Quasi-Global Assessment of Deep Learning-Based CYGNSS Soil Moisture Retrieval

M M Nabi, *Student Member, IEEE*, Volkan Senyurek, Fangni Lei, Mehmet Kurum, *Senior Member, IEEE*, and Ali Cafer Gurbuz, *Senior Member, IEEE*

***Abstract*—A high spatial and temporal resolution global soil moisture product is essential for understanding hydrologic and meteorological processes and enhancing agricultural applications. Global navigation satellite system (GNSS) signals at L-band frequencies that reflect off the land surface can convey high-resolution land surface information, including surface soil moisture (SM). Cyclone global navigation satellite system (CYGNSS) constellation generates Delay-Doppler Maps (DDMs) that contain important Earth surface information from GNSS reflection measurements. DDMs are affected by soil moisture and other factors such as complex topography, soil texture, and overlying vegetation. Including entire DDM information can help reduce the uncertainty of SM estimation under different conditions along with remotely sensed geophysical data. This work extends our previously developed deep learning (DL) framework to a global scale by utilizing processed DDM measurements (analog power, effective scattering area, and bistatic radar cross-section) and ancillary data (elevation, slope, water percentage, soil properties, and vegetation water content). The DL model is trained and evaluated using the Soil Moisture Active Passive (SMAP) mission's enhanced SM products at 9-km resolution. This study comprehensively evaluates the DL model against publicly available CYGNSS-based SM products at a quasi-global scale. In addition to the typical comparison against in-situ measurements, a robust triple collocation technique is used to evaluate the DL-based SM product and other CYGNSS-derived SM products.**

***Index Terms*—Convolutional neural network (CNN), cyclone global navigation satellite system (CYGNSS), deep learning (DL), global navigation satellite system-reflectometry (GNSS-R), soil moisture active passive (SMAP), soil moisture retrieval, triple collocation (TC).**

I. Introduction

SOIL moisture (SM) is a key variable for understanding the Earth's near-surface land–atmosphere interactions among water, energy, and biogeochemical fluxes [1], [2], [3]. An accurate high-resolution SM characterization is essential for various applications, such as flood forecasting and agricultural water and crop management [4], [5]. It has been demonstrated that microwave remote sensing is a useful technique for estimating SM with global spatial coverage and relatively high temporal frequency after decades of exploration and development [6], [7].

Currently, several dedicated satellites have been launched for obtaining SM information from the Earth's surface. The National Aeronautics and Space Administration's (NASA) Soil Moisture Active Passive (SMAP) [8], and the European Space Agency's (ESA) Soil Moisture and Ocean Salinity (SMOS) [9] are two satellite missions that operate with L-band passive radiometers. SMOS, launched in 2009, provides surface SM with a temporal interval of 2–3 days at a spatial resolution of roughly 40 km. SMAP mission was launched in 2015, and was initially designed to map SM at 3 km by integrating both passive radiometer and radar signals. However, given the SMAP radar instrument failure in July 2015, SMAP has reverted to providing only radiometer-based SM retrievals at scales that are comparable to SMOS SM resolution. Both satellites have generated SM at the spatial resolution of about 25 to 40 km for many years. Sentinel-1, an additional ESA mission, is a synthetic aperture radar that operates in the C-band and can be used to produce SM at 1-km spatial resolution with a revisit time of 6–12 days [10], [11].

In December 2016, NASA launched the Cyclone Global Navigation Satellite System (CYGNSS) mission, which consists of a constellation of eight small satellites equipped with Global Navigation Satellite System-Reflectometry (GNSS-R) receivers. In addition to its primary mission of ocean wind monitoring and hurricane tracking, CYGNSS can facilitate quasi-global SM mapping with sampling frequencies ranging from sub-daily to daily [12], [13]. In contrast to the active and passive microwave platforms, the GNSS-R technique used by CYGNSS repurposes the GNSS signals for microwave remote sensing at L-band, representing a feasible approach for obtaining global SM at high resolution with relatively low cost [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24].

Several studies have demonstrated different approaches for SM estimation using CYGNSS. For instance, Eroglu et al. [18] developed an artificial neural network (ANN) retrieval algorithm with multiple CYGNSS features over limited international soil moisture network (ISMN) sites as a proof-of-concept using a tenfold validation method. Yang et al. [19] presented another

Manuscript received 7 February 2023; revised 6 May 2023; accepted 14 June 2023. Date of publication 20 June 2023; date of current version 29 June 2023. This work was supported in part by National Science Foundation under Grant 2047771, and in part by the USDA-ARS under Grant NACA 58-6064-9-007. *(Corresponding author: Ali Cafer Gurbuz.)*

M M Nabi, Mehmet Kurum, and Ali Cafer Gurbuz are with the Department of Electrical and Computer Engineering, and Information Processing and Sensing Lab, Mississippi State University, Mississippi State, MS 39762 USA (e-mail: mn918@msstate.edu; kurum@ece.msstate.edu; gurbuz@ece.msstate.edu).

Volkan Senyurek is with the Geosystems Research Institute, Mississippi State University, Mississippi State, MS 39762 USA (e-mail: volkan@gri.msstate.edu).

Fangni Lei is with the Eversource Energy Center, University of Connecticut, Storrs, CT 06269 USA (e-mail: fangni.lei@uconn.edu).

Digital Object Identifier 10.1109/JSTARS.2023.3287591

ANN model that was subsequently used to derive long-term and continuous SM maps over Mainland China.

Chew and Small [13] developed an SM product by calibrating CYGNSS reflectivity observations to SM retrievals using multilinear regression. A machine learning (ML) model was developed in [22] to estimate global scale SM by including more SM stations and incorporating an independent validation strategy. The in-situ SM datasets provided by the 170 ISMN locations over three years were used as the learning model reference. This study evaluated the model performance against SMAP observations at 9-km gridded resolution for different temporal scales within the CYGNSS coverage. In a recent study [25], an improved random forest (RF)-based retrieval model was developed with a 72 km $\times$ 72 km delineation method using SMAP SM product as the reference data. The study showed a good correlation with SMAP SM and in-situ measurements.

Even though one of the key measurements from CYGNSS is the Delay-Doppler Map (DDM) that results from the mapping of the received power from the observed surface to delay-Doppler space, the majority of previous ML-based studies utilized only handcrafted features from DDMs such as effective reflectivity obtained from peak reflected power, trailing edge slope (TES), and leading edge slope (LES). These approaches utilize the derivative features generated from DDMs as the main variables for retrieving SM information. However, in addition to SM content, topographical characteristics and overlying vegetation canopy also contribute to entire DDM. DDM carries much more information than just several derived features. Different processed DDM observables are available such as Analog Power, Effective Scattering Area, and Bistatic Radar Cross-section (BRCS) in the CYGNSS data. A recent study [26] shows that a single DDM (analog power) can be used to estimate soil moisture. In a recent study [27], we have demonstrated that different DDM derivatives, along with other ancillary data products, can be jointly used within a deep-learning (DL) framework to estimate SM. DL is a powerful tool to learn enhanced features directly from the images by using multiple convolutional or fully connected layers. We have demonstrated that using multiple DDM images jointly with static or time-varying ancillary data within a DL framework can lead to better SM estimation performance with higher correlations [27]. Our previous analysis was conducted only over the Continental United States (CONUS) and shown benefit of DL directly using DDM images. However, a global analysis and detailed evaluation of the proposed method are conducted in this study to observe more varied SM dynamics across the world. In this work, the main goal is to extend the DL-based SM retrieval method to the global scale to extract global SM dynamics by generating dynamic features from DDMs and also evaluate the proposed global retrieval algorithm with other existing CYGNSS-based SM data products. We utilize a DL framework that uses SMAP-enhanced SM data product at 9-km spatial resolution as a label to build the relationship between CYGNSS observations and the SM estimate. In order to evaluate the performance of the DL approach, fivefold and year-based cross-validation approaches are used. Moreover, a robust and independent evaluation technique called triple collocation (TC) [28], [29] is utilized for evaluating multiple SM products at the quasi-global scale. TC is a technique that has been widely used in evaluating random errors in large-scale remote sensing or model products [30]. Kim and Lakshmi [15] presented that TC can be used as an independent evaluation method specifically for CYGNSS-based SM retrieval. In this study, an extended TC (ETC) technique is applied to characterize the correlation between the product and the theoretical truth [31]. The contributions of this article are summarized as follows.

1) We extend our previously proposed DL framework with multiple convolutional and fully connected neural network layers into an improved quasi-global SM product. Multiple DDM derivatives with ancillary geophysical data (i.e., slope, elevation, soil properties, vegetation water content, and normalized difference vegetation index) relevant to SM estimation are integrated into this framework.
2) Rigorous spatio-temporal analyses are presented to demonstrate the capability and limitation of the proposed DL approach.
3) We have published a new DL-based CYGNSS SM product at https://ssm.hpc.msstate.edu/ and this study compares it with other publicly available CYGNSS-based SM products at a global scale. To compare different remote sensing-based SM products, the ETC method is utilized to find the correlation coefficient concerning the unknown SM value.

The rest of this article is organized as follows. Section II summarizes CYGNSS, SMAP, ISMN sites, and other ancillary data needed for the DL framework. Details of the DL approach and methodologies are described in Section III. Results are presented in Section IV where CYGNSS SM retrievals are evaluated at selected regional, and quasi-global scales. The findings, difficulties, and implications for further research are discussed in Section V. Finally, Section VI concludes the article.

II. Dataset

In order to develop an efficient DL-based retrieval model for global surface SM mapping with CYGNSS measurements, various land surface products must be leveraged to characterize the underlying surface conditions. Only DDM information with DL-model will not be sufficient to extract the underlying surface condition as SM dynamics change spatially and temporally. So, different static and time-varying land surface data need to be incorporated along with DDM images. The subsections below discuss briefly the selected input data sources for the retrieval process, including the SMAP SM data as labels and other ancillary inputs that link the GNSS-R sensitivity to SM. Different quality control strategies and approaches to integrate multiresolution datasets are applied to ensure a consistent and accurate SM estimation.

A. Data for DL Model

In this study, the CYGNSS Level-1 (L1) version 2.1 product from March 18, 2017 to June 30, 2021 is used, which is available at the NASA Physical Oceanography Distributed Active Archive Center (PO.DAAC, https://podaac.jpl.nasa.gov/). CYGNSS provides one of the important measurements in the

L1 product, called DDM, which maps received power to a delay and doppler space spread by the observed surface. DDMs are processed for nonsurface related parameters by inverting the CYGNSS forward-scattering model and obtaining the surface's effective scattering area, and BRCS images [32]. Similar to our previous studies [27], [33], three types of processed DDMs (analog power, effective scattering area, and BRCS) are considered the primary inputs of the DL model.

Additionally, peak reflectivity and incidence angle of specular points are also added as ancillary inputs from the CYGNSS L1 dataset. The peak reflectivity is computed using the peak value of each DDM [13]. To facilitate the retrieval process, it is essential to obtain a prior knowledge of land surface conditions, especially topography, vegetation, and a fraction of open water bodies in pixels. Here, multiple features are derived from the ancillary datasets and used in the DL-based model, including vegetation water content (VWC), normalized difference vegetation index (NDVI), elevation, slope, water percentage, soil clay ratio, and soil silt ratio. The 16-day composite NDVI from Moderate Resolution Imaging Spectroradiometer is used to characterize vegetation conditions, and can be found in NASA Land Processes Distributed Active Archive Center (https://lpdaac.usgs.gov/products/myd13a1v006/). This NDVI data is spatially averaged to 3 km from its original 500-m resolution. VWC is computed using the NDVI and Land Cover Type (MCD12Q1) products using the same lookup table method as the SMAP VWC product [34]. The surface elevation data is collected using the GTOPO30 (1-km resolution) global digital elevation model (https://doi.org/10.5066/F7DF6PQS). Soil clay and silt ratios data are obtained from the Global Gridded Soil Information (SoilGrids) [35]. A 30-m Global Surface Water Dataset from the Joint Research Centre (GSW-JRC) [36] is utilized to indicate the presence of a surface inland water body. More details about the ancillary data for the retrieval can be found in our previous study [27].

The DL-based SM retrieval methodology is trained and evaluated using the SMAP Enhanced L3 Radiometer Global Daily 9-km EASE-Grid SM product. SMAP collects brightness temperature data with an L-band microwave radiometer to generate SM estimates. In addition to the standard SMAP SM product projected at 36-km resolution, a 9-km enhanced grid product is also created using the Backus–Gilbert optimal interpolation technique [37]. To obtain a daily SM map with sufficient SMAP data samples, the descending (a.m.) and ascending (p.m.) overpasses are combined. SM retrieval quality flags are included in the SMAP product to indicate whether or not SM retrieval is recommended and used for masking out SM retrievals with low quality. The SMAP data is publicly available and can be obtained from the National Snow and Ice Data Center (NSIDC) at https://nsidc.org/data/SPL3SMP_E/versions/3.

B. SM Data for Independent Validation

A robust evaluation of a learning-based data product requires using independent datasets for cross-validation because typical learning-based models are often impacted by the bias that existed in the training process. The ISMN dataset is one of the key datasets used for ground-based independent evaluation of the developed DL model in this study. ISMN is an integrated platform providing in-situ SM data from various networks and can be accessed from (http://ismn.geo.tuwien.ac.at).

Daily averaged SM data from 170 selected ISMN stations worldwide within the CYGNSS coverage is used to evaluate the DL model performance. ISMN sites above 2000-m altitude are not considered for comparison as CYGNSS measurements for high altitudes are unreliable for CYGNSS version 2.1. Detailed information about ISMN was reported in [38] and [39].

Besides the ground-based sparse networks validation, microwave remote sensing SM data from different satellite sources and land surface modeling products can provide additional insights for understanding the global-scale SM retrieval accuracy. Particularly, the C-band Advanced SCATterometer (ASCAT) SM product [40] is used as an independent validation dataset as compared to the passive microwave SMAP data. ASCAT SM retrievals are generated based on the change detection method developed by the Vienna University of Technology (TU-Wien) [40]. The spatial resolution of this product is 25 km with a grid spacing of 12.5 km and is resampled using the nearest-neighboring approach onto the EASE-Grid 2.0 36-km resolution. Daily averages from both a.m. and p.m. overpasses are generated from January 2017 to December 2021. SM retrievals obtained over frozen soil conditions are filtered out for quality control. In addition, the NOAH model-based SM data from the Global Land Data Assimilation System (GLDAS) [41] is also included for the independent analysis. For the GLDAS-2 NOAH v3.3 data product, the three hourly SM data is temporally averaged to daily values and spatially regridded from 0.25° to 36 km using the nearest neighboring method. SM estimates from the soil profile's top layer (0–0.1 m) are extracted.

C. Quality Control Mechanisms

This study uses CYGNSS observations from March 2017 to June 2021. Before performing SM retrieval, critical screening for the quality of CYGNSS data in underlying land surface conditions is required. There are several specific flags such as S-band powered up, black-body DDM, DDM test pattern, substantial spacecraft attitude error, poor confidence GPS EIRP estimate are applied in the CYGNSS data [14], [42]. CYGNSS observations above 600 m from the surface before December 2017 are masked out because of the altitude limitation of CYGNSS L1 data during this specified period [22]. To avoid noisy DDMs, observations with an incidence angle greater than 65° are excluded [17]. Standard CYGNSS flags are applied to remove some problematic DDMs found in CYGNSS data. Additionally, some DDM images provide no value for effective scattering area. It generally happens when the specular point bin zero-based Doppler column is less than 4 or greater than 6 [32]. These DDM images are also removed as part of data-quality control before applying it to the DL-model. It is also important to normalize the DDM images for numeric stability before applying them to the DL framework. Normalization is performed for all three types of DDMs by computing the mean and standard deviation for all pixels and scaling these values to attain zero

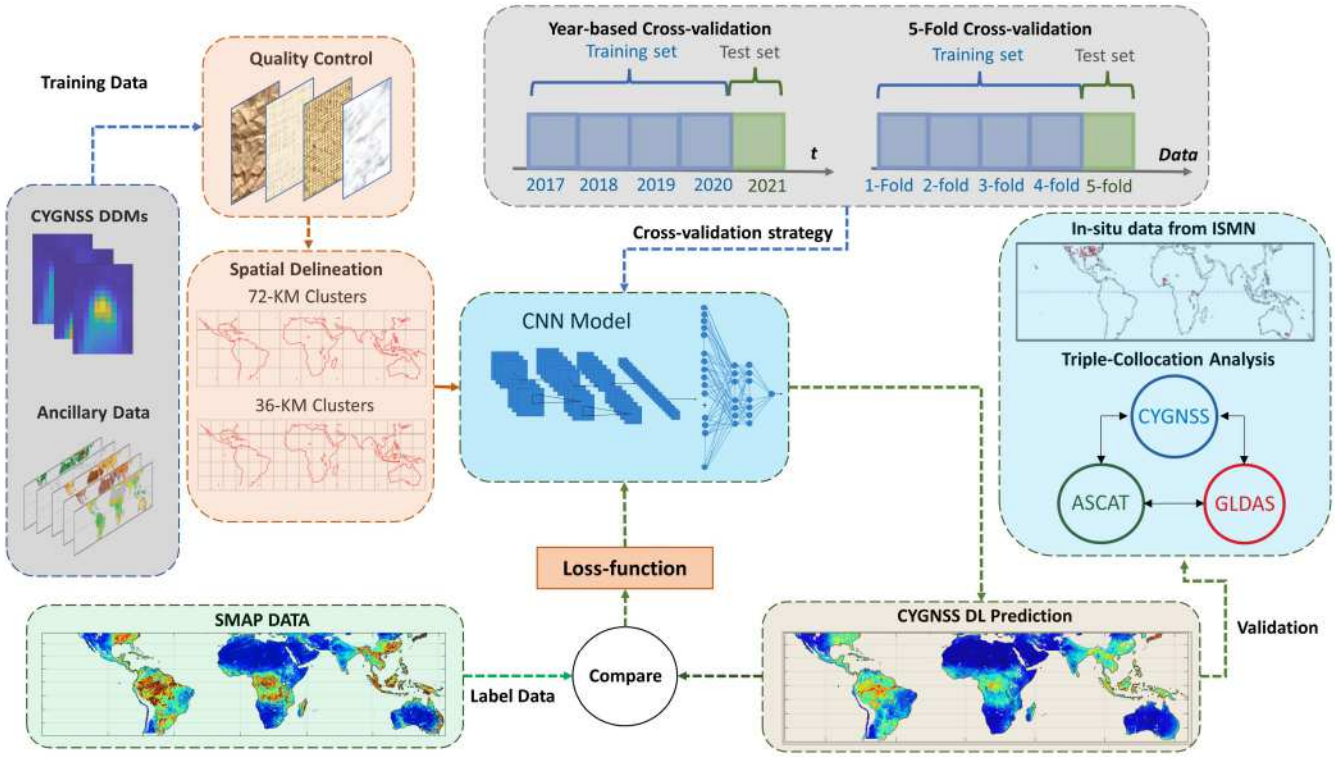

Fig. 1. Schematic framework of the DL-based SM retrieval model and validation process.

mean/unit variance. As part of SMAP quality checks, grid pixels under frozen conditions with a land surface temperature less than 273.15 K or with open water fraction greater than 10% are excluded [43]. If the surface water within the CYGNSS region is sufficiently large, SM retrieving near water bodies becomes impracticable due to the highly strong coherency over water surfaces [44]. A CYGNSS observation is removed if more than 2% of the 9 km grid is covered by permanent or seasonal water. For each 3-km block within 9-km grid, the water percent is calculated as the percentage of 30-m grids falling into the 3-km box that have either permanent or seasonal water presence [36].

III. Methods

Impacted by varying land surface conditions, such as vegetation, topography, and soil properties, the relationship between SM and the CYGNSS observables can be a complex nonlinear function. Our main goal of this current study is to build a DL model that can predict SM information from an unknown relationship. The data-driven approach, specifically the convolutional neural network (CNN) has been widely used in computer vision applications and is shown to capture essential features directly from images for the classification/regression tasks [45], [46]. Applying CNN to CYGNSS DDMs for SM retrieval is particularly attractive as it has the capability to learn directly from the DDM itself. Some existing approaches have been developed using derivative features calculated from DDM to estimate SM using ANN [18], linear regression [20], or RF [22]. Although these models produce promising results, we hypothesize that a CNN-based model can extract additional features that will further improve SM estimation, resulting in a better quality CYGNSS-based SM product. The complex information contained in the entire 2D DDM is useful in various conditions, and the DL technique is the state-of-the-art approach to retrieving features from DDMs. Our previous studies [27], [33] has demonstrated that the DL technique has a better SM estimation performance when trained and validated using the SMAP SM data over CONUS. In this paper, we extend our previous approach to the global scale within the CYGNSS coverage and evaluate DL-based SM estimation via multiple-scale assessment approaches, i.e., against in-situ measurements and other global SM products. CNN is the primary DL framework, with primary inputs from various types of CYGNSS-processed DDM images and ancillary data. It is a supervised learning framework that maps a set of input data to an SM value, which is the proposed architecture's final output. The datasets used to train and test the developed model come from CYGNSS, SMAP, and ancillary data sources over the CYGNSS coverage from March 2017 to June 2021. The overall DL framework with calibration and validation strategies are illustrated in Fig. 1, and the following subsections provide detailed methodology descriptions.

A. DL Framework

In our previous studies [27], [33], we demonstrated that the DL could be used to estimate SM using CYGNSS DDMs and ancillary data. The DL architecture comprises three major components: 1) convolutional layers; 2) concatenation layers; and

3) densely connected layers. A detailed explanation of the DL model can be found in our previous study [27]. First, the three types of processed DDMs (analog power, effective scattering area, and BRCS) together with ancillary data pass through several quality control mechanisms described in Section II-C. The generic approach in a DL model could be to train a single model that will predict SM at any given location for the whole world. Since a single model predicts SM everywhere, memory requirements are low—only 1 model is learned. However, the model should be able to address the whole complexity and variations over the world, and training a model with a huge size of dataset is highly computationally complex. Another possibility we proposed and tested in our previous paper [27] is to divide the region into smaller size clusters and learn a different DL model for each cluster using the data from that cluster region. In this case, since variations within smaller regions are less, it is easier for DL models to learn the DDM–SM relation but creating clusters increases the number of models to be stored and each model should be learned over a smaller dataset corresponding to the cluster region. Hence, there is a tradeoff between SM prediction performance, memory, and training requirements for DL models for different size regions which were analyzed in [27]. Fig. 1 shows schematic grid meshes on top of the CYGNSS coverage with two different lengths of the box, i.e., 36 km and 72 km. Hereafter, these two stratification methods are called 36-km and 72-km clusters. For the 72 km and 36-km stratification, we have approximately 16 000 and 64 000 separate clusters, respectively. 72-km clusters have been determined to be the best in the tradeoff analysis. We have shown the global analysis, and performance evaluations depending on this cluster size in this study.

After clustering, the DL-based model is trained with CYGNSS and ancillary data as inputs and SMAP-enhanced SM product as the reference. Particularly, each cluster has a separate model, and each model is validated using different cross-validation strategies, such as the fivefold cross-validation and year-based cross-validation. SM predictions are generated via the DL-trained model, which correlates the inputs with the label SM data through nonlinear relations and learned parameters. A root-mean-square error (RMSE) loss is defined as the loss between the predicted and labeled SM. The Root Mean Squared Propagation (RMSprop) is used as the optimizer, where the initial learning rate is set as 0.01. An early stopping criterion is set to reduce the computational time, which means if the model performance doesn't improve on the validation dataset within ten epochs, the system will stop training. Finally, the predicted results are validated using different independent validation methods.

B. Evaluation and Validation Methods

The K-fold approach is a popular and widely used validation technique for assessing a model's performance. This method ensures the separation of training and test data and tests each fold. Fivefold cross-validation is used over the dataset at 9-km resolution from March 2017 to June 2021. The performance of the DL model is also evaluated via the year-based cross-validation approach, where a trained model from several years of data is tested on a completely different year's data. In both validation methods, the performance metrics are computed for all available grids and the SMAP-recommended grids. In terms of model evaluation, several metrics are utilized to assess a model quantitatively. The most commonly used evaluation metrics for SM estimation are RMSE, unbiased RMSE (ubRMSE), and correlation coefficient (R-value). Additionally, the root-mean-square difference (RMSD) is also used to compute the differences between DL-based CYGNSS and the SMAP SM product as the label SMAP SM data also contain measurement errors [47]. The term "RMSE" is commonly used for in-situ evaluation because those measurements are regarded as ground truth data for SM. In our case, RMSE and ubRMSE metrics are used when comparing remote sensing products with in-situ SM stations. For independent validation at the global scale, ETC is utilized and detailed in the next Section III-C.

C. Triple Collocation (TC)

TC is widely used for evaluating large-scale remote sensing products by incorporating a minimum of three mutually independent measurement systems. Assuming each independent product is linearly related to the "Truth" and measurement errors are orthogonal, classical TC can be used to estimate error variances of three products by choosing one product as the reference. Besides, an ETC is developed to estimate the correlation coefficient of the measurement system with respect to the unknown target [31] and is used in this work.

We begin with an affine error model [48], which is commonly used in the literature on TC for relating data to a geophysical variable.

$$X_i = X_i' + \varepsilon_i = \alpha_i + \beta_i t + \varepsilon_i. \tag{1}$$

In (1), X_i $(i \in \{1,2,3\})$ are three collocated measurements from independent systems that are linearly related to the true underlying value t with additive random errors ε_i. Here, X_i, X_i', t and ε_i are all random variables. α_i and β_i represent the ordinary least squares (OLS) intercepts and slopes, respectively. The random error variance [31] can be derived as

$$\sigma_\varepsilon = \begin{bmatrix} \sqrt{Q_{11} - \frac{Q_{12}Q_{13}}{Q_{23}}} \\ \sqrt{Q_{22} - \frac{Q_{12}Q_{23}}{Q_{13}}} \\ \sqrt{Q_{33} - \frac{Q_{13}Q_{23}}{Q_{12}}} \end{bmatrix} \tag{2}$$

where σ_ε represents the error standard deviation for three measurements system and Q shows the covariance matrix. To characterize the relationship between product X and the theoretical truth, an ETC-based correlation coefficient was used [31]. Here, ρ_{t,X_i} is the correlation coefficient between t and X_i. The correlation coefficients can be derived as

$$\rho_{t,X} = \pm \begin{bmatrix} \sqrt{\frac{Q_{12}Q_{13}}{Q_{11}Q_{23}}} \\ sign(Q_{13}Q_{23})\sqrt{\frac{Q_{12}Q_{23}}{Q_{22}Q_{13}}} \\ sign(Q_{12}Q_{23})\sqrt{\frac{Q_{13}Q_{23}}{Q_{33}Q_{12}}} \end{bmatrix} \tag{3}$$

where ρ_{t,X_i} are corrected up to a sign ambiguity. In an actual scenario, it is expected that the measurement systems will almost always have a positive correlation to the unobserved truth.

The ρ_{t,X_i} keeps important additional information that is not included in σ_{ε_i} in the original TC analysis. In this paper, the correlations are used to compare different SM products generated from CYGNSS via different approaches. The evaluation using TC analysis for satellite-based SM data products is based on few assumption: 1) SM estimations are linearly related to the "true" SM value; 2) errors for each SM retrievals are uncorrelated with true SM; and 3) errors within each selected triplet are uncorrelated with each other [49]. In this work, the active microwave SM product from the Advanced SCATterometer (ASCAT) and the numerical model-based Global Land Data Assimilation System Version (GLDAS) NOAH SM products are included to complete the triplets in TC analysis. The main purpose of TC analysis is to compare the performance of different CYGNSS products. Since the ASCAT and GLDAS products are fixed in the triplet, the difference in representative depths can be neglected for this purpose. The primary limitation for conducting TC analysis is the relatively coarse spatial resolution of ASCAT and NOAH datasets. Therefore, the analysis is mainly performed at 36 km using collocated data from March 2017 to June 2021. However, to enhance the TC analysis at a higher resolution, both ASCAT and NOAH SSM products are resampled from their original spatial resolution to 9 km using the nearest neighboring approach.

IV. Results

In this section, the DL model's performance is discussed based on different cross-validation scenarios. Qualitative and quantitative performances under different validation strategies of the DL model are provided. The K-fold and year-based cross-validation is used to evaluate global performance evaluation against the SMAP SM within the CYGNSS coverage across the world. Then, the DL model performance is also compared against publicly available SM data products such as the University Corporation for Atmospheric Research (UCAR) SM product [13], and Mississippi State University Geosystems Research Institute (MSU-GRI) SM product [25], both of which use SMAP as reference for training. Besides these analyses, the DL predicted result is compared with the SM station (ISMN sites) and presents the temporal variation of different SM products for different stations. Temporal variations for several selected regions of the world are also demonstrated to examine the region-wise performance of different models. Land-cover information is also included in this section to demonstrate the performance of the DL approach for each land-cover type. Results based on ETC are also shown here, which serve as an independent technique for characterizing the product's accuracy. Details of each performance analysis approach are provided in the below subsections.

A. Quasi-Global Performance Results

In this subsection, global SM predictions generated from the DL-based approach are compared with SMAP. Previous studies [25], [27] presented clustering approaches for SM retrieval using different algorithms. Lei et al., 2022 [25] presented three different clusters such as 9-km, 72-km, and 288-km where 72-km outperformed all other clusters. In our previous work [27], we also demonstrated 36 km, 72 km, 144 km, and global clusters, where 36 km and 72 km perform similarly. The main difference between these two works is the algorithmic difference. Based on these knowledge, two different cluster cases are examined, i.e., 36-km and 72-km clusters. Both cluster methods perform very similarly, with a significant difference in computational complexity. The 36-km case has more clusters that require more memory and time to train models. On the other hand, 72-km clusters require fewer models to train, so the computational complexity is less than the 36-km clustering method. So, the 72-km cluster is chosen as the main approach, and the rest of the results in this article are generated based on the 72-km cluster. Other cluster sizes were analyzed and found to perform poorly compared to 36-km and 72-km clusters [27]. It is worth noting that a sample size limitation of 300 samples is set for the 72-km cluster. Therefore, grids with fewer than 300 samples are discarded from the training process.

Fig. 2 demonstrates ubRMSD and correlation coefficient maps generated using the 72-km cluster. The ubRMSD map [Fig. 2(a)] is the error between daily averaged SMAP and CYGNSS SM retrievals for each 9 km $\times$ 9 km grid. Regions generally flagged as poor quality by SMAP are shown to have relatively higher ubRMSD errors. On the other hand, most of the remaining regions show a high agreement with SMAP observations with lower ubRMSD results. The correlation coefficient map [Fig. 2(b)] shows high correlations in most of the world except the Sahara and Amazon regions. Slight degradations are found in regions with dense vegetation due to the masking effects of dense vegetation and arid conditions due to lower dynamic range.

B. Model Performance via Year-Based Cross-Validation

It is also necessary to assess the performance of the DL method under a year-based cross-validation scenario. This more challenging cross-validation will indicate how the developed model can be generalized for future years. Table I shows the statistics generated based on different validation years, which means the model is first trained using data from several selected years and then tested using left-over years. For example, for the validation of 2021, the DL model is trained using data from 2017 to 2020 and the learned model is tested with the data of year 2021. Similarly, if year 2019 is selected for prediction, the model is trained using 2017, 2018, 2020, and 2021 years data. The year-based cross-validations are evaluated for all grids and SMAP-recommended grids seperately. For the 2021 validation case, ubRMSD is 0.0564 m^3m^{-3} and the correlation coefficient is 0.91 when considering all available grids. When only taking into account SMAP recommended grids, ubRMSD is 0.0403 m^3m^{-3} and the correlation coefficient is 0.88. A higher correlation coefficient is obtained for all grids than the SMAP-recommended grids due to the higher number of data samples.

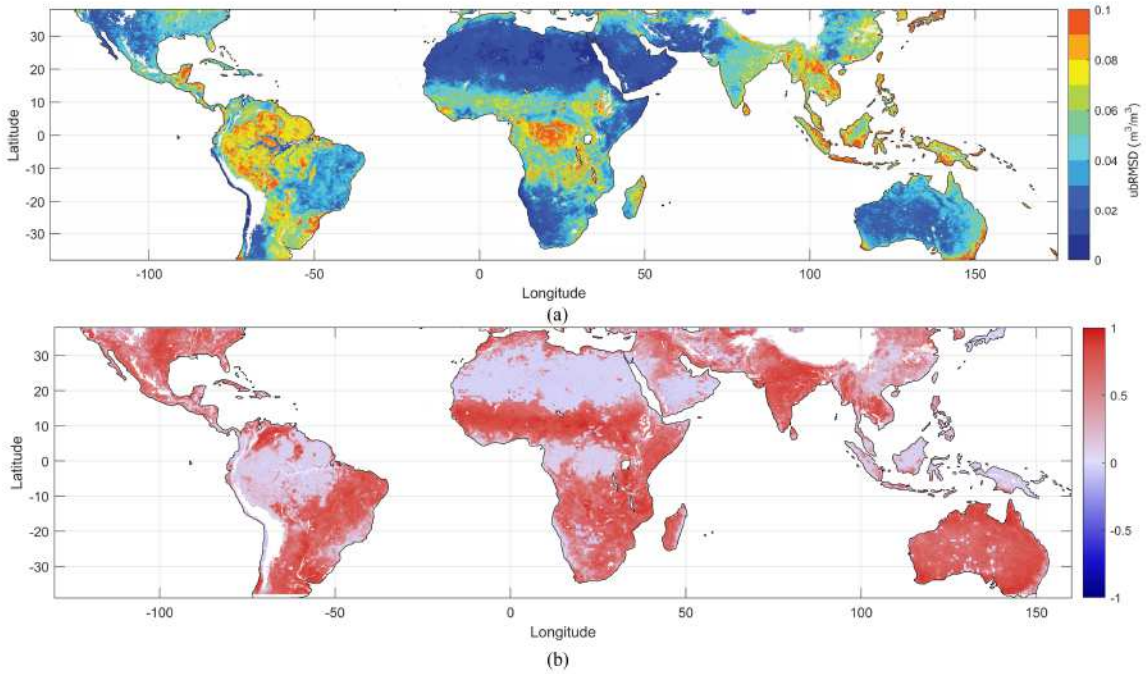

Fig. 2. (a) Unbiased root-mean-square difference (ubRMSD), and (b) correlation coefficient map between CYGNSS-based model prediction and SMAP retrievals daily averaged for each 9 km × 9 km grid over available data from 2017 to 2021. (a) ubRMSD (9 km × 9 km). (b) Correlation coefficient maps.

TABLE I
YEAR-BASED CROSS-VALIDATION RESULTS USING THE 72-KM CLUSTER APPROACH CONSIDERING ALL AVAILABLE GRIDS AND SMAP RECOMMENDED GRIDS

Training years	Validation years	Grid Selection	RMSD (m^3m^{-3})	bias (m^3m^{-3})	ubRMSD (m^3m^{-3})	R[-]
2019, 2020, and 2021	2017+2018	All grids	0.0567	0.0051	0.0565	0.92
		SMAP-recommend grids	0.0414	0.0091	0.0404	0.90
2017, 2018, 2020, and 2021	2019	All grids	0.0559	0.0028	0.0558	0.92
		SMAP-recommend grids	0.0396	0.0057	0.0392	0.90
2017, 2018, 2019, and 2021	2020	All grids	0.0602	-0.0027	0.0602	0.90
		SMAP-recommend grids	0.0441	-0.0029	0.0440	0.87
2017, 2018, 2019, and 2020	2021	All grids	0.0566	-0.0041	0.0564	0.91
		SMAP-recommend grids	0.0405	-0.0030	0.0403	0.88

TABLE II
OVERALL PERFORMANCE COMPARISON OF DIFFERENT SM PRODUCTS AGAINST SMAP SM PRODUCTS (FROM MARCH 2017 TO DECEMBER 2020)

Methods	**Grid Selection**	**RMSD** (m^3m^{-3})	**bias** (m^3m^{-3})	**ubRMSD** (m^3m^{-3})	**R [-]**
CYGNSS-DL Product	All grids	0.0488	0.0035	0.0487	0.94
	SMAP-recommend grids	0.0369	0.0056	0.0365	0.92
UCAR Product	All grids	0.0648	0.0102	0.0640	0.89
	SMAP-recommend grids	0.0457	0.0092	0.0446	0.88
MSU-GRI Product	All grids	0.0538	-0.0128	0.0522	0.93
	SMAP-recommend grids	0.0400	-0.0077	0.0392	0.90

Note: All products are resampled to 36 km for comparison.

It is observed that the correlation coefficient slightly decreases from 2017 to 2021. A slight variation of ubRMSD over the years is seen in both all grids and SMAP recommended gird cases. The year-based cross-validation shows good potential when data is trained from 2017 to 2020 and tested in 2021 as low errors with high correlations are observed. It is important to note that year-based cross-validation will be valid if the environmental condition does not change significantly. If the environmental condition is changed significantly in the future, then it will have an impact on overall data dynamics for SMAP, CYGNSS, and other ancillary data.

C. Performance Comparison Among SM Products

This section compares publicly available CYGNSS-based SM data products with respect to SMAP SM data product. Two different publicly available CYGNSS-based SM data products [13], [25] are considered that are trained using SMAP standard SM (Table II). The overall performance shows that CYGNSS-DL outperforms the existing approaches; UCAR [13] and MSU-GRI [25] products. If all grids are considered, CYGNSS-DL provides the lowest ubRMSD of 0.0487 m^3m^{-3} and the highest correlation coefficient (0.94) when compared to the other

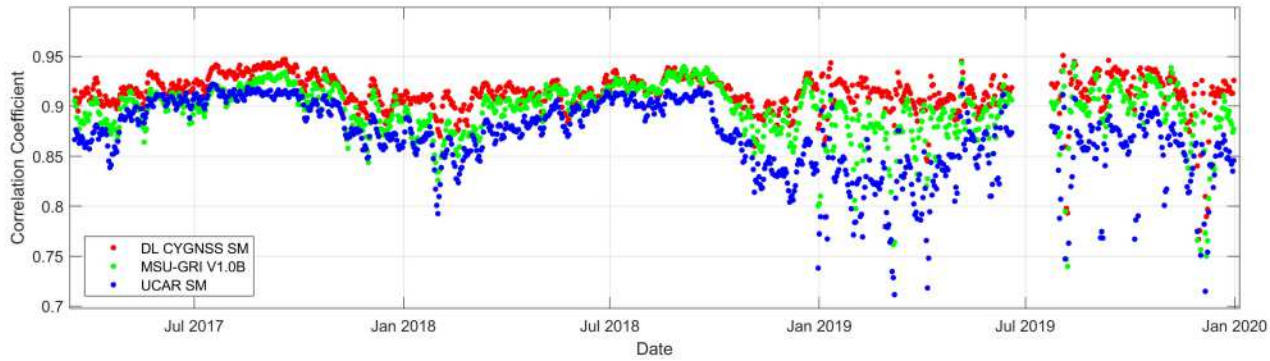

Fig. 3. Temporal evaluation of the spatial correlation between different model-based SM prediction and SMAP retrievals (from March 2017 to December 2020). CYGNSS-DL SM is presented as red dots, MSU-GRI V1.0B SM as green dots, and UCAR SM as blue dots.

two methods. For SMAP-recommended grids, the proposed approach achieves ubRMSD of 0.0365 m^3m^{-3} and the correlation coefficient of 0.92. Note that while the proposed approach learn features directly from DDMs, both UCAR and MSU-GRI products mainly use several specific features from DDMs as the primary input of their retrieval approaches. For the SMAP recommended grids, the UCAR product reaches an ubRMSD of 0.0446 m^3m^{-3} and a correlation coefficient of 0.88, and the MSU-GRI product achieves 0.0392 m^3m^{-3} ubRMSD and 0.90 correlation coefficient. These results clearly show that improved performance can be obtained when exploiting the full DDM for SM retrieval.

Besides the overall performance statistics shown in Table II, daily spatial correlations between different CYGNSS products and SMAP are compared within the time frame from March 2017 to December 2020. Specifically, the daily correlation is calculated as the correlation coefficient between CYGNSS and SMAP SM maps on a particular day. Fig. 3 illustrates the spatial correlations between the SMAP SM and the CYGNSS SM products; the proposed CYGNSS-DL, MSU-GRI, and UCAR. The CYGNSS-DL SM product provides consistently higher correlations with the SMAP SM as compared to the other two CYGNSS products. The average correlation for CYGNSS-DL SM is more than 0.90 within the study time frame. Slightly higher correlations are found during the warm seasons. On the other hand, MSU-GRI provides slightly lower correlations than the DL method. UCAR SM shows similar but slightly lower correlations from 2017 to 2018 with an average correlation greater than 0.85. However, the correlations quickly drop after mid 2018. Similarly, all three products have degraded correlations after 2018, while the proposed CYGNSS-DL sustains comparably higher correlation levels.

D. Spatial and Temporal Analysis

In this subsection, SM temporal variations of three different CYGNSS SM products and SMAP data are examined to conduct the region-wise comparison for selected regions. Since different SM products are generated using different methods under different assumptions, SM predictions can vary over different parts of the world. In order to show the region-wise SM variation, a 225-km $\times$ 225-km box is posted on several parts of the globe, i.e., midwest USA, India, and the West Africa (Ghana and Togo) regions. These regions are selected to be near the locations presented in [22]. Spatial maps are shown for SMAP and CYGNSS-DL products and daily global averaged SM values are calculated for all SM products.

First, we consider a small region in the midwest USA, which is marked as the black box in Fig. 4(a) and (b). The SM maps are generated using the daily averaged data for four years (2017–2020) time frame applying a fivefold 72-km cluster DL method. Fig. 4(c) shows the temporal variations within this selected region. SMAP SM shows strong SM dynamics (black line) and all three CYGNSS SM products generally follow the patterns. However, all three CYGNSS products provide a much smoother dynamic range as compared to SMAP. Overall, CYGNSS-DL (red-dotted line) has a slightly greater temporal variation than MSU-GRI and UCAR products. The UCAR SM (marked with a blue-dotted line) product generally overestimates during the dry-down period. UCAR product tries to capture the dynamic range rather than providing a mean value for the time frame. Specifically, in the dry-down period, it can capture the trend with SMAP SM data. Besides, MSU-GRI (green-dotted line) shows low SM variation with respect to the SMAP SM, and it shows an approximately averaged SM of 0.3 m^3m^{-3}. The MSU-GRI product shows very low SM variation during the growing season, whereas all other SM products provide greater SM variations.

Second, the western part of India is considered, which dominantly features croplands in an equatorial winter-dry climate zone. A high spatial correlation can be found with the 9-km SMAP and CYGNSS SM maps in Fig. 5(a) and (b). CYGNSS-DL tends to overestimate SM in some small areas when compared to SMAP. However, this area is also flagged by SMAP's Retrieval Quality Flag (RQF) due to high uncertainty for SM retrieval. SM temporal variations of the selected sub-region from different products are presented in Fig. 5(c). High SM dynamic ranges are seen, given the seasonal standing surface water and dense vegetation during the growing seasons in this region. Abrupt changes in SM during monsoon seasons are clear from the temporal trend plot. Aside from these monsoon events typically occurring from June to September, a high correlation between CYGNSS-DL and SMAP SM retrievals can be seen.

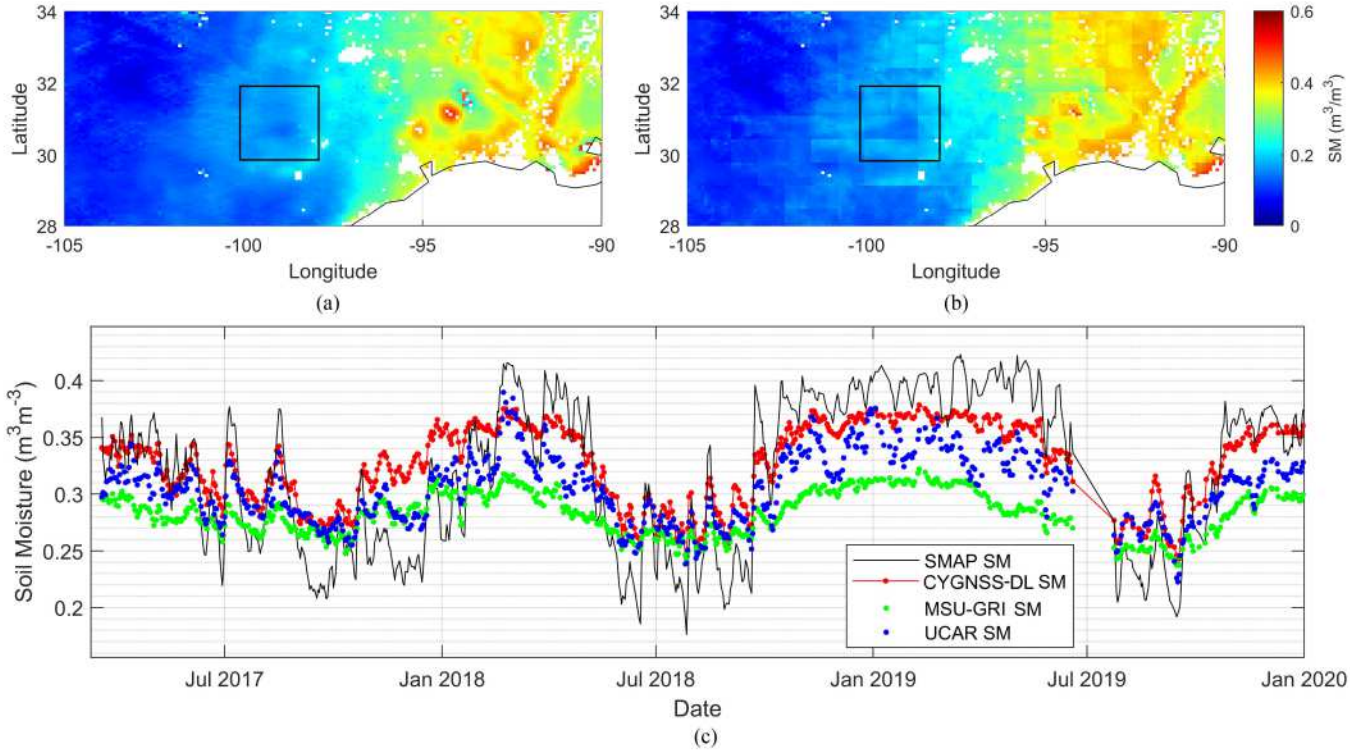

Fig. 4. Averaged SM predictions in USA (latitudes 30° to 32° and longitudes –100° to –98°) during from March 2017 to December 2020. (a) SMAP SM map of 9 km (b) CYGNSS SM map 9 km, and (c) a time-series comparison for different SM products of the selected area (black rectangle). (a) SMAP SM. (b) CYGNSS DL SM. (c) Temporal SM of a selected area in USA.

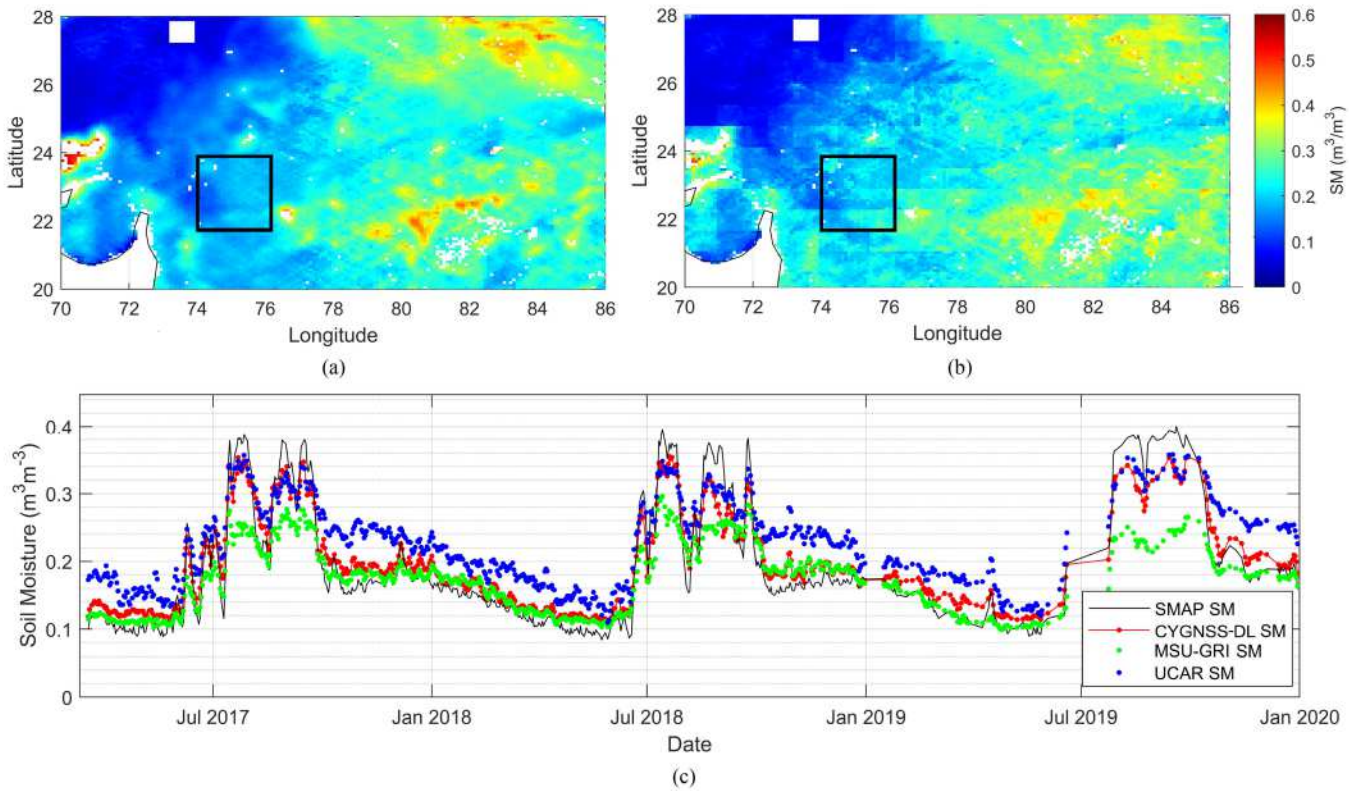

Fig. 5. Averaged SM predictions for India (latitudes 22° to 24° and longitudes 74° to 76°) during from March 2017 to December 2020. (a) SMAP SM map of 9 km (b) CYGNSS SM map 9 km, and (c) a time-series comparison for different SM products of the selected area (black rectangle). (a) SMAP SM. (b) CYGNSS DL SM. (c) Temporal SM of a selected area in India.

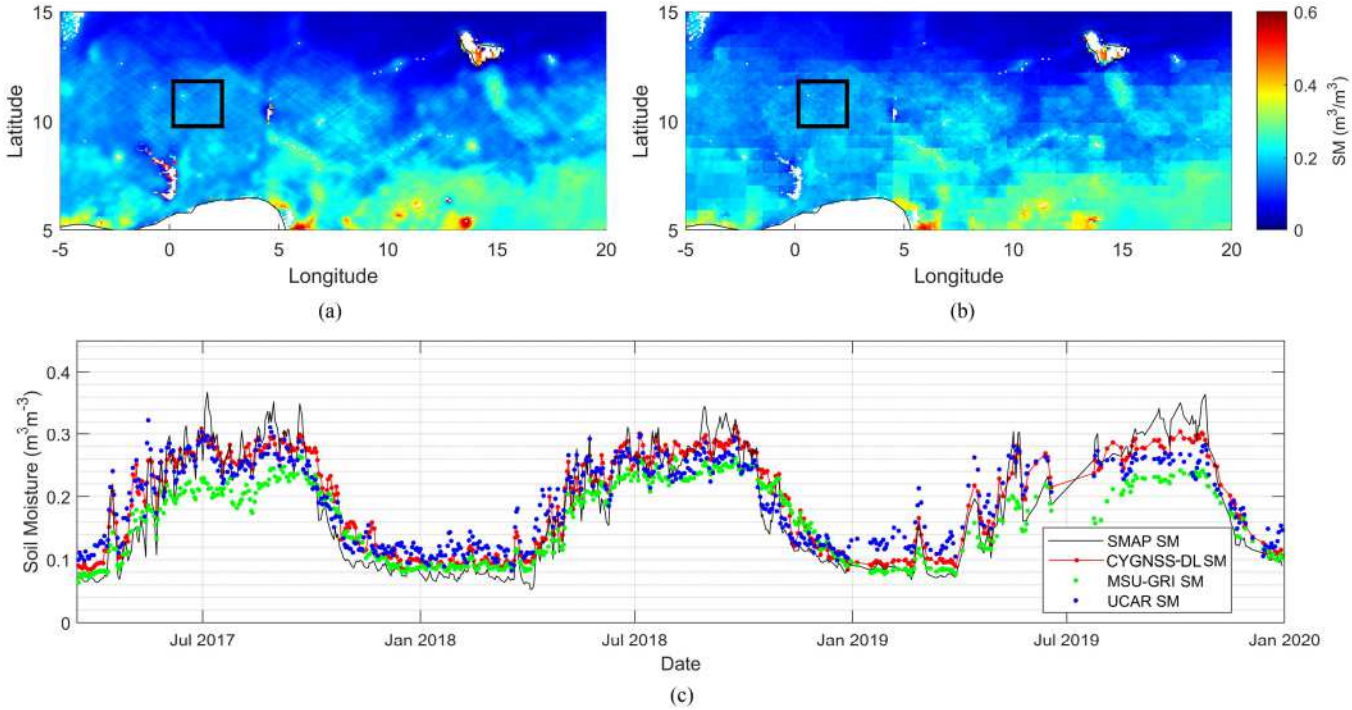

Fig. 6. Averaged SM predictions for near West Africa (latitudes 10° to 12° and longitudes 0° to 2°) regions during from March 2017 to December 2020. (a) SMAP SM map of 9 km (b) CYGNSS SM map 9 km, and (c) a time-series comparison for different SM products of the selected area (black rectangle). (a) SMAP SM. (b) CYGNSS DL SM. (c) Temporal SM of a selected area in Sahara.

The MSU-GRI follows the pattern nicely during the dry seasons but fails to capture the trend and underestimates SM during the monsoon seasons. The opposite scenario is found for the UCAR SM product. It shows good estimates with respect to the SMAP SM during the growing seasons but overestimates SM during the dry-down periods. The temporal result shows that CYGNSS-DL is able to follow dry-down and rainy patterns closely.

The last test region from West Africa is depicted in Fig. 6, representing a transitional area from barren to grass and savanna land cover types in arid and equatorial climate zones. While both SMAP and CYGNSS estimates follow a similar spatial pattern, CYGNSS SM tends to underestimate SM over some regions compared to SMAP. From the time-series analysis shown in Fig. 6(c), good correlations can be seen between SMAP and CYGNSS-DL estimates, and CYGNSS-DL follows the SMAP SM trends closely. CYGNSS-DL slightly overestimates SM during the dry-down period but captures the dynamic range during the rainy seasons. Similarly, patterns are found for the other two CYGNSS SM products over this region. UCAR SM shows relatively higher SM values than SMAP, especially during the beginning of the year, but it offers a good correlation during the rainy seasons. On the other hand, the MSU-GRI product consistently underestimates SM as compared to other products.

E. Land Cover Analysis

Additionally, the capabilities of the different SM prediction models are compared considering different land cover conditions. In total, eight primary land cover types are examined in this study. Two commonly used performance metrics (ubRMSD and correlation coefficient) are considered, and all the comparison results are made against SMAP SM. Fig. 7 shows the number of samples for each land cover type and their corresponding ubRMSDs and correlation coefficients for different CYGNSS SM products. The dominant land covers are Forest, Shrub, Savanna, Grass, Crop, and Barren. Since grids with significant surface water presence are excluded during the data quality control process, only very few samples are left for wetland land cover (approximately 55 000 samples). As clearly shown in Fig. 7, CYGNSS-DL performs better than the other method in terms of ubRMSD and correlation coefficient. On average, CYGNSS-DL provides minimal errors with consistently higher correlation coefficients. Overall, CYGNSS-DL has an average correlation greater than 0.80 for all land cover types. MSU-GRI SM product performs similarly to the DL method in terms of correlation but has higher ubRMSD errors. UCAR SM product shows higher ubRMSDs and smaller correlation coefficients than the other two approaches. Note that regions with open barren and shrub land covers are generally associated with relatively dry soil, therefore ubRMSD values are relatively small and correlations tend to be comparatively low. For the other land cover types, i.e., forest woody, savanna, grass, and croplands, relatively higher errors are found than the other two land cover types. Nevertheless, the correlations with regard to SMAP SM are moderately high. Furthermore, the sampling sizes for the forest, shrub, savanna, and grassland cover types are relatively large. A larger sample size typically helps a DL-based model to capture the empirical relationship between input features and label data. The relatively small sample sizes could be an important reason for the higher error levels over woody, wetland, and urban land cover types.

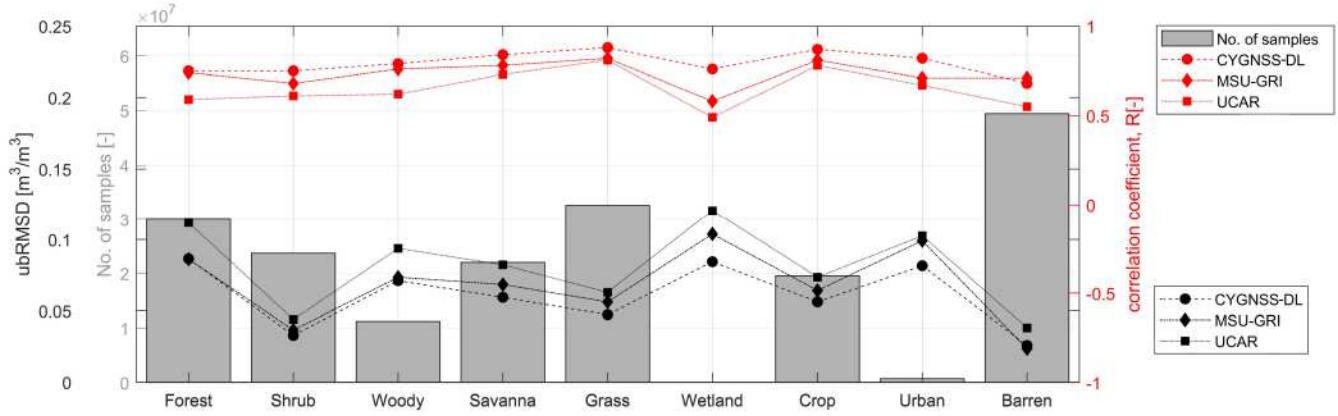

Fig. 7. Performance evaluation of different models with respect to different land cover types. The left axis is for ubRMSD (black label) and the right axis (red label) is for correlation.

F. Performance Evaluation Against ISMN

To further assess the performance of the CYGNSS-DL method, SM estimates derived through the DL method along with other SM products are compared against ISMN data. Three ISMN sites are selected to demonstrate the temporal variations of different SM products. Three sites are selected one from each of the three different ISMN network (i.e., USCRN, SCAN, and COSMOS) which have a good number of samples for all soil moisture products throughout the years. Some sites are selected that are commonly used in previous studies [22], [25]. Here, SM observations recorded at these sites from 2017 to 2020 are shown in Fig. 8.

The first representative site is Batesville-8-WNW which belongs to the USCRN network. This site is located in the mideast of the United States. SM temporal variation at this site is presented in Fig. 8(a). As shown in the figure, SM has a large dynamic range with the highest SM at around 0.4 m^3m^{-3} and the lowest SM at around 0.1 m^3m^{-3}. SMAP (black dotted line) provides higher SM values than ISMN which indicates high bias error. However, all three CYGNSS SM products have low bias error comparable to ISMN and follow very closely to SMAP SM. The comparison between CYGNSS-DL (red dot line) with ISMN gives an ubRMSE of 0.0607 m^3m^{-3} and a correlation of 0.56. MSU-GRI shows relatively small SM dynamics. UCAR SM product shows a relatively good SM range as compared to in-situ SM.

The second representative site is Knoxcity which belongs to the SCAN SM network. This site is located in the midwest of the United States. The time-series analysis for this site is presented in Fig. 8(b). SM variation at this site is slightly smaller than the first site. CYGNSS-DL provides an ubRMSE of 0.0385 m^3m^{-3} with a correlation of 0.81. UCAR SM performs almost similarly to the DL approach at this site. On the other hand, MSU-GRI shows a good correlation with SMAP but fails to capture the SM trend when the SM value is high.

The final example site is SMAP-OK which belongs to the COSMOS SM network. The SM time series from this site are presented in Fig. 8(c). This site is located in a relatively arid region with SM ranging from 0.05 to 0.2 m^3m^{-3}. All CYGNSS and SMAP SM products show much higher SM values than the ground-based measurements from ISMN. CYGNSS-DL provides an ubRMSE of 0.0354 m^3m^{-3} with a correlation of 0.72 when compared with ISMN SM.

Table III shows the statistic calculated between different SM products with respect to ISMN data at different networks. Some of the sites are removed from the comparison as there are a smaller number of samples. A total of 129 sites are considered for this comparison. Most of the sites belong to the SCAN network and the averaged metrics across sites are shown in this table. CYGNSS-DL method and UCAR provide similar results across the 70 sites in the SCAN network. In general, SMAP provides a better correlation than all CYGNSS products. CYGNSS-DL has very close ubRMSE values with SMAP but smaller correlation coefficients. When considering all sites, the average ubRMSE values are 0.050, 0.053, 0.058 and 0.053 for SMAP, CYGNSS-DL, MSU-GRI, and UCAR, respectively. The averaged correlation coefficients are 0.68, 0.51, 0.40, and 0.48, respectively.

G. Triple Collocation (TC) Analysis

In this section, the TC technique is applied to evaluate different SM products by estimating the correlations (*TCr*) with regard to the theoretical true values. Although the TC approach can be used to characterize the accuracy of SM products at the global scale, it requires at least three independent data sources and a sufficient number of collocated samples for robust estimation. In this analysis, ASCAT and GLDAS products are chosen as two fixed datasets and the third SM product varies from SMAP to CYGNSS products. It is important to note that seasonal climatologies in all SM time series are removed before applying the TC analysis to eliminate the impact of seasonal systematic errors. The analysis is conducted at both 9 km and 36 km. ASCAT and GLDAS NOAH products are resampled from their native spatial resolution to 9 km as described in Section III-C. For each grid, a minimum of 100 collocated samples are required for conducting TC analysis. The first triplet consists of SMAP, ASCAT, and GLDAS NOAH. Similarly,

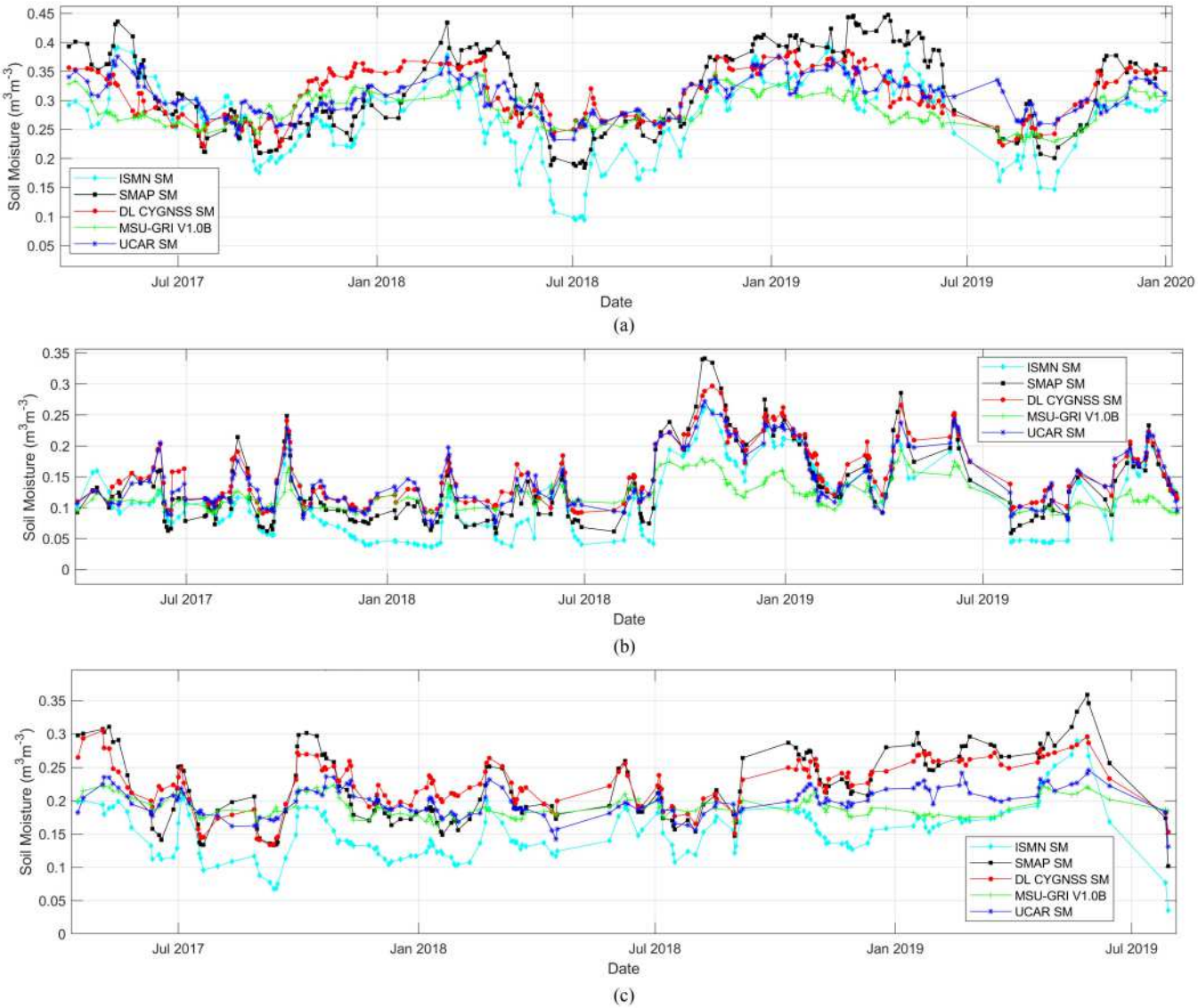

Fig. 8. Time series examples of daily averaged SMAP, CYGNSS-DL, MSU-GRI, and UCAR SM predictions against selected ISMN sites with a moderate performance. (a) Site name: Batesville, ubRMSE: 0.060729, R: 0.562473 (CYGNSS vs ISMN). (b) Site name: KnoxCity, ubRMSE: 0.038556, R: 0.811475 (CYGNSS vs ISMN). (c) Site name: SMAP-OK, ubRMSE: 0.035475, R: 0.720991 (CYGNSS vs ISMN).

TABLE III
STATISTICS BETWEEN GROUND-BASED SM MEASUREMENTS WITH DIFFERENT SM PRODUCTS FOR DIFFERENT NETWORKS WITH CORRESPONDING EXAMINED ISMN SITES

Network	No. of sites	ISMN vs SMAP			ISMN vs CYGNSS-DL			ISMN vs MSU-GRI			ISMN vs UCAR		
		RMSE	ubRMSE	R	RMSD	ubRMSE	R	RMSE	ubRMSD	R	RMSE	ubRMSE	R
SCAN	70	0.100	0.053	0.66	0.108	0.056	0.52	0.094	0.060	0.40	0.110	0.056	0.51
USCRN	32	0.097	0.048	0.70	0.102	0.052	0.54	0.085	0.060	0.41	0.112	0.056	0.49
COSMOS	12	0.096	0.045	0.70	0.100	0.048	0.049	0.070	0.038	0.40	0.129	0.044	0.45
ALL	129	0.097	0.050	0.68	0.104	0.053	0.51	0.091	0.058	0.40	0.109	0.053	0.48

Note: Metrics are calculated separately for each site with mean values shown for each network or across all sites.

CYGNSS-DL, UCAR, and MSU-GRI performances are also evaluated by keeping ASCAT and GLDAS as fixed members in the TC triplet. The global mean and median values of all valid grids are calculated and shown in Table IV. In the case of 9-km grid, SMAP offers a global mean of *TCr* is 0.7810, whereas CYGNSS-DL, MSU-GRI, and UCAR provide 0.6174, 0.4957, and 0.5356, respectively. On the other hand, the global mean *TCr* for a 36-km grid of SMAP is 0.7960 and the global averaged *TCr* of three CYGNSS products are 0.5979, 0.4928, and 0.4682 for CYGNSS-DL, MSU-GRI, and UCAR, respectively.

Fig. 9 demonstrates the 36-km grid-wise spatial distribution of *TCr* for different SM products. Fig. 9(a) shows that SMAP

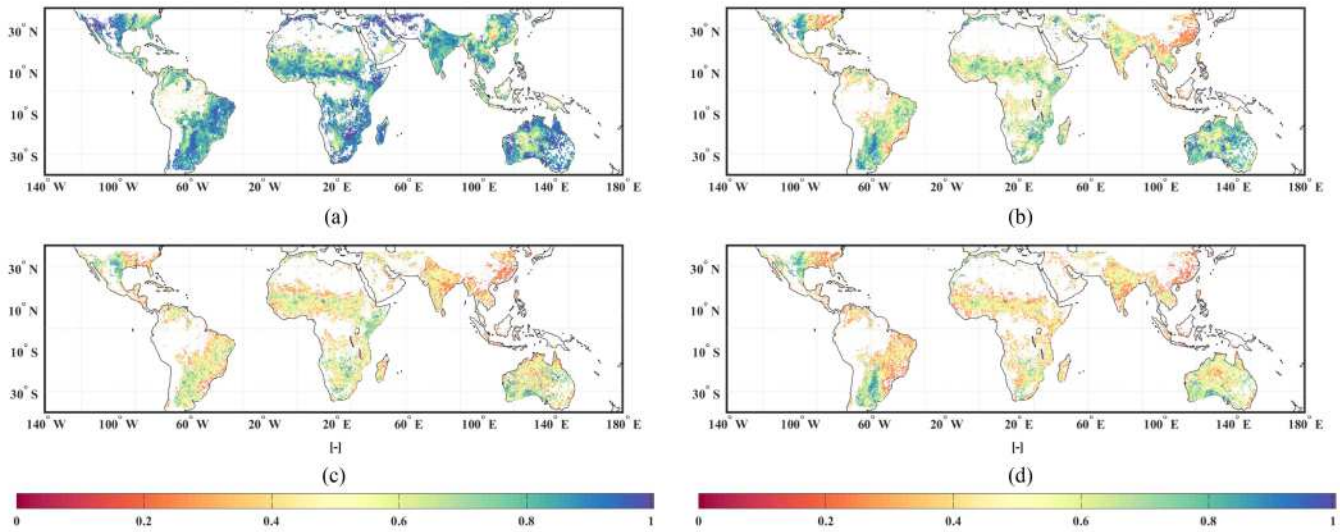

Fig. 9. TC-based correlation coefficient for different SM data products against ASCAT and GLDAS at 36-km resolution. (a) SMAP TC_T. (b) CYGNSS-DL TC_T. (c) MSU-GRI TC_T. (d) UCAR TC_T.

TABLE IV
TC ANALYSIS RESULTS FOR DIFFERENT SM PRODUCTS WITH RESPECT TO THE UNKNOWN TRUE SM VALUE

Products	9-km		36-km	
	Mean correlation	Median correlation	Mean correlation	Median correlation
SMAP	0.7810	0.8071	0.7960	0.8255
CYGNSS-DL	0.6174	0.6297	0.5979	0.6191
MSU-GRI	0.4957	0.4898	0.4928	0.4941
UCAR	0.5356	0.5323	0.4682	0.4604

All data are spatially and temporally collocated for this analysis. The result shows the mean and median results for two different grid cases (9 km and 36 km).

generally has high *TCr* over the majority of the globe, indicating superior sensitivity to SM compared to any other CYGNSS SM products. This is due most likely to intrinsic differences between SMAP and CYGNSS. The quasi-global medians for CYGNSS products are around 0.5 as listed in Table IV. Moreover, CYGSS-DL [Fig. 9(b)] has demonstrated higher correlations than the other two CYGNSS SM products [Fig. 9(c) and (d)], suggesting its improved performance in characterizing the temporal variations of soil moisture.

V. DISCUSSION

The space-borne GNSS-R observations have become popular given their wide variety of applications [50], [51], [52].

This work extends our previously developed DL framework to estimate global SM retrievals using DDMs. In previous research [20], the whole DDM frame was investigated to exploit a better understanding of CYGNSS data for SM retrieval. Different statistical moments of surface reflectivity were employed, including the maximum, mean, variance, skewness, and kurtosis. By leveraging the power of a DL model with better approximation functions and the ability to find complex nonlinearity, the entire DDMs can be easily investigated and extract more important features that contribute to the SM estimation. A reliable, accurately labeled, and well-organized dataset is also required to use a DL algorithm for proper SM estimation. Our study shows that enhanced features can be learned directly from the full DDMs leading to more accurate SM estimation performance compared to utilizing several engineered derivative DDM features.

Ancillary data plays an important role in SM retrieval algorithm. Several static and time-varying geophysical data are utilized, which can provide land dynamics for SM retrieval. But reducing the dependency on ancillary data is also important. In some previous studies, more ancillary features are used than the CYGNSS-derived features in the retrieval algorithms, so the algorithms become more dependent on auxiliary data. In our developed DL model, we used 128 features [27] that are derived directly from DDMs. In our case, the model becomes more tends to depend more on CYGNSS data for SM retrieval. It may still need some improvement by reducing the dependencies of ancillary data.

Different cross-validation strategies are used to evaluate the model's capability to predict SM accurately. A comprehensive comparison among several publicly available CYGNSS-based SM data products is provided in this work. Both spatial and temporal comparisons demonstrate a high correlation between CYGNSS products with SMAP SM data. Additionally, the robust TC results also suggest that CYGNSS-DL performs higher correlation with the underlying true SM compared to other CYGNSS-based SM products.

The proposed approach utilizes SMAP as the label SM for learning the DL model parameters. While SMAP is seen to perform best across different land covers and climate conditions in TC analysis, considering its own intrinsic instrumental and retrieval errors that could propagate to CYGNSS SM retrival process [30], a high-quality reference SM data is seen as highly important for all DL-based SM retrieval approaches.

The in-situ analysis also demonstrates CYGNSS-DL can capture temporal dynamics closely producing the lowest ubRMSE or highest correlations among CYGNSS based SM products. Regarding the temporal variation and land cover analysis, superior performance are observed in CYGNSS-DL across different land cover types with specifically low ubRMSDs and high

correlations for savanna, croplands, and grassland compared to other land covers.

Training DL models over a huge amount of data are generally computationally expensive, but when the model is trained estimating SM is only a forward pass of the model which is generally not computationally significant. We observed that rather than learning a single model for the whole world, the introduced clustering approach; learning a DL model for each cluster; both increased the performance and reduced the training complexity of the DL approach. However, such clustering artifacts can be seen in CYGNSS SM maps [Figs. 4(b), 5(b), and 6(b)], as changes on the borders of each 72-km cluster. A more dynamic DL network with smoother transitions can be used for varied soil moisture conditions in future work. It is essential to note that the true spatial resolution of CYGNSS data is subject to interpretation. The actual spatial resolutions depend on the surface dynamics because the instrument DDM can spread unevenly for a given area. The DL model trained with 9-km enhanced SMAP data products, so we refer to 9 km as our posted resolution. But, it does not necessarily have the CYGNSS native resolution as it might vary from a few kilometers depending upon the degree of coherence. Bringing DDMs into retrieval algorithms helps better replicate SMAP SM product as DDM can provide additional features that technically learn from the degree of signal spread. Full DDM images seem to help better than the handcrafted features (e.g., peak reflectivity) for SM estimation as they can cover SMAP's original resolution (36 km). However, it requires further investigation to prove this claim.

VI. Conclusion

This article evaluates the DL-based framework for quasi-global SM estimation, which uses CYGNSS DDMs and ancillary geophysical data. One of the most widely-used DL methods (i.e., CNN) is utilized in this work to learn features from DDMs. DL models are trained and validated based within regional clusters. In particular, the 72-km clusters are chosen to optimize both estimation performance and computational complexity. Separate models are trained for each cluster, and the models are validated using fivefold cross-validation and a year-based cross-validation method with data from 2017 to 2021. An ubRMSD of 0.0365 m^3m^{-3} and an R-value of 0.92 is achieved for fivefold cross-validation over all data on SMAP recommended grids. The year-wise cross-validation shows an overall performance of an ubRMSD of 0.0403 m^3m^{-3} and an R-value of 0.88 when models are learned using data from 2017–2020 and tested in 2021. An essential evaluation of the DL method is to compare the predictions with other publicly available CYGNSS-based SM products. Two additional CYGNSS SM products are included for comparison, i.e., MSU-GRI and UCAR, that are also trained with SMAP SM data. The DL method outperforms the publicly available SM products by providing smaller ubRMSD values and higher correlation coefficients. Moreover, CYGNSS-DL SM estimates are compared with ISMN SM data. Results show that CYGNSS-DL performs similarly to SMAP across the ISMN sites, suggesting that the DL model can be generalized in space and time with promising confidence. An independent global evaluation is conducted via the TC approach. Results show that good improvements can be obtained from CYGNSS-DL when compared with MSU-GRI and UCAR products. In conclusion, the DL algorithm shows clear advantages for global-scale SM retrieval using the entire DDM information.

References

[1] H. Vereecken et al., "On the value of soil moisture measurements in vadose zone hydrology: A review," *Water Resour. Res.*, vol. 44, no. 4, pp. 43–1397, 2008, doi: 10.1029/2008WR006829.

[2] D. A. Robinson et al., "Soil moisture measurement for ecological and hydrological watershed-scale observatories: A review," *Vadose Zone J.*, vol. 7, no. 1, pp. 358–389, 2008.

[3] M. E. Holzman, R. Rivas, and M. C. Piccolo, "Estimating soil moisture and the relationship with crop yield using surface temperature and vegetation index," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 28, pp. 181–192, 2014.

[4] D. Entekhabi, I. Rodriguez-Iturbe, and F. Castelli, "Mutual interaction of soil moisture state and atmospheric processes," *J. Hydrol.*, vol. 184, no. 1/2, pp. 3–17, 1996.

[5] M. Jung et al., "Recent decline in the global land evapotranspiration trend due to limited moisture supply," *Nature*, vol. 467, no. 7318, pp. 951–954, 2010.

[6] E. G. Njoku and D. Entekhabi, "Passive microwave remote sensing of soil moisture," *J. Hydrol.*, vol. 184, no. 1/2, pp. 101–129, 1996.

[7] L. Wang and J. J. Qu, "Satellite remote sensing applications for surface soil moisture monitoring: A review," *Front. Earth Sci. China*, vol. 3, no. 2, pp. 237–247, 2009.

[8] D. Entekhabi et al., "The soil moisture active passive SMAP mission," *Proc. IEEE*, vol. 98, no. 5, pp. 704–716, May 2010.

[9] Y. Kerr et al., "Overview of SMOS performance in terms of global soil moisture monitoring after six years in operation," *Remote Sens. Environ.*, vol. 180, pp. 40–63, 2016.

[10] R. Torres et al., "GMES Sentinel-1 mission," *Remote Sens. Environ.*, vol. 120, pp. 9–24, 2012.

[11] A. Balenzano et al., "Sentinel-1 soil moisture at 1 km resolution: A validation study," *Remote Sens. Environ.*, vol. 263, 2021, Art. no. 112554.

[12] C. S. Ruf et al., "A new paradigm in earth environmental monitoring with the CYGNSS small satellite constellation," *Sci. Rep.*, vol. 8, no. 1, pp. 1–13, 2018.

[13] C. Chew and E. Small, "Description of the UCAR/CU soil moisture product," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1558.

[14] C. C. Chew and E. E. Small, "Soil moisture sensing using space-borne GNSS reflections: Comparison of CYGNSS reflectivity to SMAP soil moisture," *Geophys. Res. Lett*, vol. 45, no. 9, pp. 4049–4057, 2018.

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

[16] M. M. Al-Khaldi, J. T. Johnson, A. J. O'Brien, A. Balenzano, and F. Mattia, "Time-series retrieval of soil moisture using cygnss," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4322–4331, Jul. 2019.

[17] M. Al-Khaldi, J. T. Johnson, A. J. O'Brien, A. Balenzano, and F. Mattia, "Time-series retrieval of soil moisture using CYGNSS," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4322–4331, Jul. 2019.

[18] O. Eroglu, M. Kurum, D. Boyd, and A. C. Gurbuz, "High spatio-temporal resolution CYGNSS soil moisture estimates using artificial neural networks," *Remote Sens.*, vol. 11, no. 19, 2019, Art. no. 2272.

[19] T. Yang, W. Wan, Z. Sun, B. Liu, S. Li, and X. Chen, "Comprehensive evaluation of using techdemosat-1 and CYGNSS data to estimate soil moisture over mainland China," *Remote Sens.*, vol. 12, no. 11, 2020, Art. no. 1699.

[20] Q. Yan, W. Huang, S. Jin, and Y. Jia, "Pan-tropical soil moisture mapping based on a three-layer model from CYGNSS GNSS-R data," *Remote Sens. Environ.*, vol. 247, 2020, Art. no. 111944.

[21] V. Senyurek, F. Lei, D. Boyd, M. Kurum, A. Gurbuz, and R. Moorhead, "Machine learning-based CYGNSS soil moisture estimates over ISMN sites in CONUS," *Remote Sens.*, vol. 12, no. 7, 2020, Art. no. 1168.

[22] V. Senyurek, F. Lei, D. Boyd, A. C. Gurbuz, M. Kurum, and R. Moorhead, "Evaluations of a machine learning-based CYGNSS soil moisture estimates against SMAP observations," *Remote Sens.*, vol. 12, no. 21, 2020, Art. no. 3503.

[23] F. Lei, V. Senyurek, M. Kurum, A. Gurbuz, D. Boyd, and R. Moorhead, "Quasi-global GNSS-R soil moisture retrievals at high spatio-temporal resolution from CYGNSS and SMAP data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 6303–6306.

[24] S. H. Yueh, R. Shah, M. J. Chaubell, A. Hayashi, X. Xu, and A. Colliander, "A semiempirical modeling of soil moisture, vegetation, and surface roughness impact on CYGNSS reflectometry data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2020, Art. no. 5800117.
[25] F. Lei et al., "Quasi-global machine learning-based soil moisture estimates at high spatio-temporal scales using CYGNSS and SMAP observations," *Remote Sens. Environ.*, vol. 276, 2022, Art. no. 113041.
[26] T. M. Roberts, I. Colwell, C. Chew, S. Lowe, and R. Shah, "A deep-learning approach to soil moisture estimation with GNSS-R," *Remote Sens.*, vol. 14, no. 14, 2022, Art. no. 3299.
[27] M. M. Nabi, V. Senyurek, A. C. Gurbuz, and M. Kurum, "Deep learning-based soil moisture retrieval in CONUS using CYGNSS delay–Doppler maps," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6867–6881, 2022.
[28] A. Gruber, C-H Su, S. Zwieback, W. Crow, W. Dorigo, and W. Wagner, "Recent advances in (soil moisture) triple collocation analysis," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 45, pp. 200–211, 2016.
[29] F. Chen et al., "Global-scale evaluation of SMAP, SMOS and ASCAT soil moisture products using triple collocation," *Remote Sens. Environ.*, vol. 214, pp. 1–13, 2018.
[30] X. Deng et al., "Triple collocation analysis and in-situ validation of the CYGNSS soil moisture product," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1883–1899, 2023.
[31] K. A. McColl, J. McColl, A. G. Vogelzang, D. Entekhabi, M. Piles, and A. D. Stoffelen, "Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target," *Geophys. Res. Lett.*, vol. 41, no. 17, pp. 6229–6236, 2014.
[32] C. Ruf et al., *CYGNSS Handbook*. Ann Arbor, MI, USA: Michigan Pub., Univ. of Michigan, 2022.
[33] M. M. Nabi, V.A. Senyurek, C. Gurbuz, and M. Kurum, "A deep learning-based soil moisture estimation in conus region using CYGNSS delay doppler maps," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 6177–6180.
[34] S. Chan, R. Bindlish, R. Hunt, T. Jackson, and J. Kimball, "Vegetation water content," Jet Propulsion Lab., California Inst. Technol., Pasadena, CA, USA, Tech. Rep. JPL D-53061, 2013.
[35] T. Hengl et al., "Soilgrids250 m: Global gridded soil information based on machine learning," *PLoS One*, vol. 12, no. 2, 2017, Art. no. e0169748.
[36] J.-F. Pekel, A. Cottam, N. Gorelick, and A. S. Belward, "High-resolution mapping of global surface water and its long-term changes," *Nature*, vol. 540, no. 7633, 2016, Art. no. 418.
[37] S. K. Chan et al., "Development and assessment of the SMAP enhanced passive soil moisture product," *Remote Sens. Environ.*, vol. 204, pp. 931–941, 2018.
[38] W. A. Dorigo et al., "The international soil moisture network: A data hosting facility for global in situ soil moisture measurements," *Hydrol. Earth Syst. Sci.*, vol. 15, no. 5, pp. 1675–1698, 2011.
[39] A. Gruber, W. A. DorigoS. ZwiebackA. Xaver, and W. Wagner, "Characterizing coarse-scale representativeness of in situ soil moisture measurements from the international soil moisture network," *Vadose Zone J.*, vol. 12, no. 2, 2013, Art. no. vzj2012-0170.
[40] V. Naeimi, K. Scipal, Z. Bartalis, S. Hasenauer, and W. Wagner, "An improved soil moisture retrieval algorithm for ERS and METOP scatterometer observations," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 1999–2013, Jul. 2009.
[41] M. Rodell et al., "The global land data assimilation system," *Bull. Amer. Meteorological Soc.*, vol. 85, no. 3, pp. 381–394, 2004.
[42] N. Rodriguez-Alvarez, E. Podest, K. Jensen, and K. C. McDonald, "Classifying inundation in a tropical wetlands complex with GNSS-R," *Remote Sens.*, vol. 11, 2019, Art. no. 1053.
[43] R. Zhang, S. Kim, and A. Sharma, "A comprehensive validation of the SMAP enhanced level-3 soil moisture product using ground measurements over varied climates and landscapes," *Remote Sens. Environ.*, vol. 223, pp. 82–94, 2019.
[44] A. M. Balakhder, M. M. Al-Khaldi, and J. T. Johnson, "On the coherency of ocean and land surface specular scattering for GNSS-R and signals of opportunity systems," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10426–10436, Dec. 2019.
[45] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.
[46] K. Fang, M. Pan, and C. Shen, "The value of SMAP for long-term soil moisture estimation with the help of deep learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2221–2233, Apr. 2019.
[47] R. Fernandez-Moran et al., "SMOS-IC: An alternative SMOS soil moisture and vegetation optical depth product," *Remote Sens.*, vol. 9, no. 5, 2017, Art. no. 457.
[48] S. Zwieback, K. Scipal, W. Dorigo, and W. Wagner, "Structural and statistical properties of the collocation technique for error characterization," *Nonlinear Process. Geophys.*, vol. 19, no. 1, pp. 69–80, 2012.
[49] A. Gruber, W. A. DorigoW. Crow, and W. Wagner, "Triple collocation-based merging of satellite soil moisture retrievals," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 6780–6792, Dec. 2017.
[50] N. Rodriguez-Alvarez et al., "Land geophysical parameters retrieval using the interference pattern GNSS-R technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 71–84, Jan. 2011.
[51] E. Valencia, A. Camps, N. Rodriguez-Alvarez, H. Park, and I. Ramos-Perez, "Using GNSS-R imaging of the ocean surface for oil slick detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 1, pp. 217–223, Feb. 2013.
[52] J. Bu, K. Yu, X. Zuo, J. Ni, Y. Li, and W. Huang, "Glows-Net: A deep learning framework for retrieving global sea surface wind speed using spaceborne GNSS-R data," *Remote Sens.*, vol. 15, no. 3, 2023, Art. no. 590.

M M Nabi (Student Member, IEEE) received the bachelor's degree in electrical and electronic engineering from the Ahsanullah University of Science and Technology, Dhaka, Bangladesh, in 2014. He is currently working toward the Ph.D. degree in electrical and computer engineering with the Department of Electrical and Computer Engineering, Mississippi State University (MSU), Mississippi State, MS, USA.

He is a Graduate Research Assistant with High-Performance Computing Collaboratory (HPCC) and Information Processing and Sensing (IMPRESS) Laboratory, MSU. His research interests include signal processing, remote sensing, machine learning, and deep learning.

Volkan Senyurek received the B.S., M.S., and Ph.D. degrees in electronics and communication engineering from Marmara University, Istanbul, Turkey, in 2003, 2007, and 2013, respectively.

Until 2015, he was an Assistant Professor with Marmara University. From 2015 to 2017, he was a Postdoctoral Researcher with the Department of Mechanical and Materials Engineering, Florida International University, Miami, FL, USA. From 2017 and 2019, he was a Postdoctoral Researcher with the Department of Electric and Computer Engineering, University of Alabama, Tuscaloosa, AL, USA. He is currently an Assistant Research Professor with Geosystems Research Institute, Mississippi State University, Mississippi State, MS, USA. His research interests include remote sensing, biomedical signal processing, wearable sensors, pattern recognition, fiber optic sensors, and structural health monitoring.

Fangni Lei received the B.S. degree in geographic information system and the M.S. and Ph.D. degrees in cartography and geographical information engineering from Wuhan University, Wuhan, China, in 2011, 2013, and 2016, respectively.

She was a Visiting Student with the Hydrology and Remote Sensing Laboratory, USDA-Agricultural Research Services. From 2017 to 2019, she continued her research as a Postdoctoral Associate with USDA, and then from 2019 to 2022, as a Research Assistant Professor with Mississippi State University, Mississippi State, MS, USA. She is currently an Assistant Research Professor with Eversource Energy Center, University of Connecticut, Storrs, CT, USA. Her research interests include but are not limited to quantifying soil moisture from microwave remote sensing, land surface water-energy balance modeling, watershed-scale hydrologic modeling and hydrologic data assimilation, and agricultural water management.

Mehmet Kurum (Senior Member, IEEE) received the B.S. degree in electrical and electronics engineering from Bogazici University, Istanbul, Turkey, in 2003, and the M.S. and Ph.D. degrees in electrical engineering from George Washington University, Washington, DC, USA, in 2005 and 2009, respectively.

He held a Postdoctoral position with the Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD, USA. He is currently an Associate Professor and Paul B. Jacob Endowed Chair in electrical and computer engineering with Mississippi State University, Mississippi State, MS, USA, where he is also Co-Director of Information Processing and Sensing (IMPRESS) Laboratory. His current research interests include recycling the radio spectrum to address the challenges of decreasing radio spectrum space for science while exploring entirely new microwave regions for land remote sensing.

Dr. Kurum was a recipient of the Leopold B. Felsen Award for excellence in electromagnetic in 2013 and the International Union of Radio Science (URSI) Young Scientist Award in 2014, and NSF CAREER award in 2022. From 2014 to 2021, he served as an Early Career Representative for the International URSI Commission F (Wave Propagation and Remote Sensing). He is a Member of the U.S. National Committee for the International Union of Radio Science (USNC-URSI). He is currently a Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* since 2021.

Ali Cafer Gurbuz (Senior Member, IEEE) received the B.S. degree in electrical engineering from Bilkent University, Ankara, Turkey, in 2003, and the M.S. and Ph.D. degrees from the Georgia Institute of Technology, Atlanta, GA, USA, in 2005 and 2008, respectively, both in electrical and computer engineering.

From 2003 to 2009, he researched compressive sensing-based computational imaging problems with Georgia Institute of Technology. From 2009 to 2017, he held faculty positions with TOBB University of Economics and Technology, Ankara, Turkey, and University of Alabama, Tuscaloosa, AL, USA, where he pursued an active research program on the development of sparse signal representations, compressive sensing theory and applications, radar and sensor array signal processing, and machine learning. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Mississippi State University, Mississippi State, MS, USA, where he is Co-Director of Information Processing and Sensing (IMPRESS) Laboratory.

Dr. Gurbaz was the recipient of The Best Paper Award for *Signal Processing* in 2013, the Turkish Academy of Sciences Best Young Scholar Award in Electrical Engineering in 2014, and NSF CAREER award in 2021. He was an Associate Editor for several journals such as *Digital Signal Processing*, *EURASIP Journal on Advances in Signal Processing*, and *Physical Communications*.