Available online at www.sciencedirect.com

ScienceDirect

Advances in Space Research 73 (2024) 456–473

ADVANCES IN SPACE RESEARCH (a COSPAR publication)

www.elsevier.com/locate/asr

Retrieval and evaluation of surface soil moisture from CYGNSS using blended microwave soil moisture products

Zhounan Dong [a,b,*], Shuanggen Jin [c,d], Li Li [a,b], Peng Wang [a]

[a] School of Geography Science and Geomatics Engineering, Suzhou University of Science and Technology, Suzhou 215009, China
[b] Research Center of Beidou Navigation and Environmental Remote Sensing, Suzhou University of Science and Technology, Suzhou 215009, China
[c] School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454000, China
[d] Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China

Received 20 February 2023; received in revised form 15 August 2023; accepted 28 September 2023
Available online 4 October 2023

Abstract

The Global Navigation Satellite System-Reflectometry (GNSS-R) can estimate land surface soil moisture (SSM) as a viable and promising approach. However, it has some large uncertainty in retrieving SSM. In this study, the SSM is retrieved from different Cyclone GNSS (CYGNSS) SSM retrieval models formed with different SSM reference data products, including two blended microwave SSM products from the European Space Agency's Climate Change Initiative (CCI) and the National Oceanic and Atmospheric Administration's Soil Moisture Operational Product System (SMOPS), and a single microwave sensor-derived Soil Moisture Active Passive (SMAP) Level-3 product. The performance of the developed retrieval models, characterized by spatial resolutions of 36 km × 36 km and 0.25° × 0.25°, is evaluated using K-fold cross-validation. Furthermore, the accuracy of these models is validated against ground measurements acquired from Chinese automated soil moisture observation network. In order to alleviate the impact of spatial mismatching between the predicted gridded SSM and the point-scale in-situ measurements, a cumulative distribution function (CDF) rescaling strategy is applied. The results indicated that all models are effective at capturing spatial variations in SSM, and the SMOPS-based model achieves the highest correlation coefficient (0.930) and the lowest root mean square error (RMSD, 0.028 cm^3/cm^3), followed by the CCI-based model (0.906 and 0.042 cm^3/cm^3). The SMAP-based model performs poorly in the comparison. The suboptimal performance of models evaluated with Chinese automated soil moisture measurements is largely attributed to the insufficient calibration of the original reference data in the region.

Keywords: CYGNSS; Surface soil moisture; CCI; SMOPS; CDF matching

1. Introduction

Near-surface soil moisture (SSM) is a fundamental component of the water cycle, and its precise measurement holds paramount importance for a wide range of applications, such as weather forecasting, climate modeling, irrigation practices, and monitoring crop health and productivity (Conil et al., 2006). In past decades, the use of satellite remote sensing techniques to measure SSM has become increasingly popular among the geoscience and agriculture communities, as it allows for a better understanding of the Earth's systems and improved efficiency and productivity of human activities (Wang and Qu, 2009).

L-band satellite active and passive microwave sensors are renowned for their efficacy in retrieving SSM, rendering them reliable and vital instruments for acquiring SSM dynamics on a global scale. This is attributed to the L-band signal's superior ability to penetrate the Earth's sur-

* Corresponding author.
E-mail address: zndong@mail.usts.edu.cn (Z. Dong).

https://doi.org/10.1016/j.asr.2023.09.062
0273-1177/

face and reach the soil layer, thereby facilitating accurate measurements of moisture content (Hasan et al., 2014). With the advancement of remote sensing technologies, several dedicated SSM monitoring satellites have been launched, such as Soil Moisture Active Passive (SMAP) (Entekhabi et al., 2010) and Soil Moisture and Ocean Salinity (SMOS) (Kerr et al., 2010) missions. However, these dedicated monostatic active and passive remote sensors are subject to certain limitations, primarily stemming from their low measurement frequency, leading to a revisit cycle of 2–3 days, and their relatively lower spatial resolution, typically ranging from 30 to 50 km. In-situ measurements can provide the most accurate and higher temporal resolution in comparison to satellite-based observations (Tsegaye et al., 2004). However, the measured SSM information is confined to a small area around the stations and is generally sparse when considering large-scale regions. For the calibration and validation of spaceborne scatterometer and radiometer products, core validation sites are employed. These validation networks comprise many soil moisture stations within a single satellite footprint and adhere to specific quality control criteria (Gruber et al., 2020). As a results, ground station measurements primarily serve the purpose of calibrating and validating SSM products derived from satellite remote sensing.

Spaceborne Global Navigation Satellite System-Reflectometry (GNSS-R) is a novel remote sensing technique that utilizes the signals transmitted by Global Navigation Satellite Systems (GNSS), such as GPS, BeiDou, GLONASS, and Galileo, reflected from the Earth's surface to measure various geophysical parameters (Clarizia et al., 2014). This technique has several advantages over traditional dedicated active and passive remote sensing techniques, such as radar and optical sensors, including wide coverage, high resolution, low cost, and the ability to repeat measurements over time. Additionally, it can be used to measure a wide range of parameters including soil moisture (Camps et al., 2016), ocean surface wind (Foti et al., 2015), sea ice cover (Yan and Huang, 2016), and vegetation density (Camps et al., 2016). The receiver processes the reflected GNSS signal in the form of a Delay-Doppler Map (DDM), which serves as the fundamental observation quantity in spaceborne GNSS-R. The DDM observables were used as the basis for retrieving geophysical parameters.

In the past decade, there has been a surge in spaceborne GNSS-R remote sensing studies. GNSS-R technology has been demonstrated to be effective for observing ocean and land geophysical parameters, leading to the development of retrieval algorithms. The National Aeronautics and Space Administration (NASA) launched the Cyclone Global Navigation Satellite System (CYGNSS) mission to examine tropical cyclones and factors that influence their intensity (Rose et al., 2014). The primary objective of CYGNSS is to gain a deeper understanding of the inner processes of tropical cyclones. The CYGNSS has produced vast amounts of observations which are expected to provide invaluable information for a variety of applications. Several methodologies for repurposing CYGNSS data for SSM estimation have been presented in literature. The previous spaceborne CYGNSS SSM inversion algorithms are similar to passive radiometry, forming the statistic quantified fitting model, which is heavily dependent on the quality of the collocated fiducial reference SSM values. Many successful investigations have relied on establishing a statistical model that relates the aggregated DDM-derived effective reflectivity to collocated SSM values obtained from other observing systems. Chew and Small (2018) utilized the spatial average approach in their study, where the retrieval algorithm employed a linear model to regress the variations in effective reflectivity and SSM. The retrieval algorithm assumed that effective reflectivity changes related to vegetation and roughness occurred on timescales longer than those associated with soil moisture changes. Based on this idea, Al-Khaldi et al. (2019) presents a method for retriving SSM using data from the CYGNSS constellation. The approach focuses on incoherent scattering from land surfaces. Clarizia et al. (2019) also employed a standard statistical inversion procedure that included spatial average and linear regression. The empirical statistical model between the gridded effective reflectivity, vegetation opacity, roughness coefficient, and SSM was established using trilinear regression. Yan et al. (2020) study pan-tropical SSM mapping based on a three-layer model from CYGNSS data. Some studies have sought to enhance SSM inversion accuracy by utilizing advanced artificial intelligence. This includes the use of deep learning algorithms and machine learning to estimate SSM from GNSS-R observations (Eroglu et al., 2019; Jia et al., 2021; Lei et al., 2022; Nabi et al., 2022; Senyurek et al., 2020). These approaches offer the advantage of capturing complex non-linear relationships between spaceborne GNSS-R measurements, SSM and other influencing factors, yielding promising outcomes in producing accurate and reliable SSM estimates. Nevertheless, the current results demonstrate a proximity to those achieved by traditional empirical regression methods.

One of the current debates in the field of GNSS-R terrestrial remote sensing is the relative significance of the coherent and incoherent components of scattered GNSS signals. Dong and Jin (2021) has investigated the classification of received terrestrial coherent and incoherent signals and their impact on SSM retrieval. The results showed that the proportion of signals dominated by incoherence was low and their influence on SSM retrieval was limited. In addition, prior to using the methods mentioned in this study, different pre-processing steps are required. Ground reflectivity, which is a major factor in responding to SSM in GNSS-R detection, is also influenced by various factors, including plant cover, topography, and small-scale surface roughness (Yan et al., 2020). These effects are interrelated and must be considered when conducting SSM retrieval. In previous studies, these factors have been considered, and various ancillary parameters and correction models

were employed to mitigate their effects (Al-Khaldi et al., 2019; Chew and Small, 2020; Yan et al., 2020; Yueh et al., 2022). However, most starting points have resorted directly to approaches developed for satellite-based radiometry (Chew and Small, 2020, 2018).

To date, numerous SSM products have been generated utilizing data from satellite sensors, land surface model, and merge algorithms. These products exhibit diverse spatial and temporal coverage, resolution, and quality. However, it is worth noting that the most of prior spaceborne GNSS-R land SSM retrieval methods rely solely on single-sensor-based SSM products as the ground reference values, mainly sourced from SMAP and/or SMOS missions (Chew and Small, 2020; Lei et al., 2022; Yan et al., 2020). Blended SSM products, combining data from multiple satellite sensors, have been consistently shown to outperform other single-sensor-based SSM products in terms of accuracy and spatial coverage (Wang et al., 2021). As demonstrated in Ma et al. (2019), it found that the overall performance of the European Space Agency's (EAS) CCI was superior to that of the SMOS, Advanced Microwave Scanning Radiometer 2 (AMSR2), and SMAP Level 3 products. These products are specifically designed to enhance the accuracy and spatial resolution for SSM estimates across various land cover types, making them valuable for a wide range of hydrological and climatic applications. Currently, there are only two sets of blended SSM products: the Climate Change Initiative (CCI) and the Soil Moisture Products System (SMOPS), which have different data sources, merging methods, and time intervals (Dorigo et al., 2017; Liu et al., 2016). The CCI product is distinguished for its meticulous data quality control procedures, which guarantee high accuracy and data reliability. Conversely, SMOPS provides extended spatial and temporal coverage, broadening its applicability for diverse research and monitoring purposes. The increased coverage of blended products compared to single-sensor-based products such as SMAP and SMOS products can be beneficial for studies that require a large sample size or that aim to investigate regional or temporal soil moisture trends. Additionally, validation of the retrieved SSM is crucial for evaluating the performance of the retrieval algorithm and identifying the sources of uncertainty (Gruber et al., 2020). This can be achieved through comparison with ground-based measurements or other satellite-based SSM products. The validation process helps to establish confidence in the accuracy and reliability of the data and supports various applications.

This study utilized blended SSM products from the ESA CCI and NOAA SMOPS as reference data in comparison to SMAP data for the CYGNSS SSM retrieval modeling. The following discrepancies stand out from previous research that used CYGNSS observations for SSM retrieval: (1) To make good use of blended SSM products with higher accuracy and spatial coverage in comparison to solely adopting SMAP or SMOS single-sensor-derived SSM products as inversion reference data. (2) Utilizing the Chinese soil moisture automatic measurement network stations for evaluation, a scaling strategy was used to remove the systematic differences in the spatial mismatch between the derived gridded SSM products and in-situ measurements. (3) Can merged products with a higher spatial resolution generate an accurate CYGNSS-based SSM retrieval model? (4) Determine whether the current CYGNSS-derived SSM can capture regional SSM dynamics. We expect that the derived results will be scrutinized to better understand the capabilities of spaceborne GNSS-R terrestrial SSM remote sensing, which can improve current coarse-scale satellite-based SSM products of radiometry with successive development in the near future. Current and future spaceborne GNSS-R missions can benefit and improve retrieval algorithms for quantitatively mapping global SSM for various applications. In the following Section 2, the experimental dataset, retrieval approach, and evaluation metrics are described, the main results and analysis are presented in Section 3, some discussions are shown in Section 4, and finally the summary and findings are provided in Section 5.

2. Data and method

2.1. Dataset description

2.1.1. CYGNSS dataset

The CYGNSS is a satellite mission initiated by the NASA to investigate tropical cyclones and the factors governing their intensity using GNSS-R. The mission consisted of a constellation of eight small satellites orbiting in a same orbital plane launched into a low Earth orbit in 2016. The observation coverage encompasses latitudes ranging from 38° south to north. The CYGNSS dataset became accessible in March 2017, and the experiment outlined in this study leverages CYGNSS Level-1 Version 2.1 data, sourced from the Physical Oceanography Distributed Active Archive Center (PO.DAAC, https://podaac-opendap.jpl.nasa.gov/opendap/allData/cygnss/L1/v2.1/), to extract the parameters essential for SSM retrieval. Land observations can be extracted from the product file using the per-DDM quality flags parameter, furnished with 16-bit flag masks. Invalid observation data is filtered out by utilizing a combination of various flag bits, which include criteria such as "S-band transmitter powered up," "spacecraft attitude error," "black body DDM," "DDM is a test pattern," "direct signal in DDM," and "low confidence in the GPS EIRP estimate" (Chew and Small 2020).

2.1.2. SMAP dataset

The study uses the SMAP v008 Level-3 SSM product downloaded from the National Snow and Ice Data Center, which is updated daily with a 36 km × 36 km spatial resolution gridded on the Equal-Area Scalable Earth Grids 2.0 (EASE-Grid2) grid. The SSM data from satellite descending (a.m.) and ascending (p.m.) passes, which were recorded separately in the Level-3 product files, were com-

bined by taking the average values to create a daily SMAP SSM map, which were used for SSM retrieval modeling. The valid range for SMAP SSM map is 0.02–0.5 cm^3/cm^3. It is important to emphasize that we only used the regions recommended by the data provider flag in the metadata. The vegetation opacity (''vegetation_opacity_dca'') in the SMAP product also undergo averaging for vegetation correction in the retrieval modelling (ONeill, Peggy E. et al., 2021, p. 8).

2.1.3. SMOPS soil moisture product

The Soil Moisture Products System (SMOPS), developed by the National Oceanic and Atmospheric Administration and the National Environmental Satellite Data and Information Service (NOAA/NESDIS) center, provides high-quality soil moisture data for various purposes, including weather forecasting, agriculture, and natural resource management. The dataset combines SSM products from five satellites, including GPM, SMAP, GCOM-W1, SMOS, and MetOp-B, resulting in enhanced accuracy and spatial resolution for SSM estimates over global land. SMOPS v3.0 provides daily blended products, with 0.25° × 0.25° grid SSM maps generated every 6 h and daily to meet the needs of different users. The 6-hourly product is available in GRIB2 format at standard forecast times (00Z, 06Z, 12Z, and 18Z). The maps have almost full land coverage and can fill the gaps left by most currently available single satellite soil-moisture products. The verification accuracy of the product against ground station measurements can reach 0.046 cm^3/cm^3 according to the Wang et al. (2021). The SMOPS SSM map for January 1, 2018, is shown in Fig. 1 to demonstrate the spatial coverage and dynamic range of the data.

2.1.4. CCI soil moisture product

The Soil Moisture Climate Change Initiative (CCI) is part of the European Space Agency's (ESA) program for monitoring essential climate variables. Starting in 2010, the project produced updated soil moisture data every year with an average accuracy of 0.042 cm^3/cm^3 compared with in in-situ stations. The data had a spatial resolution of 0.25° × 0.25° grids, with 27.8 km resolution in the equator. The dataset is a combination of SSM data derived from various sensors and processed through distinct algorithms, resulting in three separate SSM products from the active sensors, passive sensors, and combined product (Dorigo et al., 2017; Gruber et al., 2019). In this study, the combined product (ESA CCI COMBINED v06.1) was used as the reference data for the CYGNSS-based SSM retrieval modeling. The CCI SSM map for January 1, 2018, is shown in Fig. 2 to demonstrate the spatial coverage and dynamic range of the data.

2.1.5. Other dataset

The vegetation layer mainly causes the attenuation of the received power of the reflected GNSS signal. To decouple the effects of vegetation and SSM on the observation signals and increase the retrieval accuracy of the CYGNSS-derived SSM, external vegetation parameter data can be utilized directly to mitigate the vegetation impact. Konings et al. (2017) developed an algorithm called the multi-temporal dual-channel algorithm (MT-DCA), which retrieves SSM and vegetation optical depth from SMAP L1C brightness temperature products using a robust estimation technique. The product is currently accessible on 9 km and 36 km grids from April 2015 to July 2021. Although this is not an official product and did not go through the same calibrating validation procedures as the official SMAP criteria, it has been used effectively in several recent studies to reveal the behavior of tropical forests, semi-arid grasslands, and crops (data acquisition: https://afeldman.mit.edu/mt-dca-data). In this study, 9 km × 9 km the EASE-Grid2 data products were employed and resampled to the basic regular grid for vegetation correction in the CYGNSS SSM retrieval.

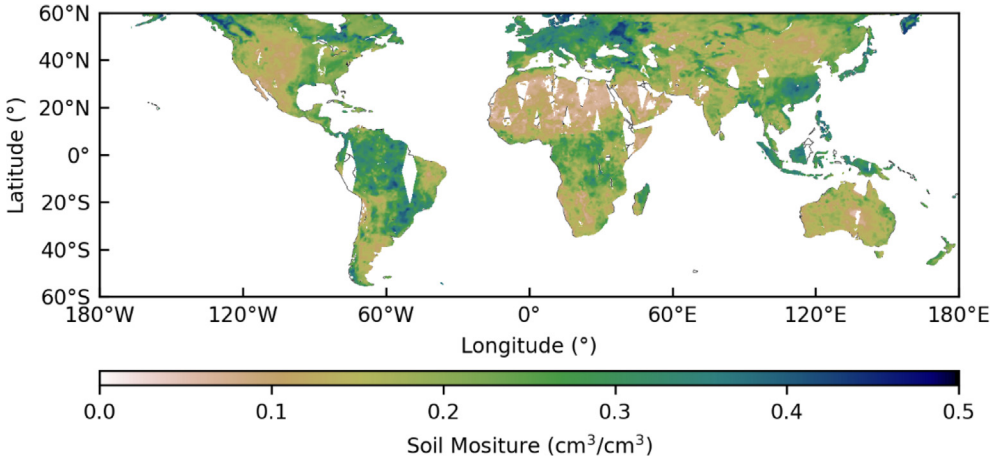

Fig. 1. SMOPS surface soil moisture product at Jan 1st, 2018.

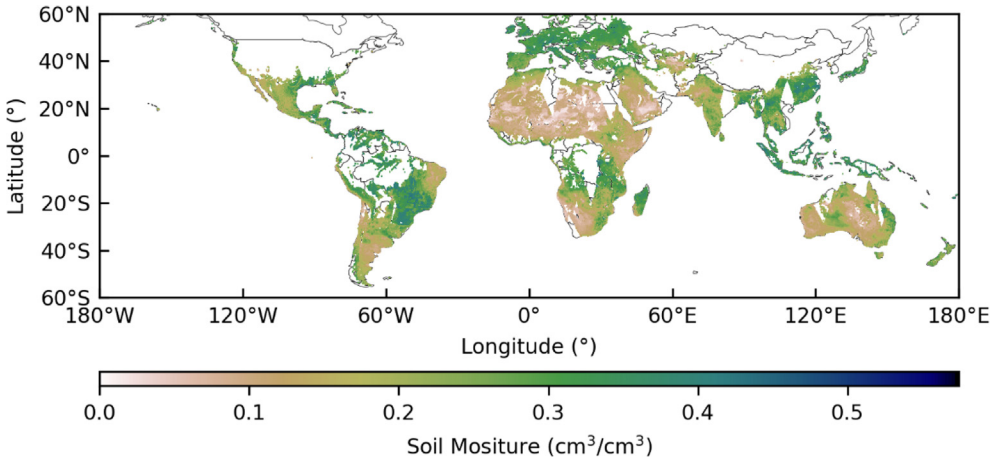

Fig. 2. CCI surface soil moisture product on Jan 1st, 2018.

2.1.6. In-situ measurements

Automatic soil moisture monitoring stations offer precise and continuous measurements of soil moisture across multiple depth layers. The ground stations operate in real time, providing valuable support for agricultural production and hydrological applications. Currently, the China Meteorological Administration has established Chinese automated soil moisture network consisting of more than 2,000 stations spanning the mainland, achieving near real-time uploading of measurements for data product integration. The published hourly element dataset of soil moisture information included four soil state parameters: volumetric water content, relative humidity, mass content, and available water at different depths. The stations are equipped with Frequency Domain Reflection (FDR) sensors, which are utilized for the measurement of volumetric water content. Simultaneously, the remaining three elements are computed through a combination of the acquired data and the soil hydrological parameters. In this study, quality control procedures were applied to in-situ measurements taken at shallow depths of 0–10 cm with hourly temporal resolution within the coverage of CYGNSS (Saeedi et al., 2021). These measurements were subsequently resampled at a daily scale and matched with the gridded SSM products to evaluate the SSM inversion model. Fig. 3 illustrates the spatial distribution of the soil moisture monitoring stations within the China region that were utilized in this study.

2.2. Methodology

2.2.1. Land GNSS-R observable

The spaceborne GNSS-R system collects scattered power of GNSS signals from the Earth's land surface to infer various soil moisture conditions. The scattered power is characterized by cross correlating the reflected signals with the local replica code of the receiver. The DDM quantifies the power distribution as a function of the time delay and Doppler frequency shifts. It serves as the primary and fundamental observation derived from GNSS-R receivers. The CYGNSS mission employs a 1 ms coherent integration time to capture the signals, followed by an incoherent integration of 0.5 to 1 s to minimize the impact of speckle and thermal noise. The Z-V model offers a useful approximation of the scattering mechanisms at the scattering surface (Zavorotny et al., 2014), with ocean surfaces exhibiting mainly incoherent scattering, whereas land surfaces tend to exhibit coherent scattering originating from the specular reflection direction. This study assumes that the land surface is dominated by coherent reflection and that the first Fresnel zones near the specular point are homogeneous. The received power can be approximated by the free-space propagation value modulated by the reflection coefficient. The total bistatic radar system path length was calculated as the sum of individual path lengths. The DDM reflectivity, also known as the effective reflectivity, can be determined through calibration using the radar equation for a coherent signal.

$$\Gamma(\theta) = \frac{(4\pi)^2 P_{coh}(R_{ts} + R_{rs})^2}{\lambda^2 G_r P_t G_t} \tag{1}$$

where P_{coh} is the received DDM peak power, $P_t G_t$ indicates the GNSS equivalent *iso*-tropically radiated power (EIRP), P_t is the GNSS satellite transmit power, G_t is the GNSS satellite antenna gain, G_r is the gain of the GNSS-R receiver antenna, λ is the carrier wavelength of the GNSS signal, R_{ts} and R_{rs} are the distances from the GNSS transmitter to the specular point and specular point to the receiver, respectively, and θ is the incidence angle of the signal. All the pertinent parameters are encompassed within CYGNSS Level-1 product file.

Effective reflectivity serves as a proxy for SSM, although it can also be influenced by other non-relevant spatial and temporal factors, including topography, surface roughness, vegetation cover type, inland water. Hence, in GNSS-R soil moisture retrieval, it becomes imperative to mitigate the impact of these influencing factors to the greatest extent feasible. Among these factors, the influence of surface veg-

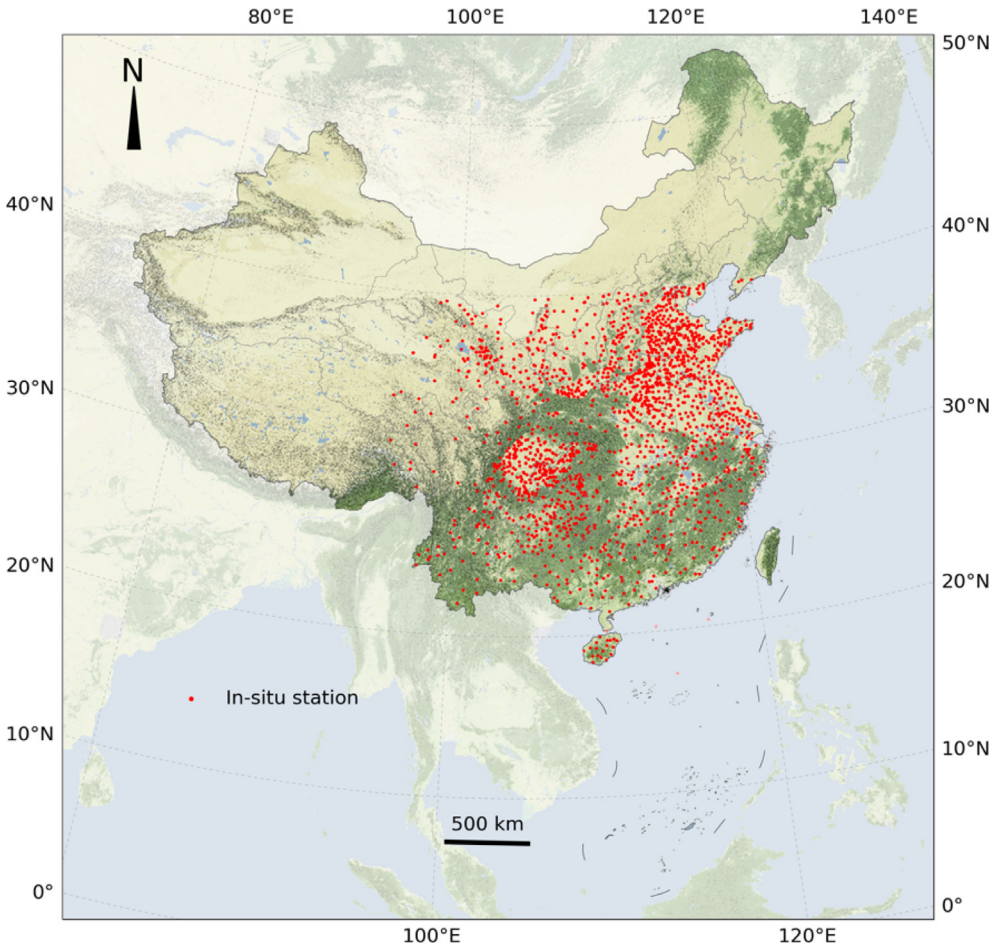

Fig. 3. Distribution of Chinese automated soil moisture observation stations within the scope of CYGNSS mission.

etation and surface roughness on effective reflectivity stands out prominently. Through the incorporation of vegetation parameters and surface roughness parameters, the effective reflectivity can be corrected using the following method:

$$\Gamma = \Re(\theta)^2\gamma^2 e^{\left(-4k^2s^2\cos^2\theta\right)} \quad (2)$$

The exponential term on the right-hand side of the equation indicates surface roughness attenuation. k is the wavenumber, s is the standard deviation of the surface height, and θ is the measurement incidence angle. The weakening of the signal caused by vegetation, can be expressed by the vegetation transmissivity $\gamma = e^{-2\tau sec\theta}$, where τ is the vegetation optical depth. It should be noted that owing to the absence of reliable small-scale surface roughness data and more refined corrective models, a surface roughness correction method similar to the SMAP SSM retrieval was applied to the CYGNSS SSM retrieval but provided limited improvement. Therefore, only the vegetation optical depth parameters provided by SMAP and the estimated vegetation parameters in Konings et al. (2017) were used to correct the CYGNSS effective reflectivity in relation to vegetation effects.

2.2.2. Retrieval algorithm

In this study, the CYGNSS SSM retrieval algorithm employs the space–time averaging approach to aggregate the discrete effective reflectivity values calculated at individual specular points into a grid format, aligning them with the reference SSM data. The semiempirical model was developed through linear regression, involving the aggregation of corrected effective reflectivity values against the reference SSM at each grid cell. In the course of SSM retrieval with CYGNSS measurements, we utilize the per-DDM quality flags parameter provided by the CYGNSS Level-1 data files to ensure the reliability of observations. Additionally, empirical quality control criteria are integrated into the process: the delay bin of the DDM peak power must fall within the 7–10 bin interval, the DDM SNR must be at least 2 dB, the receiver antenna gain at the specular point must be 0 dB or greater, and the specular incidence angle must be less than 60°. The CYGNSS observation geometry also affects the effective reflectivity; therefore, a normalization method has been implemented to adjust the effective reflectivity for different incidence angles to the nadir direction (Al-Khaldi et al., 2019).

Fundamentally, the retrieval methodology hinges on establishing a fitting regression model that correlates the

selected reference SSM values from measurement systems with the effective reflectivity deduced through GNSS-R. Therefore, the CYGNSS-based SSM estimates will be largely limited by the retrieval performance of the reference SSM source. In the aggregation and regridding step for effective reflectivity, grid cells with fewer than 5 counts of the average effective reflectivity were deemed invalid and were not used in fitting the model. Moreover, grid cells marked as urban, hilly, and inland water in the reference product were omitted. Subsequently, a pixel-by-pixel linear model was constructed using all accessible collocated training data. Using the established retrieval model, one can predict daily SSM using the following equation:

$$\mathbf{M}_{\mathrm{v}}^{\mathrm{CYGNSS}} = \mathbf{A}\Gamma_{\mathrm{gridded}} + \mathbf{B} \tag{3}$$

where $\mathbf{A}$ is the coefficient matrix of the linear model, $\mathbf{B}$ is the intercept matrix of the linear model, $\Gamma_{\mathrm{gridded}}$ is the gridded GNSS-R effective reflectivity, and $\mathbf{M}_{\mathrm{v}}^{\mathrm{CYGNSS}}$ represents the predicted SSM from GNSS-R.

2.2.3. Scaling method

The spatial mismatch between coarse-resolution gridded satellite SSM products and in-situ measurements is a critical challenge in the validation of satellite-based products. Several studies building upon the temporal stability concept introduced by the foundational research of Brocca et al. (2011), based on in-situ measurements, have shown that point-scale SSM time series can be indicative of broader regions. This demonstrates that the temporal pattern of local SSM measurements closely matches that of the spatial average. However, despite the relative comparability of temporal dynamics, systematic discrepancies between satellite-derived products and in-situ measurements are commonly observed, which are referred to as representativeness errors. Cumulative Density Function (CDF) matching is one of the most common scaling methods used to compare and adjust the differences between gridded satellite remote sensing products and in-situ measurements (Ma et al., 2019). This approach involves calculating the CDF for each dataset and then aligning the two CDFs by adjusting the satellite dataset to match the in-situ measurements. The rescaled coarse-resolution SSM products are subsequently applicable for further evaluation and wider application. To implement the CDF matching method, commence by computing the CDFs for both the satellite dataset and in-situ measurements. Next, derive the disparities between the two CDFs within each bin and plot these differences against the satellite data. Subsequently, fit a polynomial function to calculate bias corrected for the satellite dataset. The transformed satellite data will retain the distribution of the in-situ measurements, though the actual values will differ.

2.2.4. Experimental design

To evaluate the effectiveness of the blended SSM products within the CYGNSS SSM retrieval, two distinct experimental strategies were employed. Initially, the high-accuracy CCI and SMOPS data were resampled and regridded to the 36 km × 36 km EASE-Grid2 grid using nearest neighbor interpolation, aligning them with the grid used in the SMAP L3 SSM product. Furthermore, the CCI and SMOPS datasets were screened based on the spatial coverage of the recommended data in the SMAP product with the same timestamp to align the different datasets. This process enabled the evaluation of the impact of reference data quality on the retrieval results. Vegetation attenuation correction uses the vegetation opacity parameter from the SMAP L3 product produced by the dual-channel algorithm (ONeill, Peggy E. et al., 2021) to compute the two-way transmissivity of the canopy applied directly to the CYGNSS-derived grid point effective reflectivity. The models generated using the SMAP, CCI, and SMOPS reference data allowed for a comparison of the performance of different datasets based on their corresponding estimations. The 12-fold cross-validation was used in the CYGNSS SSM retrieval modeling to evaluate the performance of different reference data-generated inversion models.

Next, the raw blended SMOPS and CCI SSM products not only have high accuracy and spatial coverage but also have higher spatial resolution compared to SMAP data. To understand the capability of CYGNSS in generating higher resolution SSM products, the CYGNSS SSM retrieval model was developed using raw reference data products with the original resolution of 0.25° × 0.25°. To align with referenced blended data, individual specular effective reflectivity from CYGNSS was projected onto a cylindrical grid at a spatial resolution of 0.25° × 0.25° grids, resulting in an approximate spatial resolution of 28 km within the tropics. The model performance was also evaluated using a 12-fold cross-validation.

2.2.5. Statistical analyses

Evaluating model performance holds paramount significance within satellite remote sensing research, often facilitated by the widely adopted k-fold cross-validation method. This technique involves partitioning data into k subsets, enabling iterative model training and evaluation. In each iteration, a distinct subset is designated as the test set, while the remaining k-1 subsets serve as the training set. For this study, the value of k was set at 12, meaning that each iteration utilized approximately 11 months of data for training the SSM retrieval model and one month for testing. Nevertheless, it's important to acknowledge that the lack of established core validation stations for quality assessment of CYGNSS-derived SSM products presents a challenge in achieving absolute uncertainty measurements, which is in contrast to the situation with traditional radiometer and scatterometer products. In this study, the evaluation outcomes primarily indicate whether the retrieval results exhibit wetter or drier conditions compared to in-situ measurements. The performance evaluation of CYGNSS-derived SSM encompasses several

standard skill metrics, including mean bias (bias), Mean Absolute Error (MAE), Root-Mean-Square Deviation (RMSD), Pearson correlation coefficient (R-value), and unbiased-RMSD (ubRMSD).

3. Results and analysis

3.1. Accessibility of reference data sets

The quality and quantity of the reference data have a significant impact on the accuracy of satellite-based GNSS-R SSM retrieval modelling. To highlight the benefits of the blended soil moisture product in terms of spatial coverage, Fig. 4 shows the number of days in 2018 for which individual reference data points were available from the SMAP, SMOPS, and CCI SSM products, counted over each grid pixel. Regarding the statistics presented here, it's important to note that the SMAP data encompasses all available data points, even those from tropical rainforest areas. Nonetheless, it's a common practice to solely utilize recommended data points featuring low uncertainty in applications. This practice leads to the masking of regions that resemble undocumented areas within the CCI dataset. However, in terms of spatial and temporal coverage, the blended SSM products have a wider range than single-sensor-based SMAP data. Among the blended products, SMOPS exhibited the highest spatiotemporal coverage. The CCI data volume decreases at high latitudes, which is consistent with the characteristics of the individual products, as the generated CCI dataset has undergone meticulous quality control within its merging algorithm (Dorigo et al., 2015). The greater spatiotemporal coverage of the blended products offers more training samples and probably provides a larger dynamic range within the space and

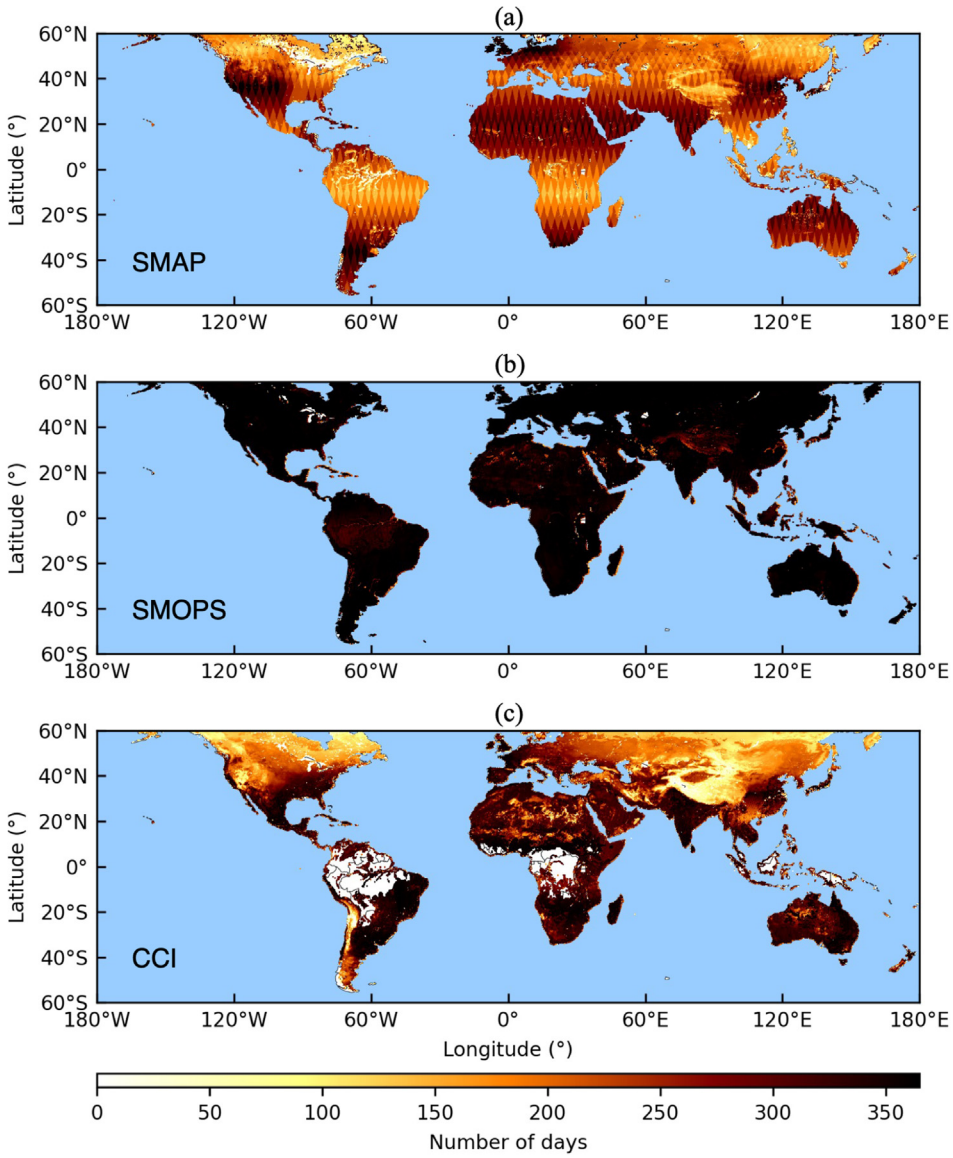

Fig. 4. Distribution of the number of available days for each pixel of SMAP (a), SMOPS (b), and CCI (c) surface soil moisture products in the year 2018.

time gaps in the SMAP data for GNSS-R SSM retrieval modeling.

3.2. Model performance of higher accuracy blended SSM products

Despite the inherent representativeness errors introduced by upscaling, for the purpose of comparing the higher-accuracy blended SSM products as reference data for CYGNSS SSM retrieval modeling with the final CYGNSS-derived estimates, the SSM maps from CCI and SMOPS were resampled. This resampling ensured spatial consistency with the SMAP data products. The spatial coverage of both CCI and SMOPS was aligned with SMAP data for the corresponding date, achieved through the application of recommended valid SMAP data to establish spatial masking. Utilizing the CYGNSS SSM retrieval algorithm outlined in Section II, we constructed the retrieval model for performance evaluation, employing three distinct sources of reference data. Vegetation opacity parameters were screened consistently with the reference datasets and CYGNSS-derived gridded effective reflectivity.

Table 1 presents the mean performance and overall standard deviation (STD) of predicted SSM values across the 12-fold cross-validation of the test dataset, as compared to the corresponding reference data. All three retrieval model predictions exhibit a mean bias near zero, with a STD of 0.006 cm^3/cm^3 for SMAP, 0.004 cm^3/cm^3 for SMOPS, and 0.005 cm^3/cm^3 for CCI, indicating that the model predictions do not have a significant systematic bias. The MAE for SMAP data-based retrieval model estimations is 0.036 cm^3/cm^3 with a STD of 0.002 cm^3/cm^3, for SMOPS is 0.021 cm^3/cm^3 with a STD of 0.001 cm^3/cm^3, and for CCI is 0.033 cm^3/cm^3 with a STD of 0.003 cm^3/cm^3. This shows that SMOPS-based model prediction has the lowest MAE. The RMSD for SMAP is 0.054 cm^3/cm^3 with a STD of 0.002 cm^3/cm^3, for SMOPS is 0.029 cm^3/cm^3 with a STD of 0.002 cm^3/cm^3, and for CCI is 0.045 cm^3/cm^3 with a STD of 0.004 cm^3/cm^3. Given the negligible bias in the predicted SSM values of all three models, the ubRMSD closely resembles the RMSD, indicating that the SMOPS-based model has the smallest unbiased error. The R-value for SMAP is 0.905 with a STD of 0.008, for SMOPS is 0.926 with a STD of 0.009, and for CCI is 0.892 with a STD of 0.019. SMOPS-based results have the highest correlation coefficient, meaning it has the highest linear relationship between predicted and reference values. In summary, the overall evaluation results show that the SMOPS dataset generated retrieval model have the best performance, followed by CCI, with the worst results for SMAP.

The ensemble predicted SSM maps derived from each fold test dataset were rigorously compared to the referenced SSM maps, and the scatter density plot of spatiotemporal collocated grid point data pairs for individual datasets is illustrated in Fig. 5. The density plots confirm a strong agreement between the model predicted SSM maps and the reference data. Notably, the scatter points on both sides of the 1:1 diagonal line are evenly dispersed, with density intensifying closer to the diagonal. Nevertheless, distinct differences remain apparent in the distribution patterns within the three figures. The blended SSM matching pairs demonstrate superior performance, displaying enhanced alignment, whereas the SMAP result display greater dispersion. Additionally, it's noteworthy that the SMAP products claimed a narrower valid range, spanning from 0.02 cm^3/cm^3 to 0.5 cm^3/cm^3, whereas the SMOPS and CCI data encompassed a broader valid range, from 0.0 cm^3/cm^3 to 1.0 cm^3/cm^3. The SSM pairs involving SMOPS and the corresponding predicted outcomes exhibited a tighter concentration within intervals compared to the matching pairs involving CCI and SMAP. The statistical results for the corresponding data pairs agreed with the average scores obtained over 12-fold cross-validation.

Fig. 6 shows the spatial maps of RMSD and R-value on 36 km × 36 km EASE-Grid2 grids of the SMAP-based, SMOPS-based, and CCI-based models from all 12 iterations of testing to determine the performance of each model. These maps, presented in Fig. 6(a), (c), and (e), represent the spatial variability of the RMSD within the CYGNSS coverage. Blue shading signifies lower RMSD values, whereas red shading indicates higher RMSD values. Notably, the models effectively captured the spatial distribution characteristics of the SSM. Across arid regions, SSM values exhibited subtle fluctuations and minimal model errors. Conversely, in moist surface regions characterized by substantial annual SSM variability, the prediction accuracy of the model declined. The SMOPS-based model consistently demonstrated reduced RMSD values compared to the SMAP-based and CCI-based models. The lowest values corresponded to areas with sparse vegetation, whereas the highest values were evident in densely vegetated regions. This trend holds true for regions like the Sudanian Savanna in western and central Africa, where notable disparities in model performance are observed. This region is covered with grassland and woodland, where receives an average of 600–800 mm of rainfall per year. In

Table 1
The average skill metrics and standard deviation of 12-fold cross-validation of retrieval models from EASE-Grid2 grid reference data (cm^3/cm^3).

Skill Metrics	Bias (STD)	MAE (STD)	RMSD (STD)	R (STD)	ubRMSD (STD)
SMAP	0.0002 (0.006)	0.036 (0.002)	0.054 (0.002)	0.905 (0.008)	0.054 (0.002)
SMOPS	0.0002 (0.004)	0.021 (0.001)	0.029 (0.002)	0.926 (0.009)	0.029 (0.002)
CCI	0.0001 (0.005)	0.033 (0.003)	0.045 (0.004)	0.892 (0.019)	0.045 (0.004)

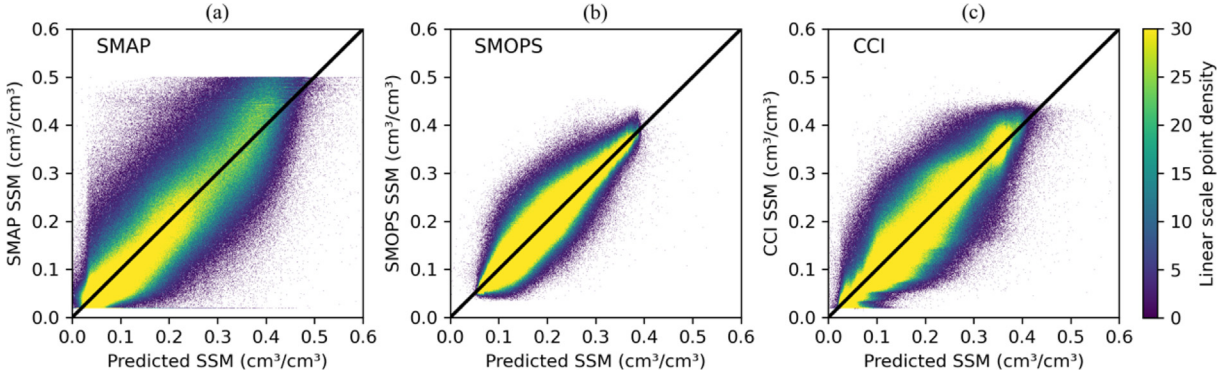

Fig. 5. Scatter density plot of ensemble data pairs from the model-predicted surface soil moisture from the test dataset, and referenced SMAP (a), SMOPS (b), and CCI (c) data.

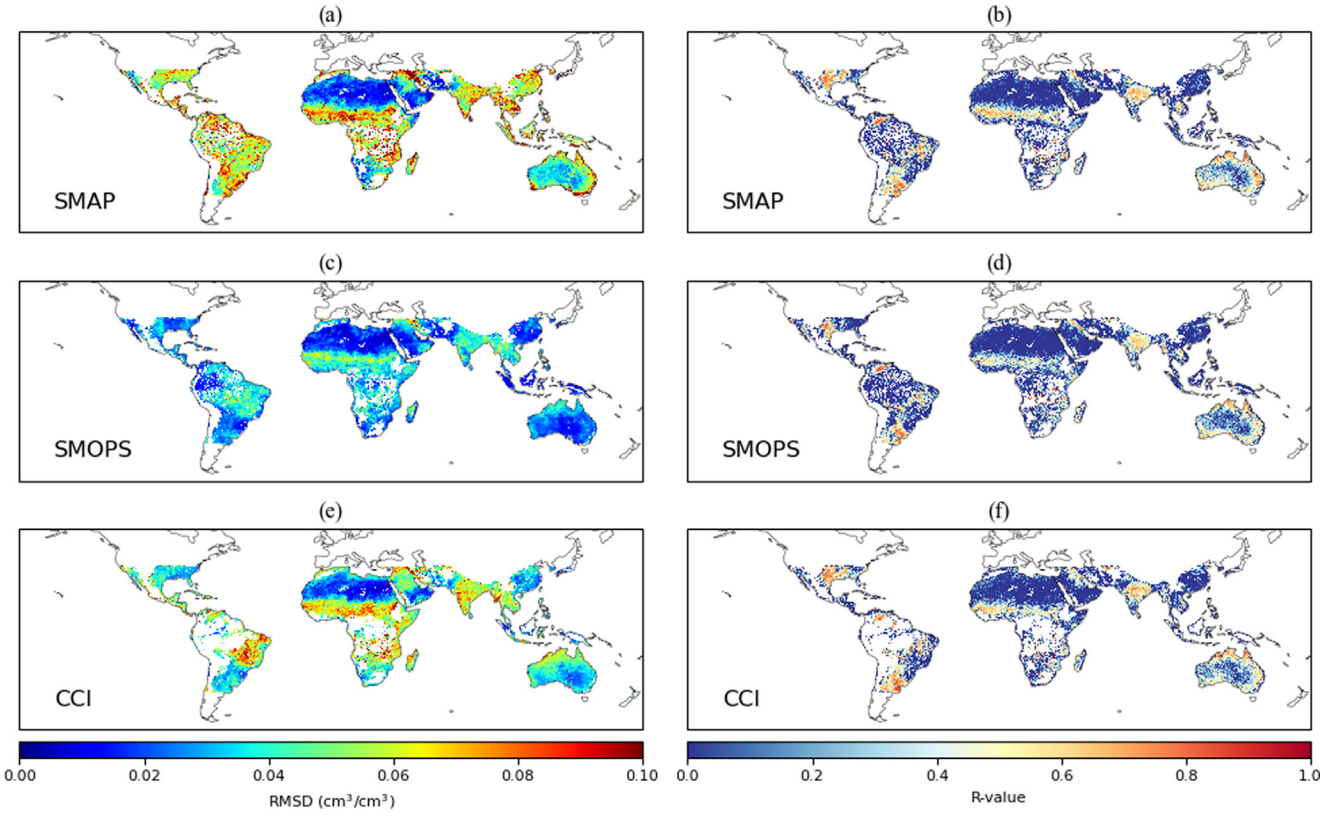

Fig. 6. Spatial distribution of the RMSD and correlation coefficient of three models based on SMAP (a, b), SMOPS (c, d), and CCI (e, f).

addition, the southern part of China revealed a relatively large difference among the three models. As for the R-value maps, prevailing lower R-values were observed across most arid regions, while wet areas displayed higher R-values. SMOPS-based model exhibited superior performance compared to the other two models, as depicted in Fig. 6(b), (d), and (f). The observed differences in model performance can be primarily attributed to the reference data and resampling algorithm employed. This underscores the critical significance of judiciously selecting reference data and employing suitable resampling techniques to ensure accurate evaluation of SSM retrieval model performance.

Temporal performance was further evaluated through quantification using daily bias, MAE, RMSD, ubRMSD, and correlation across 12-fold cross-validation for different reference data, as illustrated in Fig. 7 (a). The performance of the three models closely aligns with the ensemble average statistics mentioned earlier. Consistently, SMOPS demonstrated superior performance, followed by CCI, with SMAP showing the worst performance. All three models exhibited minimal bias, hovering around 0.0 cm^3/cm^3. Consequently, the time series of daily RMSD and ubRMSD demonstrated similarity across all three models. Among the three models, SMOPS exhibited the lowest values for MAE, RMSD, and ubRMSD, maintaining stability

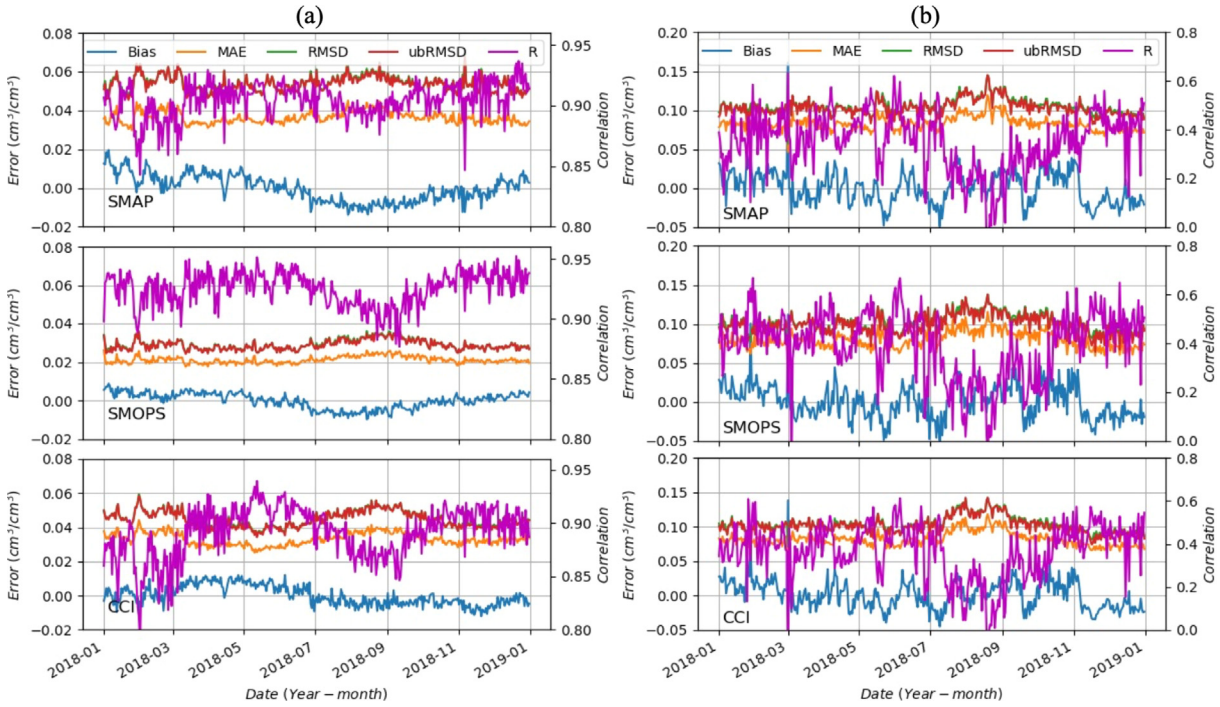

Fig. 7. Temporal skill metrics comparison between the estimated surface soil moisture from the testing dataset and referenced gridded data (a), and in-situ measurements data (b).

throughout the year. CCI and SMAP closely followed, with a similar performance trend. Nevertheless, for the R-value of the blended SSM dataset, CCI and SMOPS displayed a more pronounced overall decreasing trend between July and October.

3.3. Validation by in-situ measurements

The evaluation of the generated models was conducted using independent in-situ measurements on a daily basis. To ensure accurate comparisons, a rescaling method was applied to eliminate systematic errors caused by the differences in spatial resolution between the predicted gridded data and point-scale station measurements. As depicted in Fig. 8, a noteworthy systematic deviation is evident in the daily average SSM time series obtained from the automatic observation stations in China, in contrast to the corresponding daily average time series of SSM derived from direct predictions by different models at collocated locations. The study employed the CDF matching method for this purpose. First, the CDF was computed for the predicted gridded SSM data obtained from the three retrieval models, as well as for the in-situ measurements. Both sets of data were calculated within the same bins. Then, a fifth-order polynomial was employed to fit CDF values of the gridded SSM and the differences between the CDF of gridded data and in-site measurements. This polynomial was then applied as a transformation on the satellite data, effectively establishing a mapping between the gridded SSM values and their corresponding rescaled counterparts.

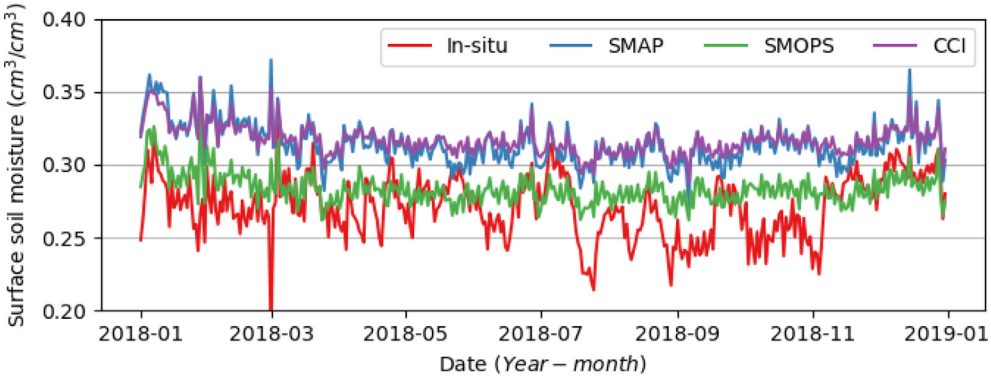

Fig. 8. Daily mean surface soil moisture time series of Chinese automated soil moisture observation stations and collocated gridded soil moisture from SMAP-based model, SMOPS-based model, and CCI-based model.

As shown in Fig. 9, the transformed satellite data exhibited and identical distribution to the in-situ measurements. This outcome confirms the suitability of the chosen polynomial degree for fitting. The results show that the CDF matching method is a robust approach to overcoming the issue of the spatial resolution difference between the sparse gridded data and the in-situ measurement.

Spatiotemporal collocation was conducted on rescaled SSM data derived from the model outputs and in-situ measurements within southern China, serving as the basis for regional evaluation. Three sets of predicted gridded SSM data were collocated with station measurements within in the same grid pixel, yielding around 130,000 data pairs for each set. These data pairs were depicted in a scatter density plot, as shown in Fig. 10, offering a visual representation of the distribution of both predicted and measured SSM values. The scatter density plots unveiled a positive correlation between all three model-predicted outcomes and the in-situ measurements, with the majority of matching points clustered on the 1:1 diagonal. For each set of data pairs, the correlation coefficient was computed, furnishing a quantifiable gauge of the association between the predicted and measured values. The calculated R-values for the SMAP, SMOPS, and CCI-based model estimations were 0.346, 0.384, and 0.362, respectively. These results suggest that the SMOPS-based model predictions exhibit the strongest correlation with the in-situ measurements, closely followed by the CCI-based model. Furthermore, the MAE and RMSD values were calculated for each model, offering an overview of the overall difference between the predicted and measured SSM. The calculated MAEs for the SMAP, SMOPS, and CCI models were 0.084, 0.082 and 0.083 cm^3/cm^3 respectively. Correspondingly, the RMSDs for the SMAP, SMOPS, and CCI models were 0.107, 0.104 and 0.106 cm^3/cm^3 respectively. The results suggest that the SMOPS-based model exhibit the smallest overall disparities between predicted and in situ measurements. To summarize, the results strongly imply that the SMOPS model outperformed the others in SSM prediction, boasting the highest correlation coefficient along with the lowest MAE and RMSD values. The CCI model also performed commendably, although not at par with the top performing SMOPS model, while the SMAP model displayed the least favorable performance.

Spatial performance evaluation of the models was conducted by using in-situ measurements aligned within the same grid pixels, as depicted in Fig. 11. The RMSD map portrays the spatial distribution of the RMSD across southern China, highlighting prevalent high RMSD values across numerous regions and a generally low correlation coefficient. This implies that the models do not effectively predict SSM in these regions. Nevertheless, it's crucial to acknowledge that this result doesn't solely stem from the CYGNSS observation and retrieval technique. Upon comparing reference data with the station measurements, it became evident that the performance of reference data in southern China mirrored the same pattern. These findings indicate that further improvements are necessary in the CYGNSS SSM retrieval model, specifically using high-quality reference data in China.

Moreover, the temporal performance of the model-predicted gridded SSM compared with in relation to in-situ measurements is portrayed in Fig. 7 (b). The daily R-values and errors, presented as scatter density plots, tended to be consistent across the three sets of model predictions against ground station measurements. However, a higher degree of variability was observed in comparison to the outcomes between the model predictions and the reference gridded values. These fluctuations in correlation and error could potentially be influenced by variations in sample size across different time periods.

To further validate the performance of the retrieval model, a scatter density plot of collocated in-situ measurements and SSM products from SMAP, SMOPS, and CCI under the EASE-Grid2 grid is presented in Fig. 12. The scatter density plots demonstrate that the three reference datasets yielded outcomes closely resembling the predictions of the model generated gridded SSM. Among the ref-

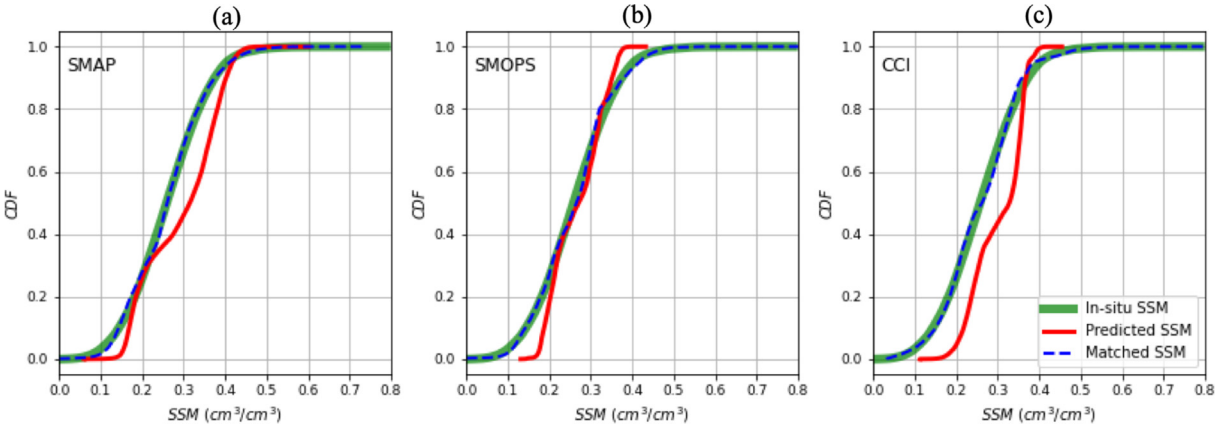

Fig. 9. Cumulative Distribution Function (CDF) matching results of the predicted soil surface moisture from the test dataset using a retrieval model, between SMAP-based model (a), SMOPS-based model (b), and CCI-based model (c), compared to spatial and in-situ measurements.

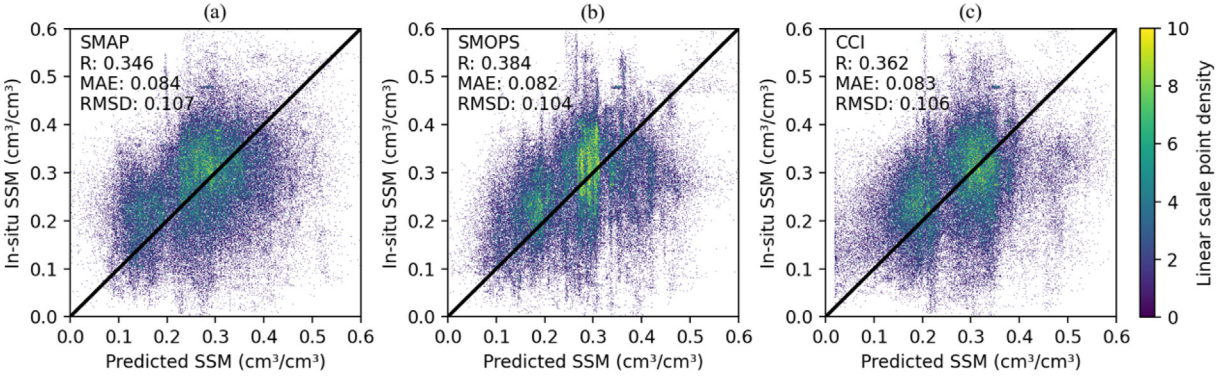

Fig. 10. Scatter density plot illustrating the relationship between rescaled soil moisture values from SMAP-based model output (a), SMOPS-based model output (b), and CCI-based model output (c), and in-situ measurement data pairs.

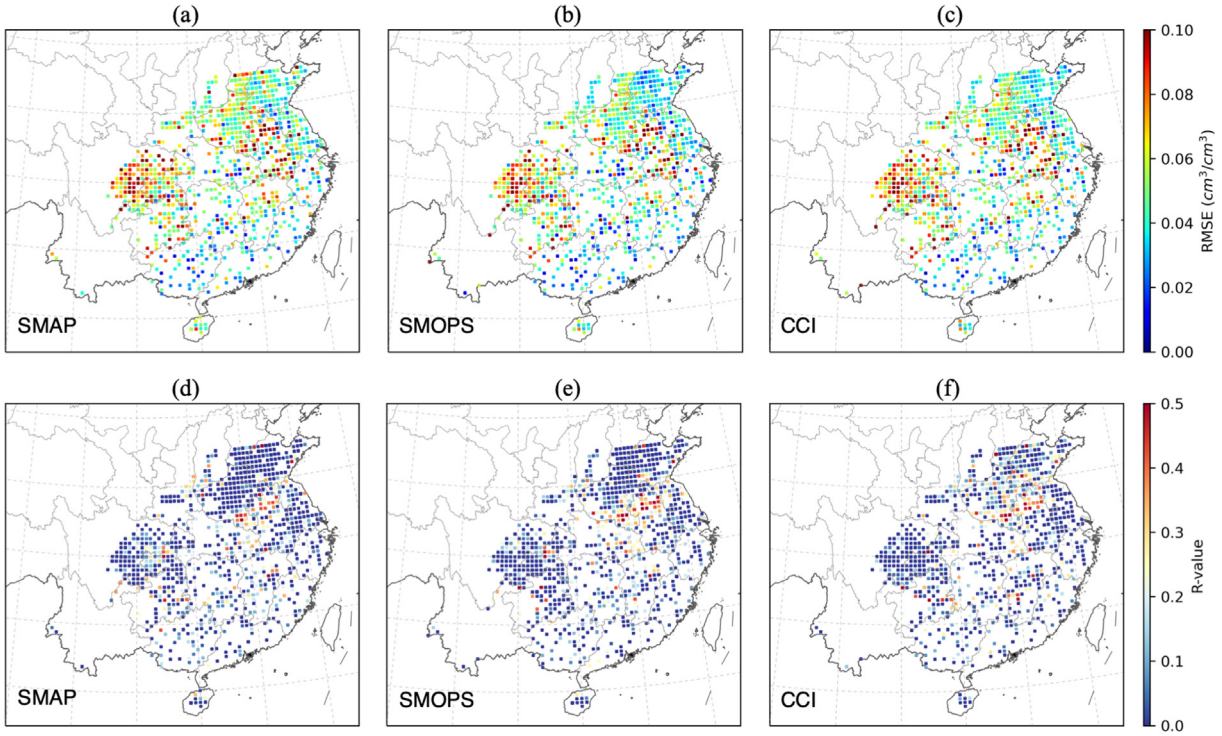

Fig. 11. Comparison of the RMSD and correlation coefficient of three models based on SMAP (a, d), SMOPS (b, e), and CCI (c, f), evaluated using collocated in-situ measurements.

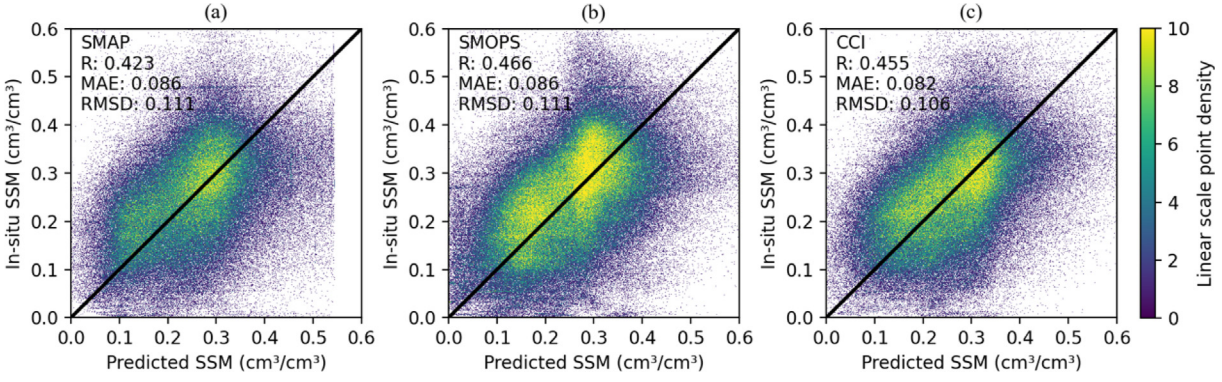

Fig. 12. Scatter density plot illustrating the relationship between rescaled soil moisture values from SMAP (a), SMOPS (b), and CCI (c) products, and in-situ measurement data pairs.

erence datasets, SMOPS exhibited the highest correlation coefficient along with the lowest MAE and RMSD values when compared to the in-situ measurements. The R-value for SMAP, SMOPS, and CCI data were 0.423, 0.466, and 0.455, respectively. While the RMSD values were slightly higher overall when compared to the respective model-predicted data, this discrepancy was primarily attributed to the considerably larger number of matched data pairs. This is due to the fact that the blended products were not subjected to the same spatial coverage screening as the SMAP data. Nonetheless, the statistics are sufficient to demonstrate that the SSM retrieved by the CYGNSS-retrieved SSM can reach levels comparable to those of traditional remote sensing techniques. Furthermore, the statistics also highlight the impact of the accuracy of the reference data on the modeling process.

3.4. Model performance using original resolution blended soil moisture products

As demonstrated by the modeling and prediction results, incorporating high-accuracy blended SSM data can enhance the accuracy of CYGNSS SSM retrieval model. Furthermore, the blended SSM product also offers broader spatial and temporal coverage, which effectively reduces the daily space–time gaps through data fusion in comparison with SSM products from SMAP or SMOS. By regridding the CYGNSS-derived effective reflectivity into a 0.25° × 0.25° grid, which is used by both blended products, the full potential of the blended SSM product can be utilized for modeling purposes. The 9 km × 9 km EASE-Grid2 raw vegetation optical depth data were resampled using nearest neighbor interpolation, upscaling it to a 0.25° × 0.25° regular grid for CYGNSS SSM retrieval vegetation attenuation correction.

The average model performances using SMOPS and CCI over 12 iterations are presented in Table 2. The evaluation demonstrated that the SMOPS model outperformed the CCI model in terms of bias, MAE, RMSD, and ubRMSD. Specifically, the SMOPS model had a bias of 0.0002 cm^3/cm^3, MAE of 0.020 cm^3/cm^3, RMSD of 0.028 cm^3/cm^3, and ubRMSD of 0.028 cm^3/cm^3, while the CCI model had a bias of 0.0 cm^3/cm^3, MAE of 0.031 cm^3/cm^3, RMSD of 0.042 cm^3/cm^3, and ubRMSD of 0.042 cm^3/cm^3. Furthermore, the SMOPS model exhibited a superior R-value of 0.930 compared with 0.906 for the CCI model. Moreover, all skill metrics displayed very low standard deviations, indicating that the models were stable. Compared to outcomes achieved through the 36 km × 36 km EASE-Grid2 grid retrieval model, the model established using the original resolution blended product demonstrated enhanced accuracy across all skill metrics. Overall, the SMOPS model outperformed the CCI model for both the spatial references. And the accuracy of the model was further enhanced under the original higher spatial resolution of the reference data. The scatter density plots of the model predictions from each fold cross-validation test dataset and reference data over grid points were very close to those in Fig. 5.

Following the CDF matching process of the SSMs predicted in each fold of the model cross-validation with the in-situ measurements, and subsequent spatiotemporal alignment with the station measurements, the scatter density plot illustrating the data pairs resulting from this matching is presented in Fig. 13. The original SMOPS-based and CCI-based models exhibit comparable prediction accuracies, evident through their similar RMSD values of 0.113 cm^3/cm^3 and 0.108 cm^3/cm^3, respectively. Given the consistency of the spatiotemporal performance with the EASE-Grid2 grid, no further analysis was pursued in this regard.

4. Discussion

This study highlights the potential enhancement of using a blended SSM product as the ground reference for CYGNSS SSM retrieval modeling. In Fig. 14, a comparison is presented regarding the spatial performance of the predicted SSM from the retrieval models and the blended SMOPS and CCI products. The results, averaged for January 2018, illustrate that the retrieval models produced SSM estimates that exhibited a stronger alignment with reference data. While deviations were generally modest across the majority of regions, notable distinctions in performance between the SMOPS- and CCI-based models were observed, particularly in the Sudanian Savanna, where SMOPS outperformed CCI. The consistently low bias levels observed in all three models, fluctuating around 0.0 cm^3/cm^3, underscore the accuracy of the developed models. Furthermore, the daily time series of RMSD and ubRMSD for the three models exhibited considerable similarity, further accentuating their accuracy. The SMOPS-based model had the smallest MAE, RMSD, and ubRMSD metrics, and remained stable throughout the year. The CCI and SMAP performed similarly, with the correlation coefficient showing a declining trend between July and October, suggesting a potential impact of seasonal variation on their performance. In summary, the results show that the SMOPS-based model is the superior model

Table 2
The average skill metrics and standard deviation of 12-fold cross-validation of raw blended product formed models (cm^3/cm^3).

Skill Metrics	Bias (STD)	MAE (STD)	RMSD (STD)	R (STD)	ubRMSD (STD)
SMOPS	0.0002 (0.002)	0.020 (0.001)	0.028 (0.001)	0.930 (0.007)	0.028 (0.002)
CCI	0.0000 (0.004)	0.031 (0.002)	0.042 (0.003)	0.906 (0.014)	0.042 (0.002)

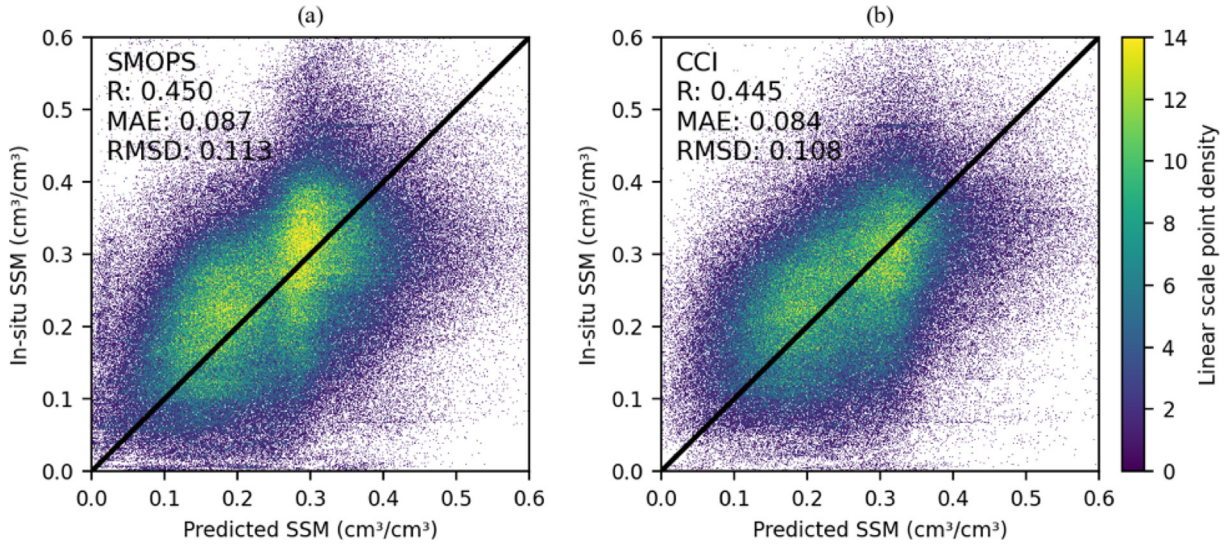

Fig. 13. Scatter density plot showing the relationship between rescaled soil moisture from SMOPS (b), and CCI-based (c) model prediction under raw spatial resolution and in-situ measurement data pairs.

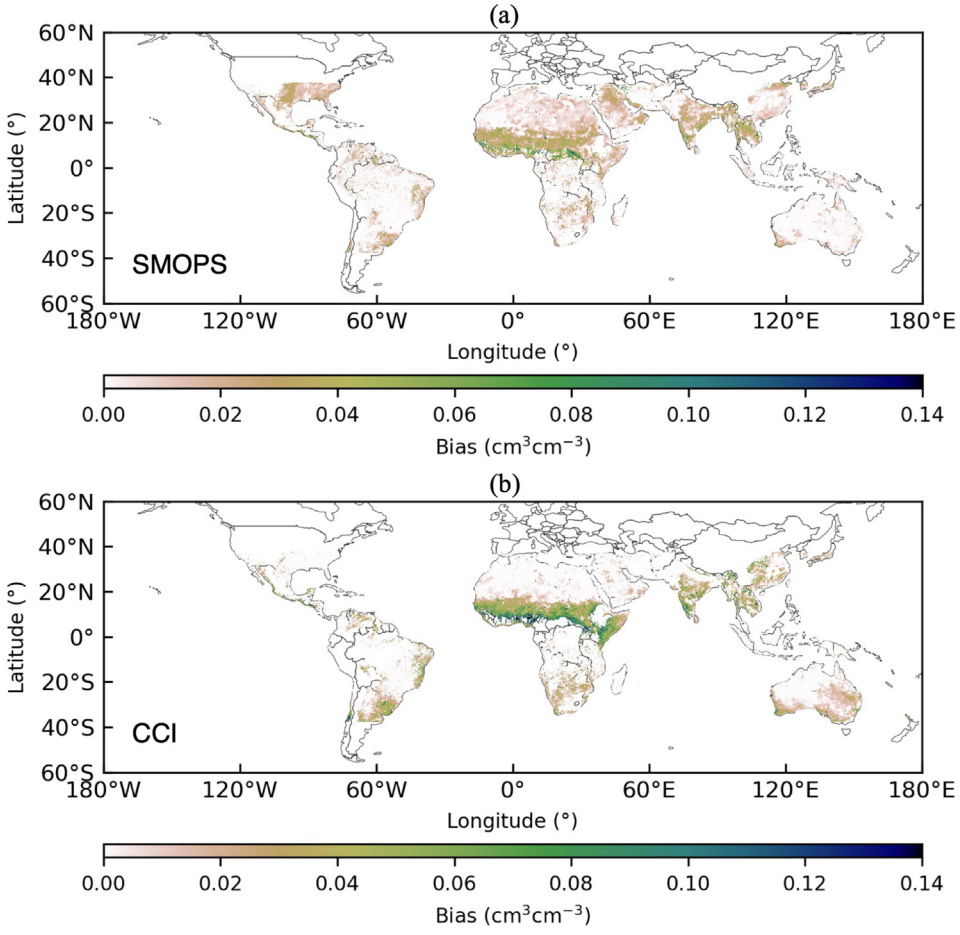

Fig. 14. The difference of monthly averaged SMOPS and CCI based model predicted surface soil moisture and corresponding reference data for the month of January 2018 with 0.25° × 0.25° spatial resolution.

among the three, whereas CCI and SMAP have comparable performances.

In certain regions, the accuracy of the CYGNSS-derived SSM is compromised due to various factors. Despite ongoing debates regarding the precise spatial resolution of CYGNSS observations, consensus suggests that it surpasses the resolutions of the two reference grids used in this study. The existence of heterogeneities within the grid pixels, such as disparities in topography, vegetation, and even small bodies of water, can cause significant representative errors. Previous studies have indicated that implementing roughness correction utilizing roughness coefficients from the SMAP product did not yield a significant impact on CYGNSS-derived SSMs. This outcome might stem from many factors, including the accuracy of the correction model, noise inherent in CYGNSS effective reflectivity, data aggregation during regridding, and spatial and temporal variability present within grid pixels. These factors collectively contribute to diminishing the efficacy of surface attenuation correction. However, in the theoretical simulation study, it is evident that surface roughness exerts substantial influence on the GNSS scattering signal (Balakhder et al., 2019); therefore, the surface small-scale roughness needs to be considered in the subsequent spaceborne GNSS-R SSM inversion focus.

When assessed against measurements from Chinese automated soil moisture observation stations, the models exhibited a slightly diminished performance in comparison to prior modeling and assessment studies that relied on measurements from the International Soil Moisture Network (ISMN) sites (Senyurek et al., 2020). The comparison between the model predictions, raw reference product, and in-situ measurements unveiled that the dissimilarity was primarily due to inadequate calibration of the original reference data within the Chinese region. The spatial distribution of the correlation coefficients, calculated from the gridded CYGNSS-derived daily effective reflectivity and spatiotemporally matched station measurements in each grid pixel, is displayed in Fig. 15. The CYGNSS-derived SSMs were calibrated linearly from the CYGNSS-derived effective reflectivity. The correlation coefficients exhibited a more favorable distribution compared to the outcomes presented in Fig. 11, providing additional validation to the aforementioned observation. Furthermore, the suboptimal model performance can also be attributed to dissimilarities in station selection. While previous studies meticulously chose ISMN stations, our approach encompassed all stations within the CYGNSS mission coverage. The geographical features of the wetter southern region of China, characterized by a larger distribution of vegetation, developed water systems, and urbanization, as depicted in the natural surface map of the Chinese region in Fig. 1, significantly impact the ultimate model performance as well.

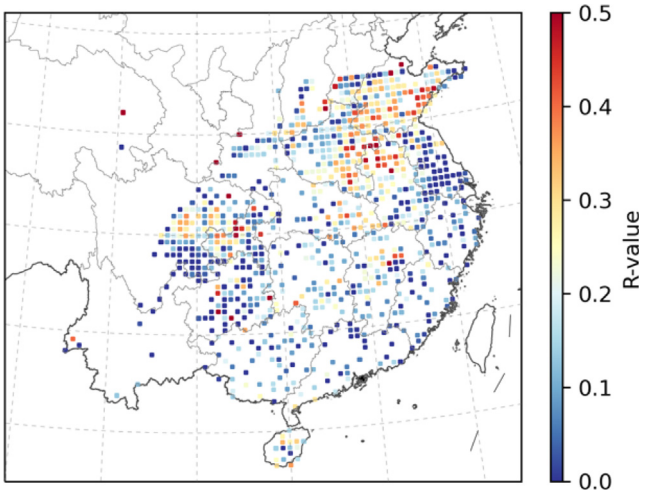

Fig. 15. Correlation coefficient between CYGNSS-derived gridded effective reflectivity and collocated in-situ measurements.

Constructing the retrieval models at the native spatial resolution of the blended products led to enhanced performance compared to when the reference data were resampled to the 36 km × 36 km EASE-Grid2 grid. Nevertheless, the evaluation results based on station measurements revealed a decline in accuracy, potentially influenced by the correction for vegetation attenuation. When the retrieval model was constructed using the full reference data without any correction for vegetation attenuation, the results of the cross-validation were in line with the findings presented in Table 2. This indicates that vegetation attenuation can be disregarded as noise in the modeling process when a larger amount of training data is available. Previous studies have showcased the rigorous quality control procedures applied to CCI data, leading to SSM products of superior accuracy compared to SMOPS. However, in this case, the model's performance using the SMOPS reference data was superior. This underscores the need to carefully deliberate on the trade-offs between the quality control rigor of the CCI product and the extended spatial and temporal coverage provided by the SMOPS product. Larger training samples generally lead to improved performance by supplying the model with a broader array of diverse and representative instances for learning. While low uncertainty and small samples might contain fewer outliers or errors that could negatively affect the performance of the retrieval model, they may not adequately capture the underlying patterns and relationships intrinsic to the data.

Space-based GNSS-R remote sensing has the unique advantage of high spatial and temporal resolutions, making it promising for obtaining high-resolution SSM datasets and assimilating them into other merged high spatial and temporal resolution products. Although the CYGNSS mission was not primarily designed for land remote sensing applications and the spatial coverage of the inversion results was influenced by the quality of the CYGNSS observations and its quality control strategy, the use of blended SSM products in this study as the ground reference true values for inversion modeling can further improve the

accuracy of the model, and the predicted SSM exhibited a comparable performance to the reference data. This study can motivate the continued development of dedicated satellite missions for GNSS-R land SSM retrieval, which will greatly improve the spatial and temporal coverage of SSM measurements.

5. Conclusion

This study examined CYGNSS-based SSM retrieval models using three reference datasets: SMAP, SMOPS, and CCI. Model performance was evaluated using in situ measurements resampled on a daily scale. The study aims to identify the optimal reference dataset for the SSM retrieval algorithm by comparing their effects on the model. Employing a CDF matching method, the study eliminates systematic errors arising from spatial resolution differences between the gridded CYGNSS-derived SSM predicted by the model and in situ measurements. To evaluate the retrieval accuracy when utilizing the blended SSM product as the reference value in the modeling, the study employed the blended SSM at its original spatial resolution (0.25° × 0.25° regular grid) and also at a reduced spatial resolution (36 km × 36 km EASE-Grid2 grid).

The results indicate that the model predicted SSM effectively captured the spatial variation at both spatial resolutions. The SMAP-based model exhibited the least favorable performance, characterized by elevated RMSD values across multiple regions and a comparatively low correlation coefficient with the EASE-Grid2 grid. The study revealed that the retrieval models exhibited higher accuracy when constructed using the native spatial resolution of the blended products, as opposed to when the reference data were resampled to the lower 36 km × 36 km EASE-Grid2 grid. Among the models tested, the SMOPS-based model demonstrated the strongest correlation coefficient (0.930) and smallest RMSD (0.028 cm^3/cm^3), followed by the CCI-based model (R = 0.906 and RMSD = 0.042 c m^3/cm^3). Upon evaluating the constructed inversion models against in-situ measurements, it was observed that all three model predictions exhibited a positive correlation with the measurements. Nevertheless, when assessed using data from Chinese automated soil moisture observation stations, the models displayed slightly diminished performance in comparison to previous studies that relied on ISMN site measurements. This difference was primarily due to the insufficient calibration of the original reference data in the Chinese region. In summary, the most precise SSM retrieval model was established using the blended SMOPS SSM product. While the outcomes of the CYGNSS mission underscore its substantial potential, it is imperative to pursue further enhancements, particularly by incorporating high-quality reference data from China, to attain heightened accuracy and reliability in results.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grand number 42204014). The authors would like to acknowledge the CYGNSS team and other institutions for providing the data set used in this study.

References

Al-Khaldi, M.M., Johnson, J.T., O'Brien, A.J., Balenzano, A., Mattia, F., 2019. Time-Series Retrieval of Soil Moisture Using CYGNSS. IEEE Trans. Geosci. Remote Sensing 57, 4322–4331. https://doi.org/10.1109/TGRS.2018.2890646.

Balakhder, A.M., Al-Khaldi, M.M., Johnson, J.T., 2019. On the coherency of ocean and land surface specular scattering for GNSS-R and signals of opportunity systems. IEEE Trans. Geosci. Remote Sensing 57, 10426–10436. https://doi.org/10.1109/TGRS.2019.2935257.

Brocca, L., Hasenauer, S., Lacava, T., Melone, F., Moramarco, T., Wagner, W., Dorigo, W., Matgen, P., Martínez-Fernández, J., Llorens, P., Latron, J., Martin, C., Bittelli, M., 2011. Soil moisture estimation through ASCAT and AMSR-E sensors: An intercomparison and validation study across Europe. Remote Sens. Environ. 115, 3390–3408. https://doi.org/10.1016/j.rse.2011.08.003.

Camps, A., Park, H., Pablos, M., Foti, G., Gommenginger, C.P., Liu, P.-W., Judge, J., 2016. Sensitivity of GNSS-R Spaceborne Observations to Soil Moisture and Vegetation. IEEE J. Sel. Top. Appl. Earth Observations Remote Sens. 9, 4730–4742. https://doi.org/10.1109/JSTARS.2016.2588467.

Chew, C.C., Small, E.E., 2018. Soil moisture sensing using spaceborne GNSS reflections: Comparison of CYGNSS reflectivity to SMAP soil moisture. Geophys. Res. Lett. 45, 4049–4057. https://doi.org/10.1029/2018GL077905.

Chew, C., Small, E., 2020. Description of the UCAR/CU Soil Moisture Product. Remote Sens. (Basel) 12, 1558. https://doi.org/10.3390/rs12101558.

Clarizia, M.P., Ruf, C.S., Jales, P., Gommenginger, C., 2014. Spaceborne GNSS-R minimum variance wind speed estimator. IEEE Trans. Geosci. Remote Sens. 52, 6829–6843. https://doi.org/10.1109/TGRS.2014.2303831.

Clarizia, M.P., Pierdicca, N., Costantini, F., Floury, N., 2019. Analysis of CYGNSS data for soil moisture retrieval. Sel. Top. Appl. Earth Observations Remote Sens. 12, 2227–2235. https://doi.org/10.1109/JSTARS.2019.2895510.

Conil, S., Douville, H., Tyteca, S., 2006. The relative influence of soil moisture and SST in climate predictability explored within ensembles of AMIP type experiments. Clim. Dyn. 28, 125–145. https://doi.org/10.1007/s00382-006-0172-2.

Dong, Z., Jin, S., 2021. Evaluation of the land GNSS-reflected DDM coherence on soil moisture estimation from CYGNSS data. Remote Sens. (Basel) 13, 570. https://doi.org/10.3390/rs13040570.

Dorigo, W.A., Gruber, A., De Jeu, R.A.M., Wagner, W., Stacke, T., Loew, A., Albergel, C., Brocca, L., Chung, D., Parinussa, R.M., Kidd, R., 2015. Evaluation of the ESA CCI soil moisture product using

ground-based observations. Remote Sens. Environ. 162, 380–395. https://doi.org/10.1016/j.rse.2014.07.023.

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., Haas, E., Hamer, P. D., Hirschi, M., Ikonen, J., de Jeu, R., Kidd, R., Lahoz, W., Liu, Y.Y., Miralles, D., Mistelbauer, T., Nicolai-Shaw, N., Parinussa, R., Pratola, C., Reimer, C., van der Schalie, R., Seneviratne, S.I., Smolander, T., Lecomte, P., 2017. ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions. Remote Sens. Environ. 203, 185–215. https://doi.org/10.1016/j.rse.2017.07.001.

Entekhabi, D., Njoku, E.G., O'Neill, P.E., Kellogg, K.H., Crow, W.T., Edelstein, W.N., Entin, J.K., Goodman, S.D., Jackson, T.J., Johnson, J., Kimball, J., Piepmeier, J.R., Koster, R.D., Martin, N., McDonald, K.C., Moghaddam, M., Moran, S., Reichle, R., Shi, J.C., Spencer, M. W., Thurman, S.W., Tsang, L., Van Zyl, J., 2010. The Soil Moisture Active Passive (SMAP) mission. Proc. IEEE 98, 704–716. https://doi.org/10.1109/JPROC.2010.2043918.

Eroglu, O., Kurum, M., Boyd, D., Gurbuz, A.C., 2019. High spatio-temporal resolution CYGNSS soil moisture estimates using artificial neural networks. Remote Sens. (Basel) 11, 2272. https://doi.org/10.3390/rs11192272.

Foti, G., Gommenginger, C., Jales, P., Unwin, M., Shaw, A., Robertson, C., Roselló, J., 2015. Spaceborne GNSS reflectometry for ocean winds: First results from the UK TechDemoSat-1 mission: SPACEBORNE GNSS-R FOR OCEAN WINDS: FIRST TDS-1 RESULTS. Geophys. Res. Lett. 42, 5435–5441. https://doi.org/10.1002/2015GL064204.

Gruber, A., Scanlon, T., van der Schalie, R., Wagner, W., Dorigo, W., 2019. Evolution of the ESA CCI Soil Moisture climate data records and their underlying merging methodology. Earth Syst. Sci. Data 11, 717–739. https://doi.org/10.5194/essd-11-717-2019.

Gruber, A., De Lannoy, G., Albergel, C., Al-Yaari, A., Brocca, L., Calvet, J.-C., Colliander, A., Cosh, M., Crow, W., Dorigo, W., Draper, C., Hirschi, M., Kerr, Y., Konings, A., Lahoz, W., McColl, K., Montzka, C., Muñoz-Sabater, J., Peng, J., Reichle, R., Richaume, P., Rüdiger, C., Scanlon, T., van der Schalie, R., Wigneron, J.-P., Wagner, W., 2020. Validation practices for satellite soil moisture retrievals: What are (the) errors? Remote Sens. Environ. 244. https://doi.org/10.1016/j.rse.2020.111806 111806.

Hasan, S., Montzka, C., Rüdiger, C., Ali, M., R. Bogena, H., Vereecken, H., 2014. Soil moisture retrieval from airborne L-band passive microwave using high resolution multispectral data. ISPRS Journal of Photogrammetry and Remote Sensing 91, 59–71. https://doi.org/10.1016/j.isprsjprs.2014.02.005.

Jia, Y., Jin, S., Chen, H., Yan, Q., Savi, P., Jin, Y., Yuan, Y., 2021. Temporal-spatial soil moisture estimation from CYGNSS using machine learning regression with a preclassification approach. IEEE J. Sel. Top. Appl. Earth Observations Remote Sens. 14, 4879–4893. https://doi.org/10.1109/JSTARS.2021.3076470.

Kerr, Y.H., Waldteufel, P., Wigneron, J.-P., Delwart, S., Cabot, F., Boutin, J., Escorihuela, M.-J., Font, J., Reul, N., Gruhier, C., Juglea, S.E., Drinkwater, M.R., Hahne, A., Martín-Neira, M., Mecklenburg, S., 2010. The SMOS mission: New tool for monitoring key elements ofthe global water cycle. Proc. IEEE 98, 666–687. https://doi.org/10.1109/JPROC.2010.2043032.

Konings, A.G., Piles, M., Das, N., Entekhabi, D., 2017. L-band vegetation optical depth and effective scattering albedo estimation from SMAP. Remote Sens. Environ. 198, 460–470. https://doi.org/10.1016/j.rse.2017.06.037.

Lei, F., Senyurek, V., Kurum, M., Gurbuz, A.C., Boyd, D., Moorhead, R., Crow, W.T., Eroglu, O., 2022. Quasi-global machine learning-based soil moisture estimates at high spatio-temporal scales using CYGNSS and SMAP observations. Remote Sens. Environ. 276. https://doi.org/10.1016/j.rse.2022.113041 113041.

Liu, J., Zhan, X., Hain, C., Yin, J., Fang, L., Li, Z., Zhao, L., 2016. In: NOAA Soil Moisture Operational Product System (SMOPS) and Its Validations. IEEE, Beijing, China, pp. 3477–3480. https://doi.org/10.1109/IGARSS.2016.7729899.

Ma, H., Zeng, J., Chen, N., Zhang, X., Cosh, M.H., Wang, W., 2019. Satellite surface soil moisture from SMAP, SMOS, AMSR2 and ESA CCI: A comprehensive assessment using global ground-based observations. Remote Sens. Environ. 231. https://doi.org/10.1016/j.rse.2019.111215 111215.

Nabi, M.M., Senyurek, V., Gurbuz, A.C., Kurum, M., 2022. Deep learning-based soil moisture retrieval in CONUS using CYGNSS delay-doppler maps. IEEE J. Sel. Top. Appl. Earth Observations Remote Sens. 15, 6867–6881. https://doi.org/10.1109/JSTARS.2022.3196658.

ONeill, Peggy E., Chan, Steven, Njoku, Eni G, Jackson, Tom, Bindlish, Rajat, Chaubell, M. Julian, 2021. SMAP L3 Radiometer Global Daily 36 km EASE-Grid Soil Moisture, Version 8. https://doi.org/10.5067/OMHVSRGFX38O.

Rose, R., Gleason, S., Ruf, C., 2014. The NASA CYGNSS mission: a pathfinder for GNSS scatterometry remote sensing applications, in: Bostater, C.R., Mertikas, S.P., Neyt, X. (Eds.). Presented at the SPIE Remote Sensing, Amsterdam, Netherlands, p. 924005. https://doi.org/10.1117/12.2068378.

Saeedi, M., Sharafati, A., Tavakol, A., 2021. Evaluation of gridded soil moisture products over varied land covers, climates, and soil textures using in situ measurements: A case study of Lake Urmia Basin. Theor. Appl. Climatol. 145, 1053–1074. https://doi.org/10.1007/s00704-021-03678-x.

Senyurek, V., Lei, F., Boyd, D., Gurbuz, A.C., Kurum, M., Moorhead, R., 2020. Evaluations of Machine Learning-Based CYGNSS Soil Moisture Estimates against SMAP Observations. Remote Sens. (Basel) 12, 3503. https://doi.org/10.3390/rs12213503.

Senyurek, V., Lei, F., Boyd, D., Kurum, M., Gurbuz, A.C., Moorhead, R., 2020. Machine learning-based CYGNSS soil moisture estimates over ISMN sites in CONUS. Remote Sens. (Basel) 12, 1168. https://doi.org/10.3390/rs12071168.

Tsegaye, T.D., Tadesse, W., Coleman, T.L., Jackson, T.J., Tewolde, H., 2004. Calibration and modification of impedance probe for near surface soil moisture measurements. Can. J. Soil Sci. 84, 237–243. https://doi.org/10.4141/S03-069.

Wang, Y., Leng, P., Peng, J., Marzahn, P., Ludwig, R., 2021. Global assessments of two blended microwave soil moisture products CCI and SMOPS with in-situ measurements and reanalysis data. Int. J. Appl. Earth Obs. Geoinf. 94. https://doi.org/10.1016/j.jag.2020.102234 102234.

Wang, L., Qu, J.J., 2009. Satellite remote sensing applications for surface soil moisture monitoring: A review. Front Earth Sci. China 3, 237–247. https://doi.org/10.1007/s11707-009-0023-7.

Yan, Q., Huang, W., 2016. Spaceborne GNSS-R sea ice detection using delay-doppler maps: First results from the U.K. TechDemoSat-1 mission. IEEE J. Sel. Top. Appl. Earth Observations Remote Sens. 9, 4795–4801. https://doi.org/10.1109/JSTARS.2016.2582690.

Yan, Q., Huang, W., Jin, S., Jia, Y., 2020. Pan-tropical soil moisture mapping based on a three-layer model from CYGNSS GNSS-R data. Remote Sens. Environ. 247. https://doi.org/10.1016/j.rse.2020.111944 111944.

Yueh, S.H., Shah, R., Chaubell, M.J., Hayashi, A., Xu, X., Colliander, A., 2022. A semiempirical modeling of soil moisture, vegetation, and surface roughness impact on CYGNSS reflectometry data. IEEE Trans. Geosci. Remote Sens. 60, 1–17. https://doi.org/10.1109/TGRS.2020.3035989.

Zavorotny, V.U., Gleason, S., Cardellach, E., Camps, A., 2014. Tutorial on remote sensing using GNSS bistatic radar of opportunity. IEEE Geosci. Remote Sens. Mag. 2, 8–45. https://doi.org/10.1109/MGRS.2014.2374220.