-5-ENE and Tucson-11-W.

Overall, the mean ubRMSE between the BF-1 SM and ISMN SM is $0.036 \text{ cm}^3\cdot\text{cm}^{-3}$ for the four sites. These results prove that the retrieval accuracy is related to the vegetation parameters (i.e.,

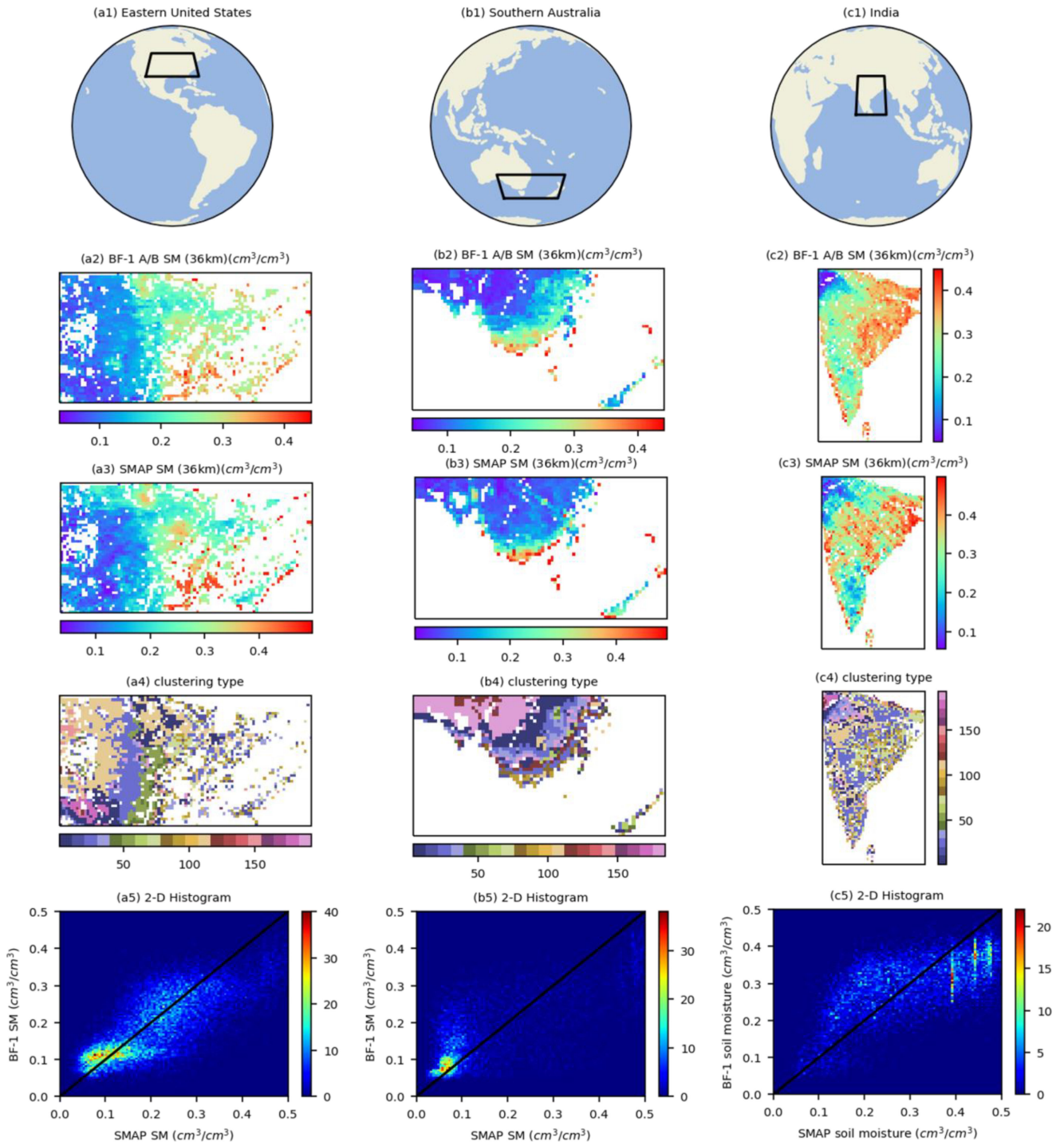


Fig. 11. Retrieval of different regions in the world (a) Southern Australia; (b) Eastern U.S.; (c) India. No.1–5 indicate the 1) Regions extent; 2) BF-1 SM; 3) SMAP SM; 4) clustering type; 5) 2-D histogram.

the VO in this study). It also shows that the BF-1 SM is not that accurate when SM variation is frequent.

D. Comparison With Other Methods

The proposed land surface clustering algorithm is further compared with the two previous algorithms, i.e., the UCAR/CU algorithm in [17] and the reflectivity–vegetation–roughness

(R–V–R) algorithm in [26]. The former is a representative grid-by-grid linear fit algorithm, and the latter is one of the global/regional linear fit algorithms. Five-month BF-1 data after being consistently preprocessed, the same as used in this study, are used as inputs to the two algorithms. The results of the three algorithms are shown in Table III. The UCAR/CU algorithm shows the highest R of 0.86 and the lowest ubRMSE of $0.57 \text{ cm}^3 \cdot \text{cm}^{-3}$, benefiting from the grid-by-grid linear relationships

TABLE II
RETRIEVAL RESULT OF FOUR ISMN SITES

| ID | ISMN site | Latitude (°) | Longitude (°) | clustering Type ID | Top 1 IGBP ID | Top 1 IGBP ratio (%) | IGBP Land Type | VO |
|----|---------------|--------------|---------------|--------------------|---------------|----------------------|-----------------|-------|
| 1 | Ithaca-13-E | 42.440 | -76.246 | 32 | 8 | 47.34 | Woody Savannas | 0.575 |
| 2 | Whitman-5-ENE | 42.068 | -101.445 | 107 | 10 | 88.94 | Grasslands | 0.083 |
| 3 | Geneva#1 | 42.883 | -77.033 | 18 | 8 | 44.12 | Woody Savannas | 0.457 |
| 4 | Tucson-11-W | 32.239 | -111.169 | 5 | 7 | 87.68 | Open Shrublands | 0.070 |

| ID | <i>In situ</i> SM average (cm ³ cm ⁻³) | <i>In situ</i> SM Standard Deviation (cm ³ cm ⁻³) | BF-1 SM average (cm ³ cm ⁻³) | BF-1 SM Standard Deviation (cm ³ cm ⁻³) | ubRMSE (cm ³ cm ⁻³) |
|----|---|--|---|--|--|
| 1 | 0.276 | 0.057 | 0.296 | 0.015 | 0.044 |
| 2 | 0.070 | 0.013 | 0.103 | 0.021 | 0.030 |
| 3 | 0.268 | 0.038 | 0.306 | 0.038 | 0.046 |
| 4 | 0.054 | 0.032 | 0.085 | 0.011 | 0.024 |

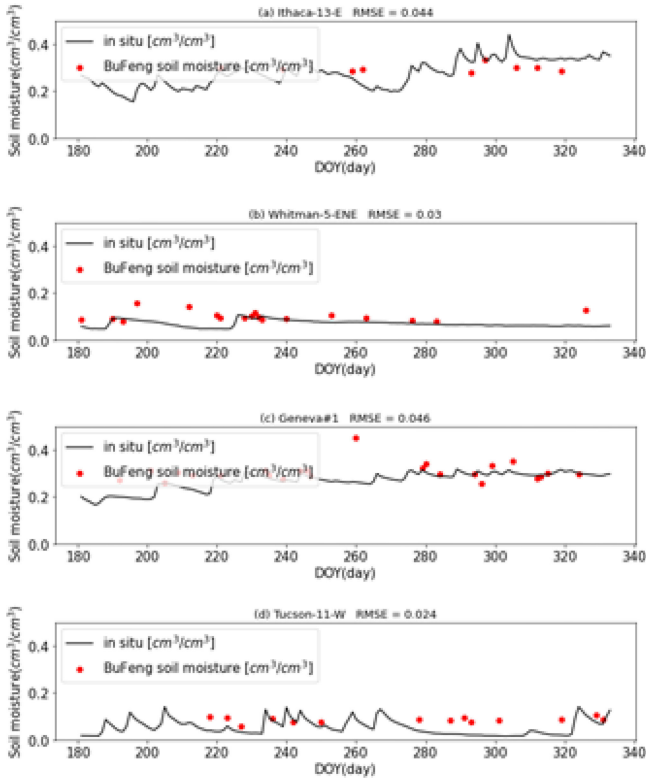


Fig. 12. Four ISMN stations. (a) Ithaca-13-E. (b) Whitman-5-ENE. (c) Geneva#1. (d) Tucson-11-W; SM and BF-1 SM retrieval for five months.

TABLE III
COMPARISON WITH UCAR/CU AND R-V-R METHODS METHODS

| Algorithm | R | ubRMSE (cm ³ cm ⁻³) | available SM area in percentage (%) |
|-------------------------|------|--|-------------------------------------|
| Land surface clustering | 0.82 | 0.070 | 35.63 |
| UCAR/CU | 0.86 | 0.057 | 17.06 |
| R-V-R | 0.69 | 0.090 | 47.38 |

between only ESR and SMAP SM. However, it sacrificed the available SM area (in percentage), with the land surface clustering algorithm being more than twice as much as that of the UCAR/CU algorithm (35.63% versus 17.06%). In the meantime, the R-V-R algorithm shows slightly poorer accuracy than the other two, with $R = 0.68$ and $ubRMSE = 0.089 \text{ cm}^3 \cdot \text{cm}^{-3}$. However, it achieves the highest percentage of 47.38% concerning the available SM area.

IV. CONCLUSION

In this study, a new land surface clustering algorithm is developed to retrieve SM using the GNSS-R technique, and the algorithm is applied to the BF-1 data to verify its effectiveness. The SM results are presented globally and in different regions, and the BF-1 SM is compared with the SMAP SM and *in situ* measurements to analyze the performance of the clustering algorithm and the BF-1 data. From the experiments of this study, BF-1 SM shows good consistency with SMAP SM with $R = 0.82$ and $ubRMSE = 0.070 \text{ cm}^3 \cdot \text{cm}^{-3}$, and BF-1 SM also shows good agreement with the *in situ* measurements from typical ISMN sites with the mean $ubRMSE = 0.036 \text{ cm}^3 \cdot \text{cm}^{-3}$. The results of this study initially prove the effectiveness of the land clustering algorithm for GNSS-R SM retrieval. Besides, after the initial evaluation in [21], this study further validated the ability of BF-1 to retrieve SM as a new spaceborne GNSS-R data source. From a new perspective, by taking full advantage of the similarity of land surface physical properties in different regions, the proposed algorithm provides an effective approach for global SM retrieval using spaceborne GNSS-R data.

REFERENCES

- [1] P. J. Wetzela and J.-T. Changb, "Concerning the relationship between evapotranspiration and soil moisture," *J. Appl. Meteorol. Climatol.*, vol. 26, no. 1, pp. 18–27, 1987.
- [2] A. Asoka and V. Mishra, "Anthropogenic and climate contributions on the changes in terrestrial water storage in India," *J. Geophys. Res.: Atmospheres*, vol. 125, no. 10, 2020, Art. no. e2020JD032470.
- [3] J. Pastor and W. M. Post, "Influence of climate, soil moisture, and succession on forest carbon and nitrogen cycles," *Biogeochemistry*, vol. 2, no. 1, pp. 3–27, 1986.

- [4] M. Scholze *et al.*, "Simultaneous assimilation of SMOS soil moisture and atmospheric CO₂ in-situ observations to constrain the global terrestrial carbon cycle," *Remote Sens. Environ.*, vol. 180, pp. 334–345, 2016.
- [5] J. Yin and X. Zhan, "Impact of bias-correction methods on effectiveness of assimilating SMAP soil moisture data into NCEP global forecast system using the ensemble kalman filter," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 659–663, May 2018.
- [6] G. Singh *et al.*, "Validation of SMAP soil moisture products using ground-based observations for the paddy dominated tropical region of India," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8479–8491, Nov. 2019.
- [7] H. Lu, W. T. Crow, Y. Zhu, Z. Yu, and J. Sun, "The impact of assumed error variances on surface soil moisture and snow depth hydrologic data assimilation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 11, pp. 5116–5129, Nov. 2015.
- [8] M. Sadeghi, E. Babaian, M. Tuller, and S. B. Jones, "The optical trapezoid model: A novel approach to remote sensing of soil moisture applied to sentinel-2 and landsat-8 observations," *Remote Sens. Environ.*, vol. 198, pp. 52–68, 2017.
- [9] D. Entekhabi *et al.*, "The soil moisture active passive (SMAP) mission," *Proc. IEEE*, vol. 98, no. 5, pp. 704–716, May 2010.
- [10] S. K. 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.
- [11] B. Liu, W. Wan, and Y. Hong, "Can the accuracy of sea surface salinity measurement be improved by incorporating spaceborne GNSS-Reflectometry?," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 1, pp. 3–7, Jan. 2021.
- [12] J. Reynolds, M. P. Clarizia, and E. Santi, "Wind speed estimation from CYGNSS using artificial neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 708–716, 2020.
- [13] 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," *Geophys. Res. Lett.*, vol. 44, no. 16, pp. 8369–8376, 2017.
- [14] M. Song *et al.*, "Study on the exploration of spaceborne GNSS-R raw data focusing on altimetry," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 6142–6154, 2020.
- [15] A. Di Simone, H. Park, D. Riccio, and A. Camps, "Sea target detection using spaceborne GNSS-R delay-Doppler maps: Theory and experimental proof of concept using TDS-1 data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4237–4255, Sep. 2017.
- [16] W. Li, E. Cardellach, F. Fabra, S. Ribó, and A. Rius, "Assessment of spaceborne GNSS-R ocean altimetry performance using CYGNSS mission raw data," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 238–250, Jan. 2020.
- [17] C. Chew and E. Small, "Description of the UCAR/CU soil moisture product," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1558.
- [18] C. Gerlein-Safdi and C. S. Ruf, "A CYGNSS-Based algorithm for the detection of inland waterbodies," *Geophys. Res. Lett.*, vol. 46, no. 21, pp. 12065–12072, 2019.
- [19] W. Li, E. Cardellach, F. Fabra, S. Ribó, and A. Rius, "Lake level and surface topography measured with spaceborne GNSS-Reflectometry from CYGNSS mission: Example for the lake qinghai," *Geophys. Res. Lett.*, vol. 45, no. 24, pp. 13332–13341, 2018.
- [20] H. Careno-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, 2020, Art. no. 1368.
- [21] W. Wan *et al.*, "Initial evaluation of the first chinese GNSS-R mission BuFeng-1 A/B for soil moisture estimation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 8017305.
- [22] 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.
- [23] C. Chew and 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.
- [24] 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, 2018.
- [25] 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.
- [26] 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.
- [27] V. Senyurek, F. Lei, D. Boyd, A. C. Gurbuz, M. Kurum, and R. Moorhead, "Evaluations of machine learning-based CYGNSS soil moisture estimates against SMAP observations," *Remote Sens.*, vol. 12, no. 21, 2020, Art. no. 3503.
- [28] 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, 2020, Art. no. 1168.
- [29] 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, 2019, Art. no. 2272.
- [30] C. Jing, X. Niu, C. Duan, F. Lu, G. Di, and X. Yang, "Sea surface wind speed retrieval from the first chinese GNSS-R mission: Technique and preliminary results," *Remote Sens.*, vol. 11, no. 24, 2019, Art. no. 3013.
- [31] V. U. Zavorotny and A. G. Voronovich, "Scattering of GPS signals from the ocean with wind remote sensing application," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 2, pp. 951–964, Feb. 2000.
- [32] P. E. O'Neill, S. Chan, E. G. Njoku, T. Jackson, R. Bindlish, and J. Chaubell, *SMAP L3 Radiometer Global Daily 36 Km EASE-Grid Soil Moisture, Version 6*, C. U. N. S. A. I. D. C. D. A. A. C. Boulder, Ed., Washington, DC, USA: NASA, 2019, doi: [10.5067/EVYDQ32FNWTH](https://doi.org/10.5067/EVYDQ32FNWTH).
- [33] S. Chan and S. Dunbar, "Level 3 passive soil moisture product specification document, version 7.0," Jet Propulsion Laboratory, NASA, Pasadena, CA, USA, Tech. Rep. JPL D-72551, Aug. 2020.
- [34] C. Carmona and A. José, *Application of Interferometric Radiometry to Earth Observation*. Barcelona, Spain: Adriano Camps, 1996.
- [35] S. H. Yueh, R. Shah, M. J. Chaubell, A. Hayashi, X. Xu, and A. Colliander, "A semiempirical modeling of soil moisture, vegetation, and surface roughness impact on CYGNSS reflectometry data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5800117.
- [36] T. Yang, W. Wan, J. Wang, B. Liu, and Z. Sun, "A physics-based algorithm to couple CYGNSS surface reflectivity and SMAP brightness temperature estimates for accurate soil moisture retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4409715.
- [37] 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. 84, no. C9, pp. 5699–5706, 1979.
- [38] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," Stanford InfoLab, Stanford, CA, USA, Tech. Rep., 2006. [Online]. Available: <http://ilpubs.stanford.edu:8090/778/>