Measurement 242 (2025) 115849

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Refined semi-empirical models for soil moisture retrieval in spaceborne GNSS-Reflectometry: Evaluation across diverse land cover types

Zhounan Dong [a,b], Qingyun Yan [c,*], Shuanggen Jin [d], Li Li [a,b], Guodong Chen [a,b]

[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 Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China
[d] School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454000, China

ARTICLE INFO

Keywords:
GNSS-R
Soil moisture
Land cover type
Vegetation attenuation
Roughness

ABSTRACT

The satellite-based Global Navigation Satellite System-Reflectometry (GNSS-R) has emerged as a promising technique for detecting surface soil moisture (SSM), which plays a pivotal role in various applications. Different approaches have been developed for SSM retrieval, including parametric semi-empirical models and non-parametric machine learning methods. This study, however, specifically focuses on constructing and evaluating semi-empirical fit models. To this end, the study compares the effectiveness of our enhanced pixel-by-pixel model, the land cover-based linear model, and the Reflectivity-Vegetation-Roughness (R-V-R) model across different land surface types, aiming to both evaluate their efficacy and identify potential limitations. In the assessment, various factors were taken into consideration, such as the correction for vegetation and roughness attenuation, fitting functions employed, and the utilization of a lookup table (LUT). The results of the evaluation showed variations in the performance of the retrieval models across different land cover types, highlighting the impact of the choice of fitting functions and attenuation correction strategies on the accuracy of soil moisture retrieval. The pixel-by-pixel model demonstrated the highest prediction accuracy, with an unbiased root mean square difference (ubRMSD) of 0.056 cm^3/cm^3 and a correlation coefficient of 0.896. By showcasing these outcomes, this research underscores the significance of accounting for surface conditions and integrating relevant data to enhance the accuracy of GNSS-R SSM retrieval, thereby contributing to the advancement of SSM monitoring methodologies.

1. Introduction

Soil moisture information is essential for diverse applications, such as water resource management, agricultural productivity enhancement, weather forecasting, and ecological health monitoring, underscoring the need for accurate and reliable remote sensing retrieval methods [1]. Conventional in situ measurements conducted on the ground are capable of delivering accurate soil moisture data only at a localized point, thereby generating sparse data that is insufficient for regional or global-scale applications [2]. On the other hand, satellite remote sensing has progressed over the last two decades and is a more efficient method of detecting surface soil moisture (SSM) at global scales. However, the limitations of early-stage shortwave-based optical remote sensing and thermal infrared remote sensing based on surface latent heat effects make it difficult to provide accurate and quantitative SSM products with high spatial and temporal resolution from individual sensors [3,4]. The successful estimation of SSM has been achieved through the subsequent advancement of microwave remote sensing operating at lower frequencies, enabling the monitoring of the dielectric constant of the near-surface soil layer, which is directly correlated with water content [5–7]. Furthermore, microwave active and passive sensors offer the advantage of continuous, all-weather operation, with 24/7 data collection capabilities, making them particularly well-suited for persistent SSM monitoring. As satellite antenna technology develops, specialized L-band soil moisture monitoring satellite missions such as Soil Moisture and Ocean Salinity (SMOS) [8] and Soil Moisture Active Passive (SMAP) [9] have been successfully launched. These satellites provide unprecedented global datasets, significantly impacting various fields. However, the

* Corresponding author at: School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China.
E-mail address: 003257@nuist.edu.cn (Q. Yan).

https://doi.org/10.1016/j.measurement.2024.115849
Received 29 February 2024; Received in revised form 19 August 2024; Accepted 29 September 2024
Available online 30 September 2024

rotation difference between satellites and Earth can result in strip-gap regions, necessitating gap-filling methods to update products in these areas [10].

In the past decade, spaceborne Global Navigation Satellite System-Reflectometry (GNSS-R) has exhibited considerable potential in remotely sensing geophysical surface parameters, thereby highlighting its capacity as an innovative data source for monitoring SSM [11–14]. This technique leverages the reflected GNSS signals from the Earth's surface to provide a cost-effective alternative to the deployment of satellite-based GNSS-R observation systems. The GNSS-R receiver processes the reflected GNSS signal into a 2-Dimensional Delay Doppler Map (DDM), which serves as a fundamental observation for spaceborne GNSS-R missions. Compared to dedicated radiometer missions, this approach offers enhanced efficiency, cost-effectiveness, and high spatiotemporal resolution, making it a desirable supplement to conventional SSM remote sensing techniques. By integrating the spaceborne GNSS-R technique, the capability and performance of SSM retrieval in earth science can be enhanced, providing new opportunities for improved understanding of SSM dynamics.

The Cyclone Global Navigation Satellite System (CYGNSS), launched by the National Aeronautics and Space Administration (NASA), is the world's first GNSS-R micro-satellite constellation initially designed for ocean surface wind speed monitoring [15]. It provides high-spatial-resolution land surface observations to compensate for the constraints associated with passive radiometers. Additionally, it offers high temporal resolution, overcoming the temporal gaps inherent in the synthetic aperture radar (SAR) remote sensing of SSM. This combination of capabilities is crucial for mitigating the individual spatial and temporal limitations of the traditional passive and active SSM sensors. Extensive exploratory validation studies have consistently validated that satellite-based GNSS-R observations are capable of accurately detecting changes in SSM. Moreover, the effective reflectivity derived from GNSS-R measurements has been established as a reliable proxy for estimating SSM with high accuracy, as evidenced by these thorough validation efforts [16–19]. As a result, GNSS-R has emerged as a valuable tool for SSM estimation. The CYGNSS mission has provided significantly high-quality and high-frequency repeated observations, attracting immediate interest for SSM inversion research [20]. These data have spurred further research in SSM inversion techniques.

Many studies have investigated SSM inversion using the CYGNSS land observations. SSM products generated by the SMAP mission are used as a reference value and modeling basis in most of these studies. The linear correlation between the changes in CYGNSS effective reflectivity and the changes in SMAP soil moisture has been discovered and explained [21]. Furthermore, an empirical statistical model incorporating trilinear regression was established to characterize the relationship between gridded effective reflectivity, vegetation opacity, roughness coefficient, and SSM [22]. To date, inversion methods are broadly classified into parametric semi-empirical fitting models and non-parametric machine learning (ML) methods. Previous research has shown that both approaches demonstrate comparable inversion accuracies [14,20–26]. The parametric semi-empirical inversion model statistically establishes the relationship between effective reflectivity and the reference SSM from other observation systems or model outputs [21,22,27,28]. The presented accuracy of the retrieval is influenced by various spatiotemporal factors, including the study area and period, the availability of reference and auxiliary data, the methods employed for data preprocessing, and the strategies used to evaluate the inversion process. Consequently, the quality of the CYGNSS SSM inversion remains uncertain. Recently, ML and convolutional neural networks have paid attention for handling complex nonlinear relationships, leading to their widespread use in space-based GNSS-R SSM retrievals. These approaches utilize ML models trained on extracted or embedded features from the DDM, combined with collocated auxiliary parameters and reference SSM labels [26,29].

It is also necessary to consider the effects of other non-interesting influences on the effective reflectivity in the inverse modeling process. Spaceborne GNSS-R observations are also susceptible to confounding factors such as vegetation, roughness, and topography. Consequently, the development and incorporation of correction models and auxiliary data are essential for mitigate these biases [30–32]. Uncertainties in characterizing these parameters can be major sources of errors in satellite based SSM retrieval [33]. Calibration and correction of radiation parameters related to surface roughness and the canopy properties are particularly crucial for GNSS-R based SSM retrieval algorithms. Several inversion methods leverage change detection, assuming that the temporal scale of variability in vegetation and surface roughness far exceed those of SSM. Consequently, these inverse models often neglect the effects of vegetation and surface roughness [16,21,28]. Furthermore, previous studies have utilized radiometers-derived parameters that account for the impacts of vegetation and surface roughness. In the correction process, the GNSS-R effective reflectivity is adjusted through the calculation of vegetation layer transmissivity and surface undulation attenuation [22,23,34].

Land cover types exhibit varying physical and electrical properties, such as dielectric constant, which directly impact the reflection and scattering of GNSS signals. Moreover, the intricate geometry of vegetation structures, including leaf density and branch morphology, can also play a crucial role in modifying the GNSS signal propagation, thereby influencing the accuracy of GNSS-R-based SSM retrieval. Therefore, it is imperative that GNSS-R SSM retrieval algorithms integrate information on land cover types and assess their inversion performance across different land cover types. Although land cover type data has been successfully incorporated into ML and CNN inversion models as physically based features for GNSS-R SSM retrieval [29], it has not yet been integrated into the semi-empirical inversion model. Nonetheless, the semi-empirical model also reflects the influence of surface attributes in its approach. A clustering algorithm has been employed to categorize the global land surface into various types based on vegetation opacity and surface roughness characteristics. Linear retrieval models have been developed for each cluster, capitalizing on the similarities in land surface physical attributes across regions and assuming the temporal stability of vegetation and roughness to minimize their impacts [35]. This underscores the significance of considering specific surface conditions and integrating pertinent data for precise land cover classification in GNSS-R SSM retrieval.

The advancement of satellite-based GNSS-R terrestrial remote sensing has accelerated with the deployment of the CYGNSS mission, spurring the development of various soil moisture inversion algorithms. The upcoming HydroGNSS mission, featuring a new generation of GNSS-R sensors tailored for land applications, is set to launch in 2024, highlighting the growing interest in utilizing GNSS-R for SSM retrieval within the scientific community [36]. In this context, we aim to develop and evaluate semi-empirical fit models, with a particular focus on comparing the pixel-by-pixel model, land cover-based linear model, and trilinear regression model across various land surface types. The goal is to assess their effectiveness and identify potential limitations. To achieve this objective, the study addresses several key challenges in spaceborne GNSS-R SSM retrieval modeling, including: (1) optimal vegetation and roughness attenuation corrections and evaluating the impact of exponential function fitting within the pixel-by-pixel model; (2) constructing and assessing land cover-based semi-empirical inversion models; (3) evaluating the impact of Look-Up Table (LUT) improvements on the pixel-by-pixel model; and (4) comparing the performance of various semi-empirical inversion models to offer insights for their application and refinement. The findings of this study are expected to make a valuable contribution to the enhancement of more precise and dependable SSM retrieval algorithms. The subsequent sections of this paper are organized as follows: Section 2 discusses the datasets, methodologies, and experimental approaches employed; Section 3 presents the results, with analyses and discussions provided in Section 4; and finally, Section 5 offers the conclusions derived from this

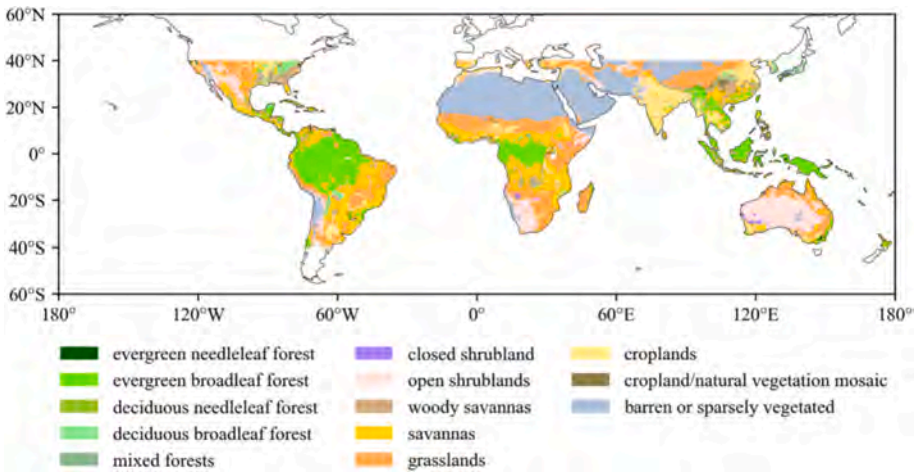

Fig. 1. The distribution of the IGBP land cover types over pan-tropical regions in the year of 2019.

Table 1
Model training and testing scores using pixel-by-pixel retrieval model with different vegetation and roughness correction strategies (unit for bias and ubRMSD is cm^3cm^{-3}).

Vegetation and roughness correction	Model Training			Model Testing		
	Bias	R	ubRMSD	Bias	R	ubRMSD
On specular point	0.000	0.912	0.047	−0.002	0.886	0.054
On grid level	0.000	0.914	0.051	−0.004	0.897	0.058

Table 2
Model training and testing scores using pixel-by-pixel model with different fitting functions on grid pixels (unit for bias and ubRMSD is cm^3cm^{-3}).

Fitting function	Model training			Model testing		
	Bias	R	ubRMSD	Bias	R	ubRMSD
Linear function	0.000	0.914	0.051	−0.004	0.887	0.058
Logarithmic function	0.000	0.914	0.051	−0.004	0.886	0.058

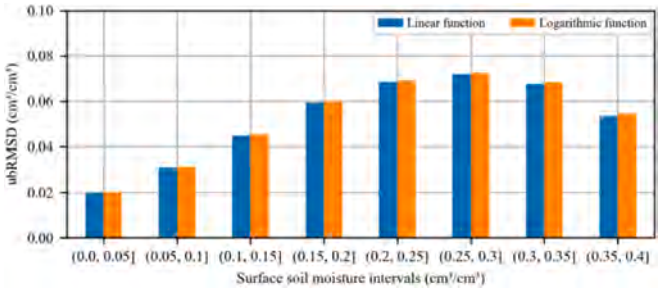

Fig. 2. The ubRMSD of surface soil moisture prediction using pixel-by-pixel model with linear and logarithmic function at different humidity intervals.

study.

2. Datasets

2.1. CYGNSS dataset

The CYGNSS mission offers observational coverage between 38°S to 38°N latitude, with sub-daily revisit periods over land. After completing in-orbit testing on 23 March 2017, CYGNSS transitioned to scientific operations, maintaining an average of 6–8 normally operating satellites

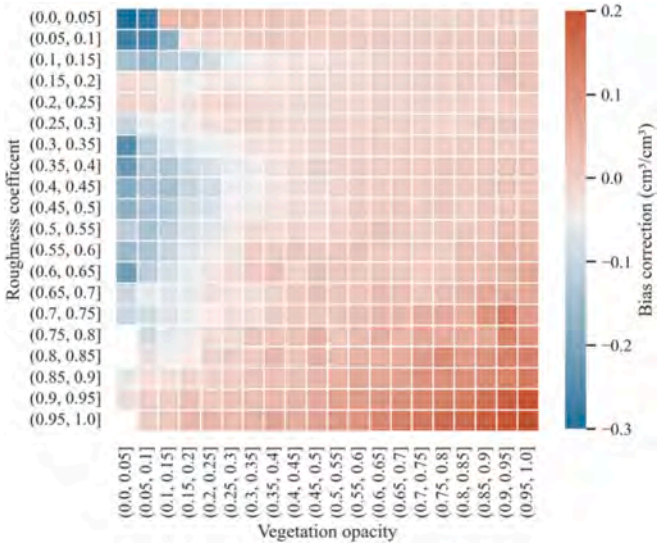

Fig. 3. Heatmap of prediction residuals from trilinear model as a function of vegetation opacity and roughness coefficient.

under typical conditions. The Delay Doppler Mapping Instrument (DDMI) aboard each CYGNSS satellite generates DDMs in real-time using1 ms coherent integration and 0.5–1 s incoherent integration periods, with a multi-stage process of analog radio frequency, digital processing, and calibration [37]. The downloaded DDMs are compressed to 17 delay bins by 11 Doppler bins, with delay resolution of 0.244 μs and Doppler frequency shift resolution of 500 Hz. Each satellite captures up to four reflected signal sets simultaneously, enabling extensive daily land sampling. Prior to August 2018, the DDMI operated in fixed gain mode, which is typically used during commissioning phase of a mission. After that, the receiver switched to adaptive gain mode to optimize the signal quality by automatically adjusting to variations in signal strength. In July 2019, CYGNSS increased its sampling frequency from 1 s to 0.5 s, enhancing the temporal resolution of land applications. All CYGNSS data products are open-access in the Physical Oceanography Distributed Active Archive Center (PODAAC) (https://podaac.jpl.nasa.gov/dataset). This study mainly uses the Level 1 Science Data Record Version 3.0 (CYGNSS_L1_V3.0) product, which is available in netCDF4 format. Each satellite offers a separate data file daily, and the data publication delay is about 6 days. The data product file contains the receiver power DDM calibrated by geographic location, the bistatic radar scattering cross-section DDM, and other scientific and engineering

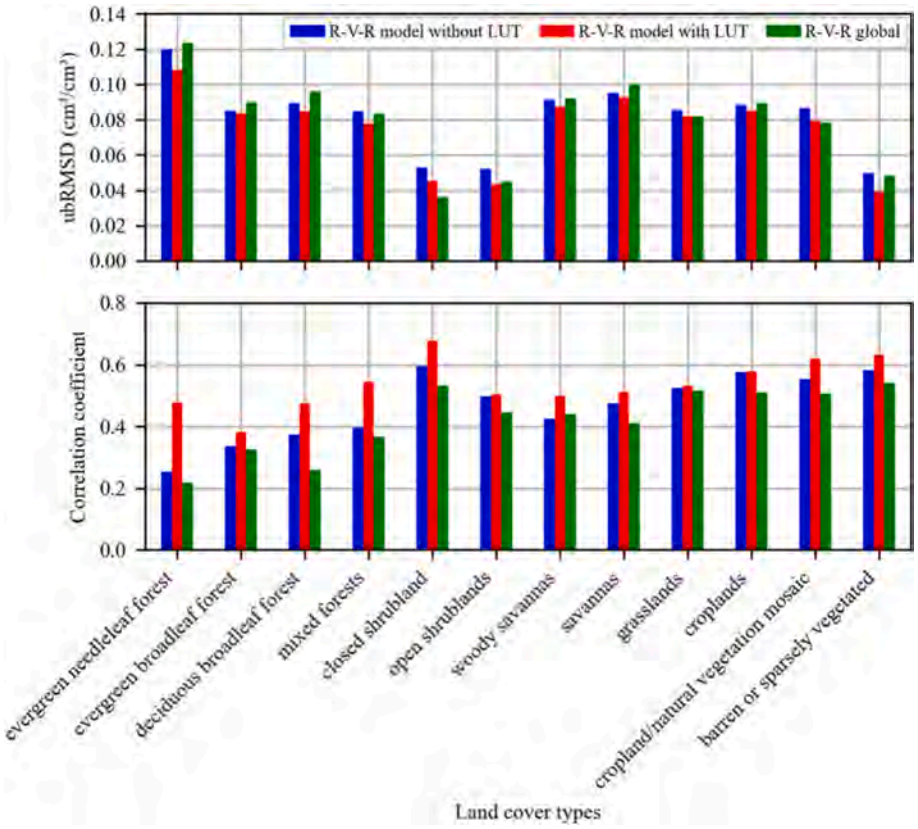

Fig. 4. Comparison of the ubRMSD and correlation coefficient for the R-V-R model without LUT correction, the R-V-R model using LUT-corrected R-V-R, and the global model in predicting surface soil moisture over various land cover types.

parameters.

2.2. SMAP soil moisture product

The SMAP Level-3 Version 8 radiometer-only soil moisture product (SPL3SMP) is a daily composite of half-orbit swaths, generated from the SMAP L2 Radiometer Half-Orbit 36 km Equal-Area Scalable Earth Grid, Version 2.0 (EASE-Grid 2.0) Soil Moisture, Version 8 (SPL2SMP), updated the dual channel algorithm (DCA) as the baseline algorithm. It provides comprehensive information on all data fields within the SPL3SMP product, available in HDF5 format (https://nsidc.org/data/spl3smp/versions/8). The product file also includes soil moisture parameters retrieved by the single-channel algorithm (SCA) used in the pre-Version 8 data products, along with essential vegetation and surface roughness parameters. The SMAP Level-3 radiometer SSM product has been widely applied in weather and climate forecasting, water resource management, and hydrological research. This product entails SSM that has been resampled to a global, cylindrical 36 km × 36 km EASE-Grid 2.0 projection, with a valid range spanning from 0.02 to 0.5 cm^3cm^{-3}. The dataset is updated on a daily basis and spans from April 1, 2015, to the present. In addition to SSM data, the product incorporates supplementary variables such as vegetation opacity and surface roughness parameters. For this study, we utilized combined daily averages of morning (AM) and afternoon (PM) product parameters. selecting only data flagged as "recommended" by the provider in the metadata [38].

2.3. Land cover type data

The Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover Climate Modeling Grid (MCD12C1), Version 6.1, is a global annual land cover type distribution dataset available from https://e4ftl01.cr.usgs.gov. It combines data from Terra and Aqua satellites to provide maps of the land cover types according to the International Geosphere-Biosphere Programme (IGBP), University of Maryland (UMD), and Leaf Area Index (LAI) classification schemes [39]. The MCD12C1 product provides global coverage from 2001 to 2021 at a native spatial resolution of 0.05° (approximately 5,600 m). To ensure consistent spatial resolution, the 2019 MCD12C1 data were resampled from the original 0.05° × 0.05° grid to a 36 km × 36 km EASE-Grid2 projection. Resampling was accomplished by uniformly mapping the finer resolution source grid points to the coarser target grid. The dominant land surface cover type within each coarse grid pixel was determined by aggregating the plurality of cover types from the underlying finer resolution cells. This aggregated modal cover type was then assigned as the representative value for each coarse grid cell. Given the focus on SSM retrieval in this study, certain observations were deliberately omitted based on the IGBP land cover type dataset. Specifically, data pertaining to water surfaces, permanent wetlands, urban and built-up areas, as well as snow/ice regions were excluded from our analysis. Deciduous needleleaf forests were also excluded, as they are absent in the pan-tropical regions covered by the CYGNSS mission. Fig. 1 illustrates the land cover types within the study area for 2019.

2.4. In-situ measurement

In situ measurements from the International Soil Moisture Network (ISMN), which collects and distributes high-quality soil moisture data sourced from internationally shared in situ monitoring networks for hydrology, meteorology, agriculture, and environmental research, were used to evaluate the performance of the retrieval model. The ISMN offer a reliable benchmarking platform for satellite-based SSM products due to its comprehensive coverage and accuracy. To evaluate the performance of the retrieval models developed in this work, we collected hourly soil moisture time series datasets at a depth of 0–0.1 m from all

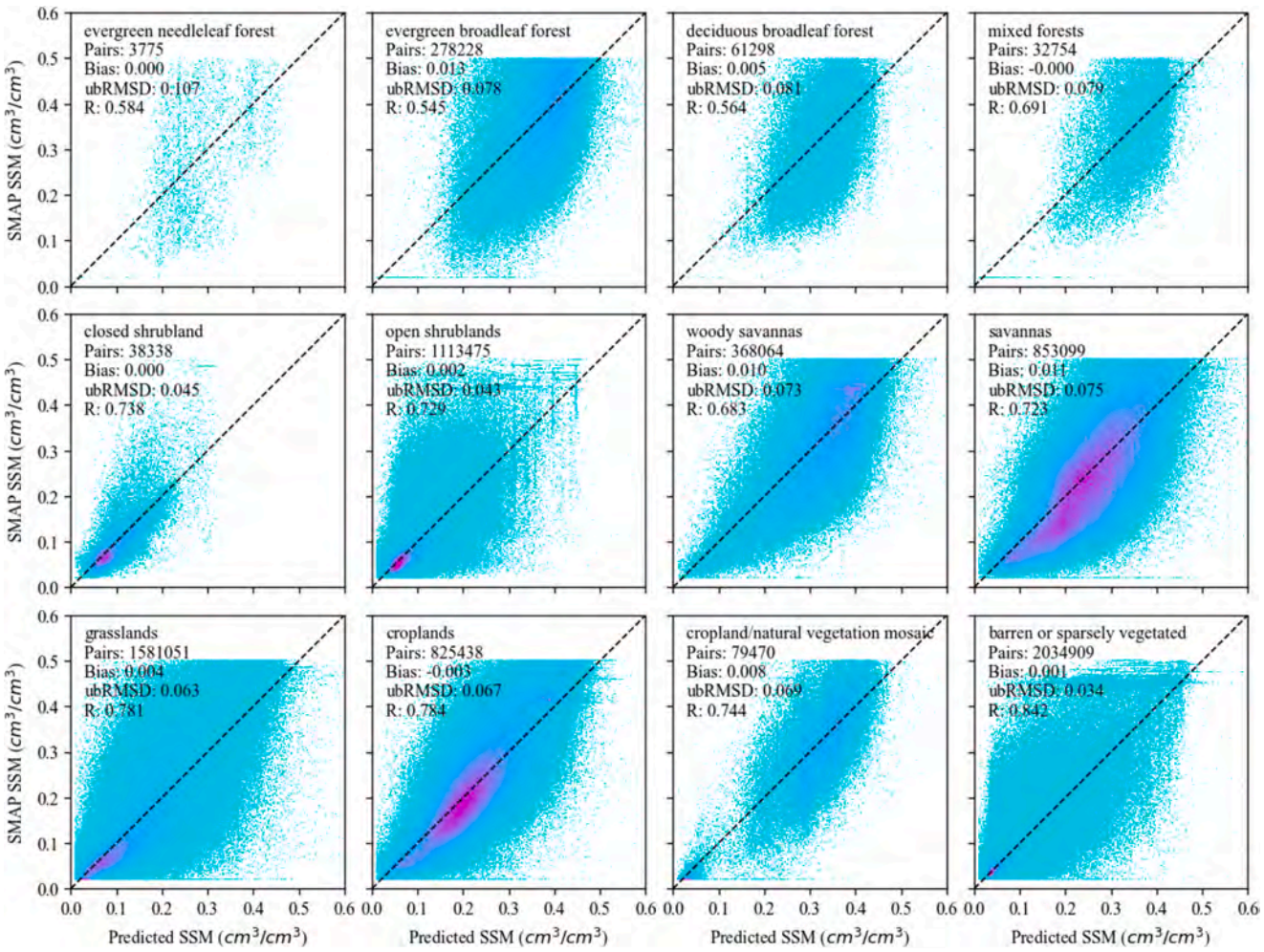

Fig. 5. Scatter density plot of pixel-by-pixel model predicted surface soil moisture and referenced SMAP L3 soil moisture over different land cover types.

ISMN network stations. Specifically, we downloaded observations from August 1, 2018, to July 31, 2020, that were marked as "good" by the quality control process setting on the ISMN website download panel [40]. To evaluate the gridded SSM data from the SMAP and CYGNSS SSM retrieval models, high-quality site measurements from the same network were temporally and spatially averaged within a grid pixel on a daily scale based on the flags provided by the ISMN.

3. Methodology

3.1. CYGNSS surface soil moisture proxy

The mathematical representation of surface reflectivity includes the Fresnel reflection coefficient as well as attenuation due to surface vegetation and roughness [29,41]. The dielectric constant of the land surface, which is primarily influenced by SSM, directly impacts the Fresnel reflection coefficient [5]. Thus, surface reflectivity can be viewed as proxy responding to SSM. This theoretical foundation underpins the utilization of GNSS-R technology for SSM remote sensing.

$$\Gamma_{\mathrm{RL}}(\epsilon_s,\theta) = \Re_{\mathrm{RL}}(\epsilon_s,\theta)\gamma^2 \exp\left(-h\cos^2\theta\right) \tag{1}$$

where Γ is surface reflectivity, *RL* stands for the left circular polarized scattering with the incoming right circular polarized signal, ϵ_s is the dielectric constant, θ is the incidence angle of the signal, $\Re$ is the circle polarized Fresnel reflection coefficient, γ is the transmissivity of the overlying canopy layer, $\gamma = \exp\left(-\tau_p \sec\theta\right)$, τ_p represents the nadir vegetation opacity, the parameter h is assumed linearly related to the root mean square height [42].

For the CYGNSS mission, land observations can be derived from the raw L1 V3.0 data product file using the per-DDM quality flags parameter, which is provided in 16-bit flag masks. Specifically, invalid observation data were filtered using a combination of different flag bits [28]. These include: "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". In the GNSS-R land applications, it is assumed that only coherent reflection occur on the land surface [21]. The surface reflectivity can be calibrated using the coherent reflection equation, as follows.

$$\Gamma_{\mathrm{RL}}(\epsilon_s,\theta) = \frac{(4\pi)^2 P_{coh}(R_T+R_R)^2}{\lambda^2 G_R P_T G_T} \tag{2}$$

where P_{coh} is the received DDM peak power, which was determined by the peak value of 17×11 array of DDM bin analog power, P_T is the GNSS satellite transmit power, G_T is the GNSS satellite antenna gain, G_R is the gain of the receiver antenna, λ is the carrier wavelength of the GNSS signal, R_T and R_R are the distances from the transmitter to the specular point and specular point to the receiver, respectively, and θ is the incidence angle of the incoming signal.

The requisite parameters were extracted from the CYGNSS L1 product files. Unlike theoretical reflectivity, which assumes perfect observations, surface reflectivity derived from CYGNSS is subject to various errors, such as observation errors, calibration errors, and other factors. Therefore, it is commonly referred to as the effective reflectivity, considering the combined impact of these factors on the retrieved

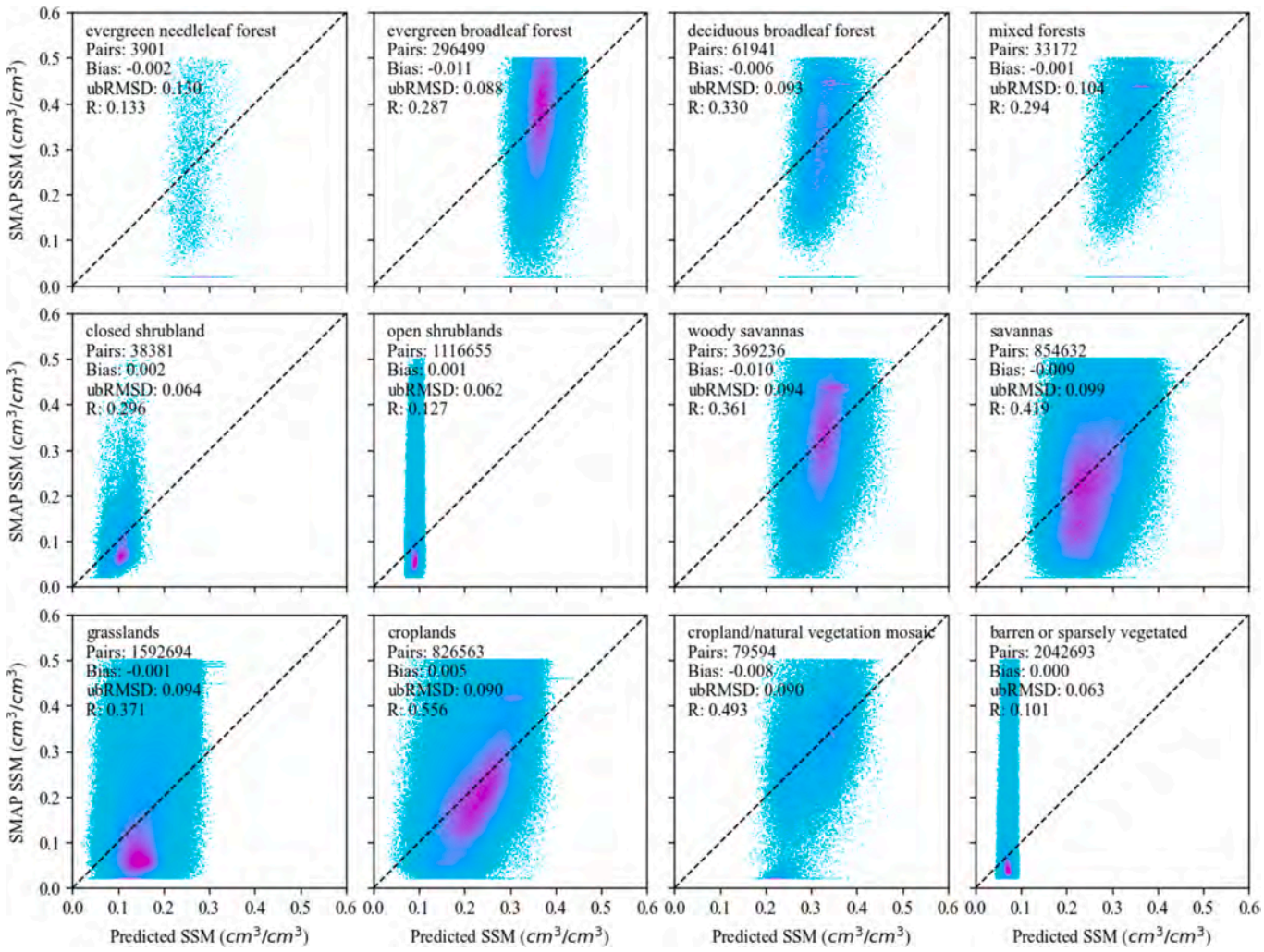

Fig. 6. Scatter density plot of land cover-based linear model predicted surface soil moisture and referenced SMAP data over different land cover types.

reflectivity values. In this study, a geometry correction methodology was employed to normalize the CYGNSS observations by accounting for the effects of varying incidence angles relative to the nadir direction [43]. The spatial resolution of the effective reflectivity at the specular point was mainly determined by observation geometry and surface roughness. It is generally assumed that the spatial footprint of CYGNSS effective reflectivity over land is approximately 7 × 0.5 km, or 3.5 × 0.5 km after mid-2019. Given the 36 km × 36 km EASE-Grid2 resolution, which exceeds the spatial scale of CYGNSS-derived effective reflectivity, this study treated CYGNSS observations as point-scale data. The gridded map was then created by rasterizing the effective reflectivity of individual specular points on a single day and aggregating the average values. This map maintains the same spatiotemporal resolution as the SMAP L3 reference SSM for semi-empirical modeling. Additionally, a minimum of five specular point effective reflectivity values within the same pixel is required. The gridded processing enhances the signal-to-noise ratio and reduces the representative error of the effective reflectivity relative to the reference data.

3.2. Enhanced pixel-by-pixel retrieval model

The prior pixel-by-pixel retrieval model was limited to estimating relative changes in SSM, as it modeled shifts in effective reflectivity and SSM rather than absolute values [21,28]. The model assumes that changes in vegetation and surface roughness occur on a much longer timescale than SSM. Therefore, modeling effective reflectivity variations can help minimize their influence. However, vegetation and surface roughness, particularly for crops and forests, are subject to continuous changes, including rapid growth and seasonal fluctuations. To accurately capture the complex dynamics influencing SSM, it is essential to incorporate these changes into the retrieval model. The enhanced model developed in this study rectifies biases in CYGNSS-derived effective reflectivity by integrating the tau-omega model [44], which incorporates parameters for vegetation opacity and surface roughness. The correction of the model to the effective reflectivity is given in Eq. (1), using the vegetation opacity and roughness coefficient parameters from the SMAP L3 data product. The raw point-scale effective reflectivity is rasterized to create an effective reflectivity image matching the spatial resolution of the SMAP soil moisture product. Corrections to the effective reflectivity can be applied in two ways: by first correcting the raw effective reflectivity of the specular points and then gridding them, or by correcting the aggregated effective reflectivity after gridding. The choice of a correction method can affect inversion results, as the process is sensitive to GNSS-R observation geometry, aggregation errors, and the accuracy of auxiliary parameter interpolation.

In the enhanced pixel-by-pixel retrieval model, the least squares regression coefficient estimates a linear relationship between the time series of the gridded CYGNSS-derived effective reflectivity and the SMAP reference SSM for each grid pixel.

$$u_{(i,j)} = a_{(i,j)}\overline{\Gamma}_{RL(i,j)} + b_{(i,j)} \tag{3}$$

where $u_{(i,j)}$ indicates the reference SSM values at grid point index (i, j),

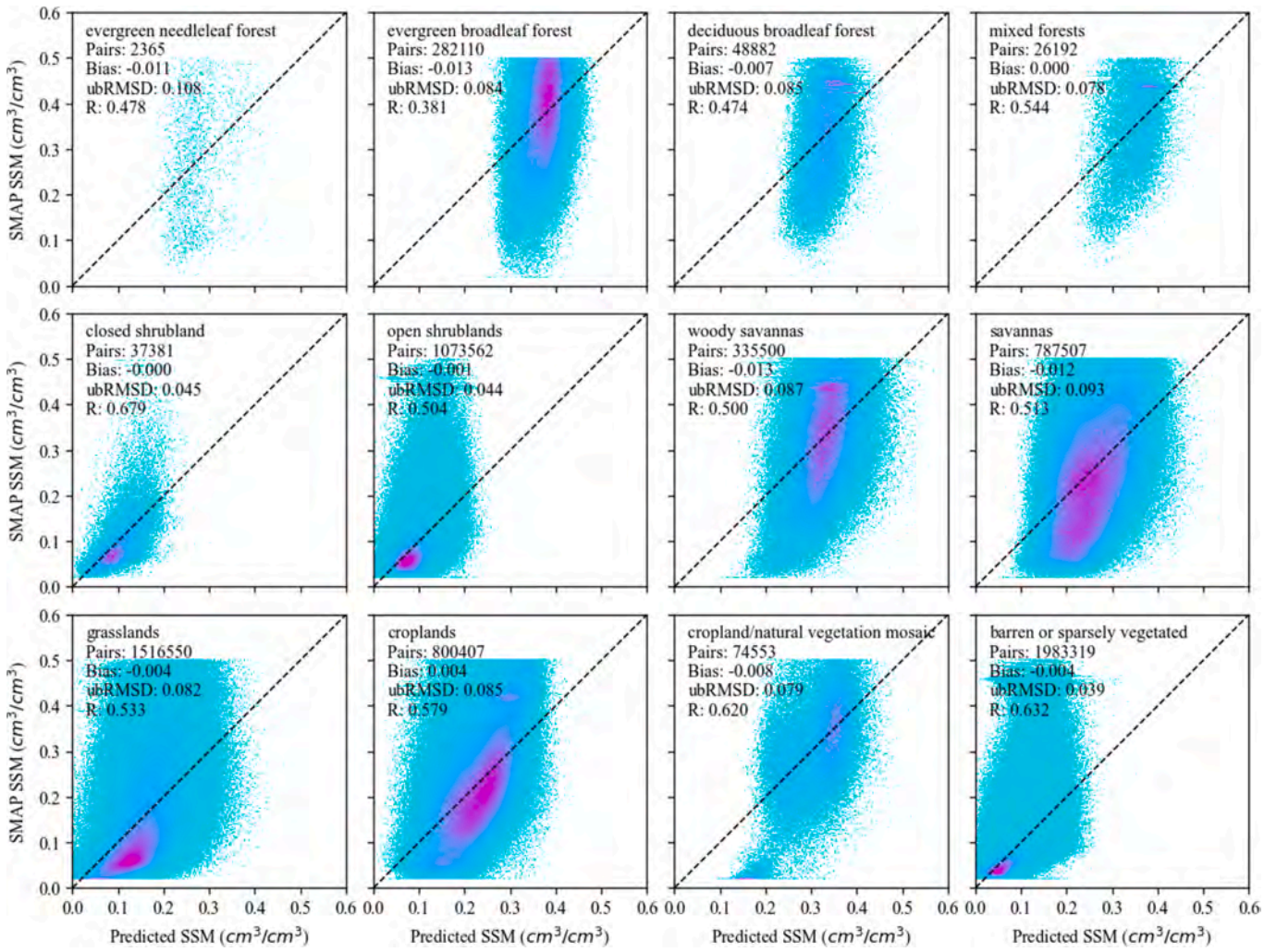

Fig. 7. Scatter density plot of R-V-R model predicted surface soil moisture and referenced SMAP data over different land cover types.

$\overline{\Gamma}_{RL(i,j)}$ represents gridded effective reflectivity with vegetation and surface roughness attenuation correction, $a_{(i,j)}$ denotes the slope of the linear model, and $b_{(i,j)}$ represents the intercept. The established linear relationship can calibrate the absolute SSM using effective reflectivity. The entire model for quasi-global CYGNSS SSM mapping can be represented in matrix form.

$$\boldsymbol{U} = \boldsymbol{A}\overline{\Gamma} + \boldsymbol{B} \tag{4}$$

where $\boldsymbol{U}$ indicates the reference SSM image for modeling or CYGNSS SSM image from retrieval model prediction, $\overline{\Gamma}$ represents the observations matrix of gridded CYGNSS-derived effective reflectivity, $\boldsymbol{A}$ and $\boldsymbol{B}$ are the coefficient matrix and intercept matrix of the formed retrieval model.

Based on theoretical simulations using the dielectric constant model and the Fresnel equations [5], the Fresnel reflection coefficient increases continuously with increasing SSM up to the point of soil saturation. The logarithmic function provided a better fit for the correlation between the two variables. Prior studies did not examine the incorporation of a logarithmic function in the pixel-by-pixel SSM retrieval model. This study also assessed the effectiveness of applying a logarithmic function at the pixel level.

$$u_{(i,j)} = a_{(i,j)} ln(-b_{(i,j)}\overline{\Gamma}_{RL(i,j)}) + c_{(i,j)} \tag{5}$$

where a, b, and c are parameters to be estimated in the model.

In order to conduct pixel-by-pixel modeling, at least five aligned time series of the reference SSM and CYGNSS effective reflectivity for a given pixel are required. Without these, the prediction model cannot be generated for that specific pixel.

3.3. Land cover-based linear model

In the study by [35], a representative model was developed to separately model SSM inversion across various surface features driven by the diverse surface vegetation and roughness characteristics of different land properties. The model clusters the surface using SMAP L3 vegetation opacity and roughness, establishing linear relationships between CYGNSS-derived effective reflectivity and reference SSM within each cluster. Building on the finding that the roughness parameters included in the current SMAP Level 3 product have minimal impact on the retrieval outcomes [45], this result aligns with previous research [35], which showed that clustering the surface based on SMAP Level 3 vegetation opacity and roughness closely mirrored the distribution of land cover types from the IGBP. Leveraging this insight, the study developed dedicated linear regression models for each land cover type.

$$u = a_{lc} \cdot \overline{\Gamma}_{RL} + b_{lc} \tag{6}$$

where lc indicates the land cover type, a_{lc} and b_{lc} are the slope and intercept, $\overline{\Gamma}_{RL(i,j)}$ represents gridded effective reflectivity after vegetation and surface roughness attenuation corrections. An optimization method is used to enhance modeling accuracy by removing extreme outliers

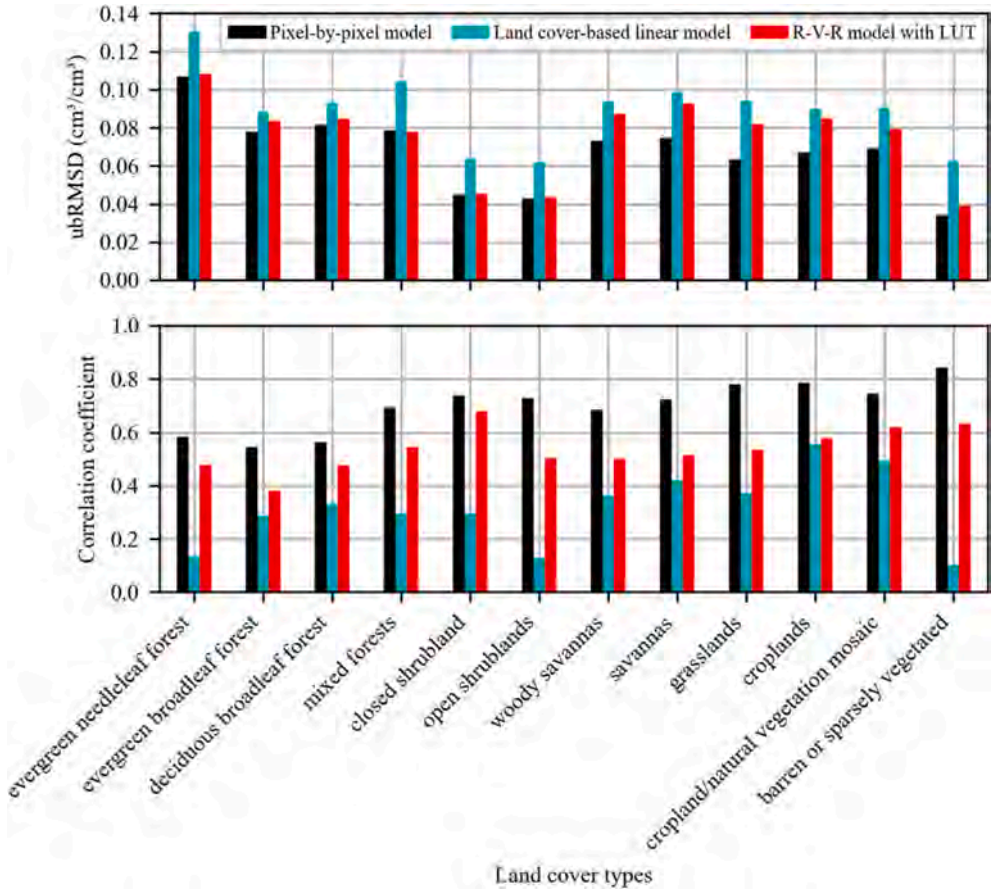

Fig. 8. Comparison of the ubRMSD and correlation coefficient for predicting surface soil moisture using the pixel-by-pixel model, land cover-based linear model, and R-V-R model with LUT correction over various land cover types.

Table 3
The total scores of three model predictions (unit for bias and unRMSD is cm^3cm^{-3}).

SSM retrieval models	Bias	R	ubRMSD
Pixel-by-pixel model	−0.004	0.887	0.058
Land cover-based Linear model	−0.002	0.756	0.082
R-V-R model	−0.004	0.806	0.073

from the training dataset of each land cover type. This method trims the lowest and highest 2.5 % of the training samples.

3.4. Reflectivity-Vegetation-Roughness retrieval model

The Reflectivity-Vegetation-Roughness (R-V-R) GNSS-R SSM retrieval model was first developed in [22] and later improved by incorporating additional DDM observables and applying multivariate linear regression [23]. The original R-V-R model functions as a trilinear regression, modeling SSM as a function of effective reflectivity, vegetation opacity, and surface roughness. However, the regression residuals analysis revealed a significant correlation between vegetation opacity and surface roughness, exposing a limitation of the trilinear model in accurately capturing their independent effects on SSM estimation. As a result, predictions from trilinear model may exhibit systematic biases under varying vegetation and roughness conditions. To address this issue, regression errors were grouped by intervals of vegetation opacity and surface roughness, with the average error calculated within each interval. This data was used to create a LUT to correct the systematic biases in the model's predictions. The R-V-R model, along with its corresponding correction LUT, enables both global and regional applications with just a single trilinear regression model needed for the study area.

The original R-V-R model was improved by incorporating land cover type information. Using IGBP land cover type data, individualized R-V-R models were developed for each surface type. Spatiotemporal matching techniques were applied to gather training data specific to each of the 12 land cover categories.

$$u = a_{lc}\cdot\overline{\Gamma}_{RL} + b_{lc}\cdot\tau + c_{lc}\cdot\sigma + d_{lc} \tag{7}$$

where $\overline{\Gamma}_{RL}$ represents the aggregated effective reflectivity on grid pixel without vegetation and surface roughness correction; τ represents the vegetation opacity, and σ represents the roughness coefficient of ground surface; a, b, c and d were model parameters that need to be estimated in the trilinear regression; the subscript lc represents a specific land cover type. The bottom and top 2.5 % of training samples for each land cover type are excluded to improve the data quality. The ordinary least squares (OLS) method was used to perform a correlation test between the regression residuals and the vegetation opacity and roughness coefficients before establishing the LUT. The formed enhanced R-V-R model for SSM prediction initially uses the trilinear function over a specific land cover type. This function predicts SSM using vegetation opacity, surface roughness, and effective reflectivity parameters. Subsequently, a corresponding LUT integrates surface roughness and vegetation opacity to calculate correction factors for SSM prediction.

3.5. Experimental design

To achieve the research aims, a flexible inversion program was

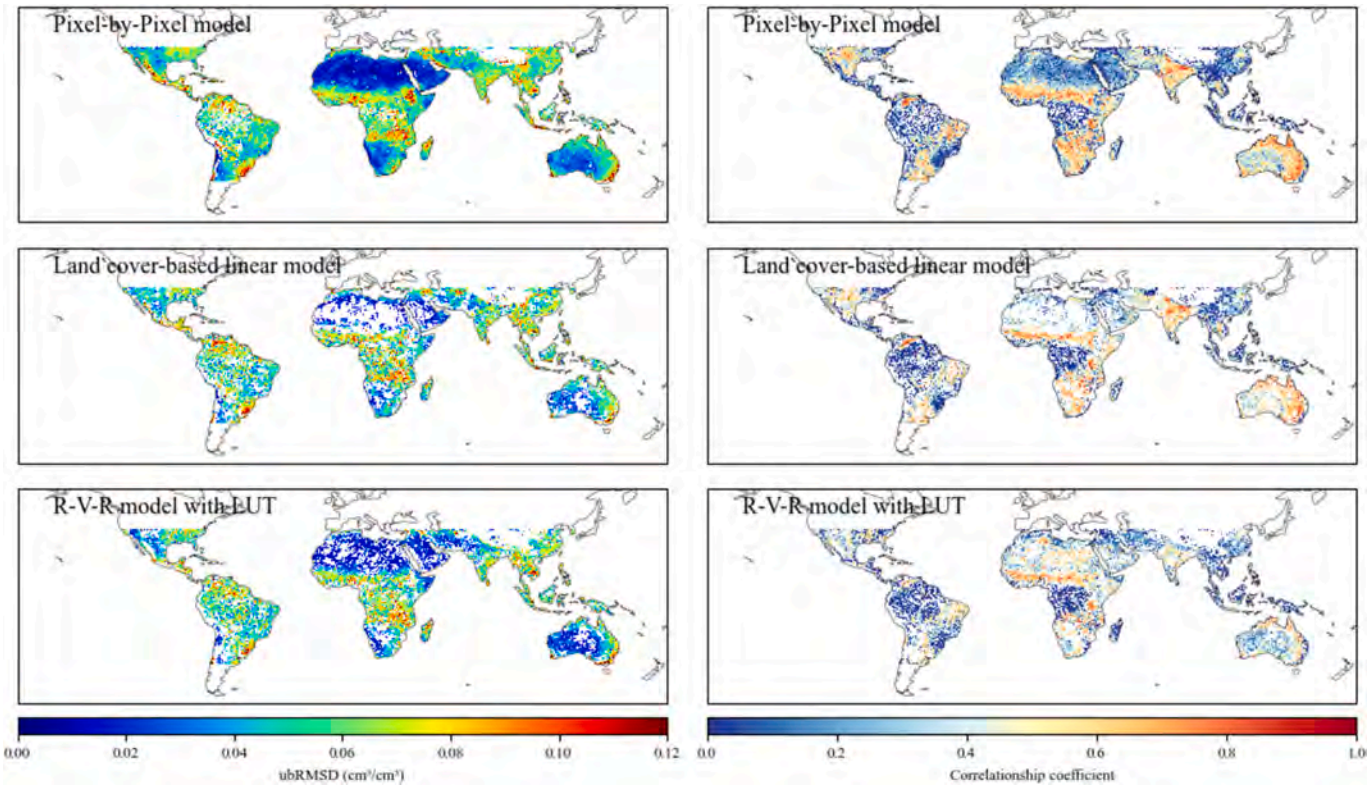

Fig. 9. Spatial distribution of the ubRMSD and correlation coefficient for the soil moisture retrieval using pixel-by-pixel model, land cover-based linear model, and R-V-R model with LUT correction. The color bars on the bottom represent the scale for the ubRMSD and R values, respectively.

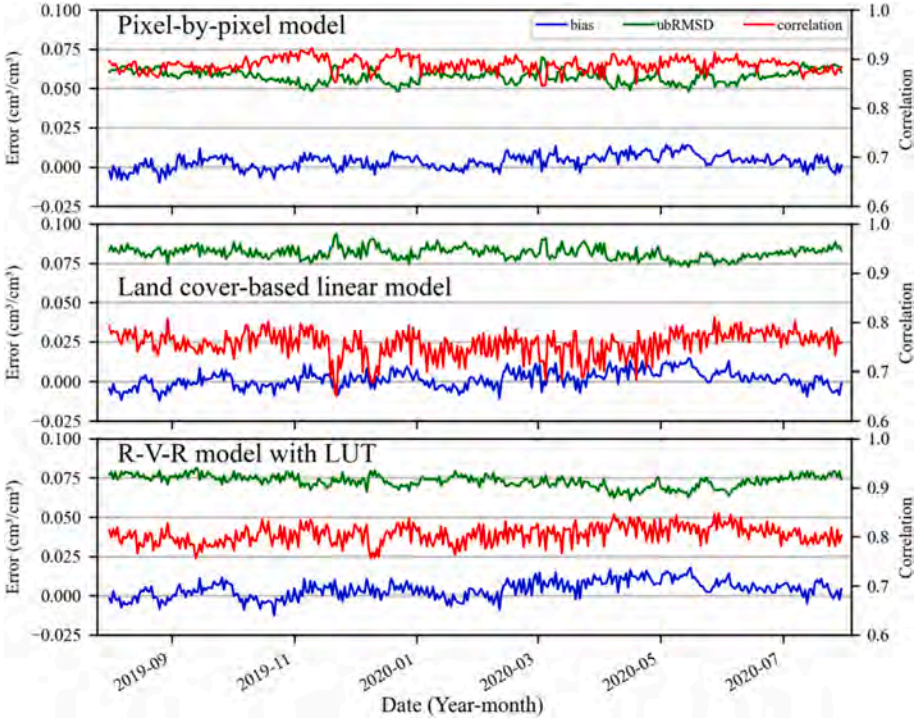

Fig. 10. Temporal skill metrics comparison between the estimated surface soil moisture from pixel-by-pixel model, land cover-based linear model, and R-V-R model with LUT correction.

developed to implement pixel-wise and land cover-dependent models. This enabled the rapid incorporation of varied data sources and adjustment of correction strategies for vegetation and surface roughness.

To develop and assess the SSM retrieval models, we utilized CYGNSS data spanning two years (August 1, 2018, to July 31, 2020). Specifically, the first year (August 1, 2018, to July 31, 2019) was used for model

Table 4
Error metrics median of SMAP product, predictions from pixel-by-pixel model and R-V-R model from independent ISMN networks (unit for bias and unRMSD is cm^3cm^{-3}).

SSM product	Network	Collocated grid pixel Num.	bias	ubRMSD	R
Prediction from pixel-by-pixel model	SCAN	33	−0.014	0.060	0.282
	ARM	3	0.002	0.038	0.597
	OZNET	7	−0.008	0.042	0.547
	SNOTEL	8	−0.011	0.074	0.222
	TAHMO	12	−0.002	0.053	0.504
	USCRN	10	−0.002	0.055	0.260
Prediction from R-V-R model	SCAN	33	0.011	0.059	0.244
	ARM	3	−0.011	0.060	0.439
	OZNET	7	0.010	0.053	0.468
	SNOTEL	8	−0.018	0.072	0.105
	TAHMO	12	0.023	0.046	0.512
	USCRN	10	0.003	0.055	0.217
SMAP	SCAN	33	−0.002	0.050	0.650
	ARM	3	0.014	0.042	0.780
	OZNET	7	−0.018	0.053	0.739
	SNOTEL	8	0.010	0.073	0.444
	TAHMO	12	0.000	0.052	0.733
	USCRN	10	0.005	0.047	0.557

training, while the second year (August 1, 2019, to July 31, 2020) served as validation dataset.

Initially, a comparative analysis was conducted between vegetation and surface roughness attenuation corrections of effective reflectivity at the individual specular point or rasterized grid point level in the improved pixel-by-pixel retrieval model. Following this, an evaluation of linear and logarithmic fitting functions employed in pixel-by-pixel retrieval modeling was undertaken. Subsequently, R-V-R retrieval models were generated for different land cover types, with a focus on assessing the enhancement achieved through bilinear interpolation LUT in the R-V-R model. Building on these investigations, we formulated optimal inversion models based on the most effective approaches and evaluated their performance on a test dataset using three distinct inversion models. Furthermore, we compared the improvement in inversion accuracy achieved through bilinear interpolation LUTs and examined the impact of incorporating LUTs to correct inversion results within the pixel-by-pixel model.

To further validate the retrieval performance, ISMN in-situ measurements were employed. Only matched pairs within the same grid pixel were considered, which included the predicted SSM from the retrieval model, SMAP SSM maps, and ISMN station measurements with a minimum of 31 samples. To maintain fairness and rigor in the statistical comparison, the same stations with complete valid scores across all SSM sources were retained [46,47]. The evaluation of performance entailed utilizing various skill metrics, such as mean bias, Pearson correlation coefficient (R), and unbiased root mean square difference (ubRMSD).

4. Results and analysis

4.1. Vegetation and surface roughness correction evaluation

The CYGNSS mission offers L1 data products provided necessary parameters that allow the calibration of the effective reflectivity at the specular point using Eq. (2). Current inversion techniques commonly use space–time averaging, which involves reprojecting and rasterizing the effective reflectivity of specular points to create CYGNSS-derived effective reflectivity images at different spatiotemporal resolutions. This process is essential for reducing observation noise and improving the signal-to-noise ratio. The resultant effective reflectivity map from this stage captures contributions from both surface vegetation and surface roughness. One approach involves adjusting the specular point effective reflectivity obtained from raw CYGNSS observations by interpolating SMAP-derived vegetation opacity and roughness parameters at specular points for Eq. (1). However, a thorough data analysis revealed that about 10 % of the calibrated effective reflectivity would fall between −40 to −5 dB outside its valid range due to errors in interpolation and representativeness, necessitating their exclusion before rasterization. Notably, the most substantial reduction in data points occurred predominantly in forested regions. Such instances may accentuate representativeness errors in gridded effective reflectivity and result in under-sampling for specific grid pixels, potentially causing modeling inadequacies. As an alternative, it is proposed to directly correct the aggregated average effective reflectivity at the grid points of the gridded effective reflectivity map.

Both solutions were developed to model 55,855 land grid cells. As shown in Table 1, the first solution, which applied corrections at the specular point level, exhibited slight improvement in training and testing accuracy over the second solution, which corrected the aggregated reflectivity map. Notably, neither approach significantly impacted on the number of accurately modeled grid pixels. The correction process involves interpolating the vegetation opacity and roughness coefficient to rectify the discrete effective reflectivity values obtained from GNSS-R observations separately. It is essential to recognize that the spatial and temporal resolution of the vegetation and roughness parameters can introduce anomalies in the correction process. Thus, it is recommended to directly adjust the calibrated specular point effective reflectivity using high-resolution vegetation and roughness parameters to ensure accuracy. Conversely, correcting vegetation and roughness at the grid level is deemed a simpler and more efficient process. Due to the spatial resolution limitations of the vegetation opacity and roughness coefficient

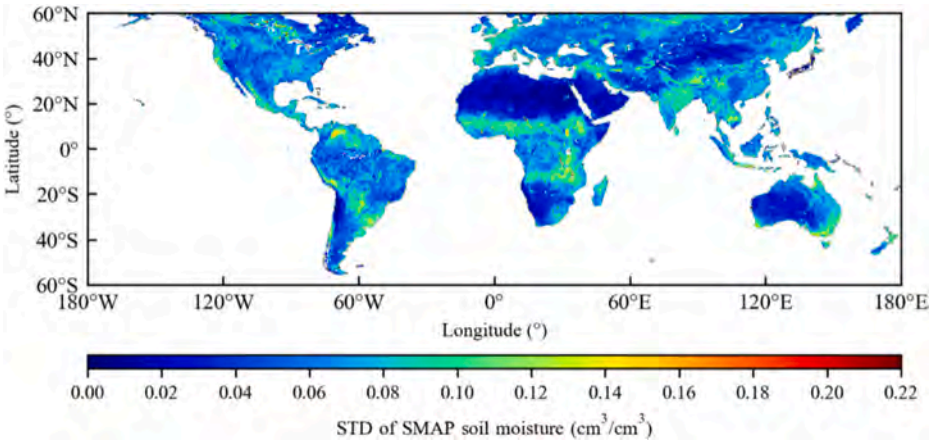

Fig. 11. Spatial map of the standard deviation of surface soil moisture from the SMAP product.

data used in this study, the second solution was implemented in subsequent experiments involving a pixel-by-pixel retrieval model. Additionally, the use of vegetation opacity and surface roughness parameters from the SMAP SCA versus DCA algorithms in the SMAP L3 product had minimal impact on SSM retrieval results using the pixel-by-pixel model. Future experiments will incorporate the parameters derived from the DCA algorithm for further analysis and refinement.

4.2. Comparison of pixel-level modeling functions

The Fresnel reflection coefficient exhibits a non-linear, monotonic relationship with SSM within its dynamic range [45], which can be accurately modeled using logarithmic functions. However, it is essential to differentiate between the physically relationships governing reflectance and soil moisture in models and the statistically derived correlation observed between CYGNSS effective reflectivity and SMAP reference SSM at specific area. This distinction is necessary due to differences in errors and uncertainties associated with actual observation and reference SSM. Furthermore, some regions, such as arid and semi-arid areas, experience a limited annual range of SSM. In these regions, the connection between effective reflectivity and SSM can be properly represented by a linear model. Therefore, the fitting function derived from the physical mechanism simulation cannot be entirely applicable to actual observation data, but rather serves as a reference for analysis. Thus, SSM retrieval models were compared to evaluate the efficiency of linear and logarithmic functions used at the pixel level in the pixel-by-pixel retrieval model.

Table 2 displays the skill metrics for employing linear and logarithmic functions in the pixel-by-pixel retrieval model to assess their performance in calibrating SSM from the CYGNSS-derived effective reflectivity. The evaluation metrics reveal that both models exhibited comparable performance with negligible differences in bias and accuracy. Specifically, both models demonstrated near-zero biases, indicating no systematic over- or under-prediction, while the high correlation coefficients suggested a strong correlation between model predictions and the reference values. The training ubRMSD was consistent at 0.051 cm^3cm^{-3} for both models. The prediction error on the test set was found to be 0.058 cm^3cm^{-3}, highlighting the considerable accuracy of both linear and logarithmic models. Ultimately, this analysis shows that both linear and logarithmic functions can be used accurately calibrate SSM from the effective reflectivity at the pixel level, with similar performance. Furthermore, to compare the performance of the two function models, prediction errors on the test set were analyzed across 0.05 cm^3cm^{-3} SSM intervals within the dynamic range. As Fig. 2 shows, when SSM was under 0.2 cm^3cm^{-3} the functions exhibited comparable performance. However, above 0.2 cm^3cm^{-3}, the linear model slightly outperformed the logarithmic model, even under humid conditions.

4.3. Intercomparison of different reflectivity-vegetation-roughness models

In contrast to the original global R-V-R model, the enhanced R-V-R model was specifically developed for diverse land cover types using IGBP land cover classification data. The trilinear model and the corresponding corrected LUTs on different surface types were simultaneously generated. The heatmap corresponding to the LUTs created when building the global R-V-R model is shown in Fig. 3. Subsequently, the performance of the land cover-based R-V-R inversion models was compared with the global-scale R-V-R model, both before and after LUT correction, across various surface cover types. The trilinear regression analysis indicates that effective reflectivity was the most influential variables, exhibiting the strongest correlation with model predictions. Notably, this finding implies that vegetation and surface roughness characteristics may have contributed to the observed prediction bias. The significance testing found no statistically significant correlation between the residuals of the model fit and vegetation and surface roughness, either in building R-V-R models for different land cover types or at a global scale. However, comparing the trilinear regression model predictions to the reference SSM shows apparent prediction bias. To address this issue, a 2D LUT was established by calculating the mean residuals for different vegetation and roughness intervals.

The performance of the enhanced R-V-R model across various surface types follows a consistent pattern with the global-scale R-V-R retrieval model, as shown in Fig. 4, which compares histograms of ubRMSD and correlation coefficients for predicted SSM values on the test set. Inversion errors were higher for moist surfaces, such as forests, grasslands, and agriculture areas, and lower for deserts. The comparisons involve the land cover-based R-V-R model before and after applying LUT correction, alongside the global-scale R-V-R retrieval model. The analysis that while the R-V-R model derived from trilinear regression fitting cannot fully eliminate the influence of vegetation and surface roughness, prediction accuracy significantly improves with the application of LUT correction. The land cover-based R-V-R model demonstrated high prediction accuracy. Apart from closed shrubland and open shrubland surfaces, the predictive accuracy of the trilinear function fitted on other land cover types is already close to the global-scale R-V-R model. This highlights the benefits of developing land cover-specific R-V-R models, especially in wooded areas where the greatest improvement in correlation coefficients were observed. Hereafter, the R-V-R model refers to the methodology developed for various land-cover types.

4.4. Comparison of different inversion models for each land cover type

Given the consensus in the previous studies that SSM retrieval is sensitive to vegetation and surface roughness, both of which exhibit complex spatiotemporal variations, this study aimed to evaluate the accuracy of empirical inversion models for SSM estimating across diverse land cover surfaces. The pixel-by-pixel model used a unary linear function for each grid pixel. The land cover-based linear model and R-V-R model were built based on land cover types by upscaling the MCD12C1 IGBP land cover data. The auxiliary parameters, including vegetation optical thickness and surface roughness, were sourced from the estimations provided by the DCA algorithm within the SMAP L3 product.

Scatter density plots of the predicted SSM versus reference SSM by the pixel-by-pixel model, land cover-based linear model, and R-V-R model on the test dataset for different land covers are depicted in Figs. 5, 6, and 7. The inversion results reveal significant variations across land cover types. Notably, for certain land covers, such as evergreen needleleaf forests in mountainous regions, limited CYGNSS coverage resulted in reduced accuracy due to the smaller test dataset. The scatter distribution demonstrates higher SSM values for forest covers (including evergreen broadleaf forest, deciduous broadleaf forest, and mixed forest), predominantly exceeding 0.2 cm^3cm^{-3}, albeit with slightly elevated inversion errors. In contrast, the inversion accuracy for herbaceous plant-covered surfaces (grasslands, croplands, savannahs, closed shrublands, and open shrublands) was notably superior, particularly for grasslands and croplands, with strong agreement shown with reference data across all SSM levels. This finding provides additional support for the assertion that GNSS-R remote sensing displays significant resilience to the impact of vegetation attenuation effects [19]. Furthermore, both land cover-based linear and R-V-R models demonstrate a distinctive limitation in predicting quasi-saturated SSM values.

Fig. 8 presents histograms of ubRMSD and correlation coefficients for SSM estimates from three retrieval models across diverse land covers, facilitating an intuitive assessment of model performance. The comparison of the inversion models revealed that the land cover-based linear model produced the least accurate results, with considerable limitations. Notably, both the R-V-R and land cover-based linear models exhibited greater bias than the pixel-by-pixel model across a range of land surface types. The pixel-by-pixel model achieved lower ubRMSD compared to SMAP reference SSM across most land covers, except for evergreen

needleleaf forests. The land cover-based linear model showed weak agreement with the reference SSM on various surfaces, often producing anomalous results, with correlation coefficients typically below 0.4. Its performance was particularly poor in shrublands and barren or sparsely vegetated areas. In contrast, the pixel-by-pixel model generally achieved correlation coefficients above 0.6. Although the R-V-R model outperformed the land cover-based linear model, its results remained inferior to those of the pixel-by-pixel model.

Table 3 summarizes the total prediction error statistics of three models on the test dataset, offering insights into their performance across different land cover types. Model evaluation indicates that all models exhibit minimal prediction biases, with the pixel-by-pixel model demonstrating superior prediction accuracy. Specifically, the pixel-by-pixel model exhibited an ubRMSD of 0.058 cm^3cm^{-3} and a correlation coefficient of 0.887 in comparison to the reference SMAP SSM. On the contrary, the land cover-based linear model displayed the weakest performance, with an ubRMSD of 0.082 cm^3cm^{-3} and a correlation coefficient of 0.756. The R-V-R model, although surpassing the land cover-based linear model, showed inferior performance compared to the pixel-by-pixel model, with ubRMSD and correlation coefficients of 0.073 cm^3cm^{-3} and 0.806, respectively.

Fig. 9 provides a comprehensive assessment of the spatial distribution of inversion errors and inter-model variability in SSM estimation accuracy through spatial maps of ubRMSD and Pearson correlation coefficients for three retrieval models. The pixel-by-pixel model proves to be the most reliable for grid-level SSM prediction, demonstrating superior performance with a broader spatial coverage, even in arid regions where CYGNSS observations are more prone to noise, resulting in lower ubRMSD. However, while the pixel-by-pixel model excels in prediction accuracy based on ubRMSD, the correlation coefficient map shows lower values for this model. After applying quality control procedures to the inversion results and enforcing a valid range of $0.0 \sim 0.65$ cm^3cm^{-3} for model prediction, it was found that both the land cover-based linear model and the R-V-R model showed larger errors compared to the pixel-by-pixel model. In arid regions prone to noise, all these models produced unreliable predictions, resulting in a reduction in the coverage area of the metrics spatial map. Notably, the R-V-R model outperformed the land cover-based linear model, although both struggled in arid areas due to the data noise. Conversely, the pixel-by-pixel model showcased higher correlation between estimates and references, surpassing the other two models, particularly in wet areas. The comparative analysis reveals that the land cover-based linear model is significantly less effective, primarily attributed to the spatiotemporal variability of surface vegetation and roughness. This variability makes it challenging for land cover-based linear models in constructing inversion models that consistently deliver accurate and reliable estimates across diverse surface coverage conditions.

Fig. 10 compares the time series of daily bias, ubRMSD, and correlation between the predicted SSM maps from the three retrieval models and the reference SMAP SSM maps during the testing period. All three models show stable performance over time, with no significant degradation as the extrapolation period extends, despite minor fluctuations in the daily statistics, particularly in correlation coefficients. The time-domain bias sequences consistently converge to approximately 0.0 cm^3cm^{-3} across all three models. Notably, the pixel-by-pixel model consistently maintains an ubRMSD below 0.07 cm^3cm^{-3}, with a daily correlation statistic around 0.88. In contrast, the land cover-based linear model shows an ubRMSD time series exceeding 0.075 cm^3cm^{-3} overall, with daily correlation statistics consistently below 0.80. The R-V-R model with LUT maintains a daily ubRMSD time series of roughly 0.075 cm^3cm^{-3}, alongside correlation statistics around 0.80. The pixel-by-pixel model displays markedly superior time-domain performance compared to the other two models. This exhaustive analysis reveals the significant performance lag of the land cover-based linear model compared to the other models, leading to its exclusion from subsequent validation and evaluation processes.

4.5. Pixel-by-pixel model with LUT

The analysis revealed that the global R-V-R model had lower retrieval accuracy compared to the land cover-based R-V-R model. Notably, without LUT correction, the accuracy of the land cover-based model declined to match that of the global model, highlighting the critical role if LUTs in improving prediction accuracy, especially when tailored to specific land cover types. To address potential systematic biases in the pixel-by-pixel model from insufficient vegetative and surface roughness attenuation corrections, a correction LUT was created using prediction residuals and normalized vegetation opacity and surface roughness data. This correction LUT was then integrated into the pixel-by-pixel model with refined predictions incorporated interpolating corrections based on observed surface vegetation opacity and roughness values. The efficacy of this enhanced approach was rigorously validated through assessment of its prediction performance on the test set, demonstrating robust and reliable estimates of SSM across diverse land cover classes, as reflected in the results presented in Fig. 5. Despite the integration of LUT corrections, minimal deviations in the overall predictive accuracy were observed compared to the original pixel-by-pixel model, as detailed in Table 3, where the overall bias was determined at -0.002 cm^3cm^{-3}, RMSD at 0.054 cm^3cm^{-3}, and the correlation coefficient at 0.886. These findings suggest that the regression fit might not comprehensively account for vegetation and surface roughness effects resulting from observation errors, temporal variations, and geographical differences in auxiliary factors.

4.6. Evaluation of retrieval performance with in-situ measurement

After spatiotemporal matching and screening, six sparse networks comprising 73 ground stations within the CYGNSS footprint were used to validate the performance of two SSM retrieval models. Table 4 presents the median evaluation metrics of the collocated grid time series between in-situ measurements from each ISMN network and the gridded SSM data, including the SMAP SSM dataset, pixel-by-pixel model predictions, and the R-V-R model predictions during the test period. The model performance varied across different in situ soil moisture networks, with the most significant performance discrepancies observed in the SNOTEL network. Specifically, for the pixel-by-pixel model, the median ubRMSD stands at 0.074 cm^3cm^{-3}, and the median correlation coefficient at 0.222 when considering SNOTEL stations. On the other hand, the R-V-R model with LUT correction exhibits a median ubRMSD of 0.072 cm^3cm^{-3} and a median correlation coefficient of 0.105 over the same SNOTEL stations. Additionally, the analysis indicates weak correlations with station measurements when considering the SCAN and USCRN networks. This result is consistent with previous study [48]. However, the performance over ARM, OZNET, and TAHMO stations is relatively more satisfactory. Overall, although both models display room for enhancement compared to SMAP SSM products, the pixel-by-pixel model consistently outperforms the R-V-R model.

5. Discussion

This study compares and analyzes three extended GNSS-R semi-empirical models for SSM retrieval, with a focus on their distinct approaches to handing land cover heterogeneity, accounting for vegetation-induced signal loss, and mitigating the effects of surface roughness on signal attenuation. Accurate reference SSM values are crucial for semi-empirical fitting methods, while vegetation and surface roughness significantly affect inversion outcomes. Among the models evaluated, the pixel-by-pixel model stands out as the most sophisticated version, as it generates a unique model for each grid pixel compared to the land cover-based linear model and R-V-R model. The distinctive feature of the pixel-by-pixel model lies in its higher level of granularity, providing detailed modeling at the individual pixel level, unlike the broader corrections used in other models. Performance varies

significantly across different surfaces, as reflected in the results. To improve prediction accuracy, the pixel-by-pixel model uses a two-step approach: first, geometric correction is applied to the point-scale effective reflectivity of specular points, followed by vegetation and surface roughness correction on the aggregated EASE-Grid 2.0 pixel effective reflectivity. This sequential process has yielded better inversion results and greater efficiency. Correcting for vegetation and surface roughness attenuation significantly improved the pixel-by-pixel model's performance on moderately vegetated land surfaces. Linear and logarithmic function performed similarly at the pixel level, even at high SSM levels. As show in Fig. 11, which presents the spatial map of SMAP L3 SSM standard deviation from August 1, 2018, to July 31, 2020, most land surface exhibited relatively small annual SSM variation, allowing linear models to effectively capture the correlation between the variables.

Based on the total evaluation and performance comparison across different surface types, the pixel-by-pixel model outperformed the R-V-R model in both modeling and prediction accuracy. This finding aligns with previous studies [22,23], which reported a retrieval error of 0.07 cm^3cm^{-3}. However, in this study, the accuracy was observed to be 0.08 cm^3cm^{-3}. The disparity may be due to differences in quality control of inversion results and the sizes of the training and testing datasets. It is worth noting that the study by [22] utilized only five months of data, splitting it equally for modeling and testing.

The R-V-R algorithm is useful in areas where matched pairs for individual grid pixels are scarce, making it ideal for missions with limited number of GNSS-R receivers in orbit that struggle to obtain frequent, repeated observations. The R-V-R algorithm demonstrated its effectiveness in SSM retrieval during the BuFeng-1 mission [35]. In contrast, the pixel-by-pixel retrieval model requires extensive training data for development. To increase training samples and enhance prediction accuracy, alternative reference datasets from other satellite observations or model outputs can be employed based on the CYGNSS retrieval scenario. These datasets can play a crucial role in improving the accuracy of the pixel-by-pixel model. The research used ISMN in situ measurements for evaluation and demonstrated the success of the GNSS-R semi-empirical model in estimating SSM. However, discrepancies between the predicted and reference data may be due to calibration errors between the datasets, underscoring the importance of meticulous calibration procedures in ensuring the accuracy of GNSS-R data retrieval and modeling processes.

Previous scattering simulations have shown that densely vegetated areas and surface roughness significantly affect GNSS-R received reflected signal. While GNSS-R observations are also influenced by terrain, noise, and calibration errors, current correction methods for vegetation and surface roughness do not fully mitigate their impact on the GNSS-R SSM proxy. Errors in both the land cover-based linear model and the R-V-R retrieval model are largely due to spatiotemporal heterogeneity in surface vegetation and roughness. Additionally, this study uses SSM data from two microwave sensors: CYGNSS, which relies on active radar reflectometry, and SMAP, which uses passive radiometric measurements, both in the same microwave band. Two methods may sense different soil conditions, especially when moisture varies with depth, necessitating caution when combining data from both sensors. Incoherent radiometric sensing may better account for vegetation effects using a homogeneous layer, whereas radar is more sensitive to vegetation heterogeneity, such as differences in structure between forests and grasslands. The disparity between the land cover-based and R-V-R models highlights the importance of improving vegetation and roughness corrections and reducing uncertainty in auxiliary parameters. Future research should focus on developing pixel-level models or inversion techniques that bypass the need for auxiliary vegetation and roughness corrections, to better capture environmental complexities.

6. Conclusion

In conclusion, reliable and accurate methods for estimating SSM are essential for advancing our understanding of hydrological processes, optimizing agriculture practices, and gaining insights into the effects of climate change on water resources and ecosystems. This study aimed to enhance SSM estimation by developing semi-empirical retrieval algorithms using CYGNSS observations across various land cover types. Three enhanced semi-empirical inversion models, the pixel-by-pixel model, the land cover-based linear model, and the R-V-R model, were evaluated to assess their performance. The study focused on identifying optimal strategies for correcting diverse influencing factors, such as vegetation and surface roughness attenuation. Additionally, the study explored the efficacy of different fitting functions in the pixel-by-pixel model and the utilization of LUT correction in the various inversion models. The models were trained on data from August 2018 to July 2019 and validated on data from August 2019 to July 2020. The inversion results varied significantly among different land cover types, with wetter forest floors having slightly higher errors than other land cover types. Herbaceous plant cover, such as grasslands and croplands, had notably good inversion accuracy. Statistical results indicated that the CYGNSS observation coverage had an insignificant SSM variation for most land surfaces. The pixel-by-pixel model demonstrated the highest prediction accuracy with minimal bias, achieving an ubRMSD of 0.058 cm^3cm^{-3} and a correlation coefficient of 0.887. The land cover-based linear model performed the poorest with an ubRMSD of 0.082 cm^3cm^{-3} and a correlation coefficient of 0.756, while the R-V-R model outperformed the land cover-based linear model with an ubRMSD of 0.073 cm^3cm^{-3} and a correlation coefficient of 0.806.

The importance of enhanced SSM retrieval algorithms using satellite-based GNSS-R observations with consideration of surface conditions is emphasized in the study. The findings provide valuable insights for researchers and practitioners to improve SSM retrieval from GNSS-R data for different land surface types, which help develop more accurate and reliable SSM retrieval algorithms. Further research in this area can help advance our understanding of SSM dynamics and improve our ability to monitor and manage water resources in different land cover types.

Funding

This research was supported by the National Natural Science Foundation of China [grant number 42204014] .

CRediT authorship contribution statement

Zhounan Dong: Writing – original draft, Software, Methodology, Funding acquisition, Conceptualization. **Qingyun Yan:** Writing – review & editing, Resources, Data curation. **Shuanggen Jin:** Writing – review & editing, Supervision, Formal analysis. **Li Li:** Validation, Software. **Guodong Chen:** Writing – review & editing.

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.

Data availability

Data will be made available on request.

Acknowledgments

We would like to acknowledge the following institutions for providing valuable datasets and resources that supported this research: the National Oceanic and Atmospheric Administration (NOAA) for

providing CYGNSS data, the Boulder, Colorado USA and NASA National Snow and Ice Data Center (NSIDC) for providing SMAP L3 soil moisture products, the United States Geological Survey (USGS) for providing MCD12C1 data, and International Soil Moisture Network (ISMN) providing the in situ measurements. Without their contributions, this research would not have been possible.

References

[1] J. Peng, C. Albergel, A. Balenzano, L. Brocca, O. Cartus, M.H. Cosh, W.T. Crow, K. Dabrowska-Zielinska, S. Dadson, M.W.J. Davidson, P. de Rosnay, W. Dorigo, A. Gruber, S. Hagemann, M. Hirschi, Y.H. Kerr, F. Lovergine, M.D. Mahecha, P. Marzahn, F. Mattia, J.P. Musial, S. Preuschmann, R.H. Reichle, G. Satalino, M. Silgram, P.M. van Bodegom, N.E.C. Verhoest, W. Wagner, J.P. Walker, U. Wegmüller, A. Loew, A roadmap for high-resolution satellite soil moisture applications – confronting product characteristics with user requirements, Remote Sens. Environ. 252 (2021) 112162, https://doi.org/10.1016/j.rse.2020.112162.

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

[3] J.C. Price, The potential of remotely sensed thermal infrared data to infer surface soil moisture and evaporation, Water Resour. Res. 16 (1980) 787–795, https://doi.org/10.1029/WR016i004p00787.

[4] Y.H. Kerr, Soil moisture from space: Where are we? Hydrogeol J 15 (2007) 117–120, https://doi.org/10.1007/s10040-006-0095-3.

[5] M. Dobson, F. Ulaby, M. Hallikainen, M. El-rayes, Microwave dielectric behavior of wet soil-part II: dielectric mixing models, IEEE Trans. Geosci. Remote Sensing GE-23 (1985) 35–46, https://doi.org/10.1109/TGRS.1985.289498.

[6] E.G. Njoku, D. Entekhabi, Passive microwave remote sensing of soil moisture, J. Hydrol. 184 (1996) 101–129, https://doi.org/10.1016/0022-1694(95)02970-2.

[7] T.E. Ochsner, M.H. Cosh, R.H. Cuenca, W.A. Dorigo, C.S. Draper, Y. Hagimoto, Y. H. Kerr, K.M. Larson, E.G. Njoku, E.E. Small, M. Zreda, State of the art in large-scale soil moisture monitoring, Soil Sci. Soc. Am. J. 77 (2013) 1888–1919, https://doi.org/10.2136/sssaj2013.03.0093.

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

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

[10] H. Mao, D. Kathuria, N. Duffield, B.P. Mohanty, Gap filling of high-resolution soil moisture for SMAP/Sentinel-1: a Two-layer machine learning-based framework, Water Resour. Res. 55 (2019) 6986–7009, https://doi.org/10.1029/2019WR024902.

[11] M.P. Clarizia, C.S. Ruf, Wind speed retrieval algorithm for the cyclone global navigation satellite system (CYGNSS) mission, IEEE Trans. Geosci. Remote Sensing 54 (2016) 4419–4432, https://doi.org/10.1109/TGRS.2016.2541343.

[12] D. Comite, L. Cenci, A. Colliander, N. Pierdicca, Monitoring freeze-thaw state by means of GNSS reflectometry: an analysis of TechDemoSat-1 data, IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 13 (2020) 2996–3005, https://doi.org/10.1109/JSTARS.2020.2986859.

[13] E. Santi, S. Paloscia, S. Pettinato, G. Fontanelli, M.P. Clarizia, D. Comite, L. Dente, L. Guerriero, N. Pierdicca, N. Floury, Remote sensing of forest biomass using GNSS reflectometry, IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 13 (2020) 2351–2368, https://doi.org/10.1109/JSTARS.2020.2982993.

[14] Y. Jia, S. Jin, P. Savi, Q. Yan, W. Li, Modeling and theoretical analysis of GNSS-R soil moisture retrieval based on the random forest and support vector machine learning approach, Remote Sens. (Basel) 12 (2020) 3679, https://doi.org/10.3390/rs12223679.

[15] C.S. Ruf, S. Gleason, Z. Jelenak, S. Katzberg, A. Ridley, R. Rose, J. Scherrer, V. Zavorotny, The CYGNSS nanosatellite constellation hurricane mission, 2012 IEEE International Geoscience and Remote Sensing Symposium IEEE, Munich, Germany 2012 (2012) 214–216, https://doi.org/10.1109/IGARSS.2012.6351600.

[16] C. Chew, R. Shah, C. Zuffada, G. Hajj, D. Masters, A.J. Mannucci, Demonstrating soil moisture remote sensing with observations from the UK TechDemoSat-1 satellite mission, Geophys. Res. Lett. 43 (2016) 3317–3324, https://doi.org/10.1002/2016GL068189.

[17] H. Carreno-Luengo, A. Camps, J. Querol, G. Forte, First results of a GNSS-R experiment from a stratospheric balloon over boreal forests, IEEE Trans. Geosci. Remote Sensing 54 (2016) 2652–2663, https://doi.org/10.1109/TGRS.2015.2504242.

[18] A. Camps, M. Vall·llossera, H. Park, G. Portal, L. Rossato, Sensitivity of TDS-1 GNSS-R reflectivity to soil moisture: global and regional differences and impact of different spatial scales, Remote Sens. (Basel) 10 (2018) 1856, https://doi.org/10.3390/rs10111856.

[19] H. Carreno-Luengo, G. Luzi, M. Crosetto, Sensitivity of CyGNSS bistatic reflectivity and SMAP microwave radiometry brightness temperature to geophysical parameters over land surfaces, IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 12 (2019) 107–122, https://doi.org/10.1109/JSTARS.2018.2856588.

[20] M. Rahmani, J. Asgari, M. Asgarimehr, Soil moisture retrieval using space-borne GNSS reflectometry: a comprehensive review, Int. J. Remote Sens. 43 (2022) 5173–5203, https://doi.org/10.1080/01431161.2022.2128927.

[21] C.C. Chew, E.E. Small, Soil moisture sensing using spaceborne GNSS Reflections: comparison of CYGNSS reflectivity to SMAP soil moisture, Geophys. Res. Lett. 45 (2018) 4049–4057, https://doi.org/10.1029/2018GL077905.

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

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

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

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

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

[27] H. Kim, V. Lakshmi, Use of cyclone global navigation satellite system (CyGNSS) observations for estimation of soil moisture, Geophys. Res. Lett. 45 (2018) 8272–8282, https://doi.org/10.1029/2018GL078923.

[28] C. Chew, E. Small, Description of the UCAR/CU soil moisture product, Remote Sens. (Basel) 12 (2020) 1558, https://doi.org/10.3390/rs12101558.

[29] Y. Jia, S. Jin, Q. Yan, P. Savi, R. Zhang, W. Li, An effective land type labeling approach for independently exploiting high-resolution soil moisture products based on CYGNSS data, IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 15 (2022) 4234–4247, https://doi.org/10.1109/JSTARS.2022.3176031.

[30] O. Eroglu, M. Kurum, J. Ball, Response of GNSS-R on dynamic vegetated terrain conditions, IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 12 (2019) 1599–1611, https://doi.org/10.1109/JSTARS.2019.2910565.

[31] L. Dente, L. Guerriero, D. Comite, N. Pierdicca, Spaceborne GNSS-R signal over a complex topography: modeling and validation, IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 13 (2020) 1218–1233, https://doi.org/10.1109/JSTARS.2020.2975187.

[32] Y. Liu, Y. Yang, Advances in the quality of global soil moisture products: a review, Remote Sens. (Basel) 14 (2022) 3741, https://doi.org/10.3390/rs14153741.

[33] L. Gao, M. Sadeghi, A. Ebtehaj, J.-P. Wigneron, A temporal polarization ratio algorithm for calibration-free retrieval of soil moisture at L-band, Remote Sens. Environ. 249 (2020) 112019, https://doi.org/10.1016/j.rse.2020.112019.

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

[35] Z. Guo, B. Liu, W. Wan, F. Lu, X. Niu, R. Ji, C. Jing, W. Li, X. Chen, J. Yang, Z. Bai, Soil moisture retrieval using BuFeng-1 A/B based on land surface clustering algorithm, IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 15 (2022) 4680–4689, https://doi.org/10.1109/JSTARS.2022.3179325.

[36] M.J. Unwin, N. Pierdicca, E. Cardellach, K. Rautiainen, G. Foti, P. Blunt, L. Guerriero, E. Santi, M. Tossaint, An introduction to the HydroGNSS GNSS reflectometry remote sensing mission, IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 14 (2021) 6987–6999, https://doi.org/10.1109/JSTARS.2021.3089550.

[37] S. Gleason, C.S. Ruf, A.J. O'Brien, D.S. McKague, The CYGNSS level 1 calibration algorithm and error analysis based on on-orbit measurements, IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 12 (2019) 37–49, https://doi.org/10.1109/JSTARS.2018.2832981.

[38] O'Neill, P. E., S. Chan, E. G. Njoku, T. Jackson, R. Bindlish, and J. Chaubell. (2021). SMAP L3 Radiometer Global Daily 36 km EASE-Grid Soil Moisture, Version 8 [SPL3SMP]. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. doi: 10.5067/OMHVSRGFX38O. Date Accessed 3-15-2023.

[39] Friedl, M., Sulla-Menashe, D. (2022). MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 0.05Deg CMG V061 [MCD12C1]. NASA EOSDIS Land Processes Distributed Active Archive Center. Accessed 2023-3-20 from doi: 10.5067/MODIS/MCD12C1.061.

[40] W.A. Dorigo, A. Xaver, M. Vreugdenhil, A. Gruber, A. Hegyiová, A.D. Sanchis-Dufau, D. Zamojski, C. Cordes, W. Wagner, M. Drusch, Global automated quality control of in situ soil moisture data from the international soil moisture network, Vadose Zone J. 12 (2013) 1–21, https://doi.org/10.2136/vzj2012.0097.

[41] A. Camps, H. Park, M. Pablos, G. Foti, C.P. Gommenginger, P.-W. Liu, J. Judge, Sensitivity of GNSS-R spaceborne observations to soil moisture and vegetation, IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 9 (2016) 4730–4742, https://doi.org/10.1109/JSTARS.2016.2588467.

[42] B.J. Choudhury, T.J. Schmugge, A. Chang, R.W. Newton, Effect of surface roughness on the microwave emission from soils, J. Geophys. Res. 84 (1979) 5699, https://doi.org/10.1029/JC084iC09p05699.
[43] M.M. Al-Khaldi, J.T. Johnson, A.J. O'Brien, A. Balenzano, F. Mattia, Time-series retrieval of soil moisture using CYGNSS, IEEE Trans. Geosci. Remote Sensing 57 (2019) 4322–4331, https://doi.org/10.1109/TGRS.2018.2890646.
[44] T.J. Jackson, T.J. Schmugge, Vegetation effects on the microwave emission of soils, Remote Sens. Environ. 36 (1991) 203–212, https://doi.org/10.1016/0034-4257(91)90057-D.
[45] Z. Dong, S. Jin, G. Chen, P. Wang, Enhancing GNSS-R soil moisture accuracy with vegetation and roughness correction, Atmos. 14 (2023) 509, https://doi.org/10.3390/atmos14030509.
[46] J. Zeng, P. Shi, K.-S. Chen, H. Ma, H. Bi, C. Cui, Assessment and error analysis of satellite soil moisture products over the third pole, IEEE Trans. Geosci. Remote Sensing 60 (2022) 1–18, https://doi.org/10.1109/TGRS.2021.3116078.
[47] C. Yi, X. Li, J. Zeng, L. Fan, Z. Xie, L. Gao, Z. Xing, H. Ma, A. Boudah, H. Zhou, W. Zhou, Y. Sheng, T. Dong, J.-P. Wigneron, Assessment of five SMAP soil moisture products using ISMN ground-based measurements over varied environmental conditions, J. Hydrol. 619 (2023) 129325, https://doi.org/10.1016/j.jhydrol.2023.129325.
[48] X. Deng, L. Zhu, H. Wang, X. Zhang, C. Tong, S. Li, K. Wang, Triple collocation analysis and in situ validation of the CYGNSS soil moisture product, IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 16 (2023) 1883–1899, https://doi.org/10.1109/JSTARS.2023.3235111.