Deep Learning-Based Soil Moisture Retrieval in CONUS Using CYGNSS Delay–Doppler Maps

M M Nabi, Volkan Senyurek, Ali C. Gurbuz, *Senior Member, IEEE*,
and Mehmet Kurum, *Senior Member, IEEE*

***Abstract*—National Aeronautics and Space Administration's Cyclone Global Navigation Satellite System (CYGNSS) mission has gained significant attention within the land remote sensing community for estimating soil moisture (SM) by using the Global Navigation System Reflectometry (GNSS-R) technique. CYGNSS constellation generates Delay-Doppler Maps (DDMs), containing important Earth surface information from GNSS reflection measurements. Previous studies considered only designed features from CYGNSS DDM, whereas the whole DDM image is affected by SM, inundation, and vegetation. This paper presents a deep learning (DL) based framework for estimating SM in the Continental United States by leveraging spaceborne GNSS-R DDM observations provided by the CYGNSS constellation along with remotely sensed geophysical data. A data-driven approach utilizing convolutional neural networks (CNNs) is developed to determine complex relationships between the reflected measurements and surface parameters which can provide improved SM estimation. The model is trained jointly with three types of processed DDM images of analog power, effective scattering area, and bistatic radar cross-section with other auxiliary geophysical information such as elevation, soil properties, and vegetation water content (VWC). The model is trained and evaluated using the Soil Moisture Active Passive (SMAP) mission's enhanced SM products at a 9 km resolution with VWC less than 5 kg/m^2. The mean unbiased root-mean-square difference between concurrent CYGNSS and SMAP SM retrievals from 2017 to 2020 is 0.0366 m^3/m^3 with a correlation coefficient of 0.93 over fivefold cross-validation and 0.0333 m^3/m^3 with a correlation coefficient of 0.94 over year-based cross-validation at spatial resolution of 9 km and temporal resolution similar to CYGNSS mission.**

***Index Terms*—Convolutional neural network (CNN), Cyclone Global Navigation Satellite System (CYGNSS), deep learning (DL), Global Navigation Satellite System (GNSS)-reflectometry, Soil Moisture Active Passive (SMAP), soil moisture (SM) retrieval.**

I. Introduction

SOIL moisture (SM) is essential for crop harvesting, rain forecasting, hydrology, meteorology, and different Earth science applications [1], [2], [3]. High-resolution and precise SM estimation is required for many application for forecasting floods and agriculture yields [4], [5]. Remote sensing techniques have been used widely for SM retrieval [6]. There are some dedicated satellites that have been used in order to retrieve SM from the Earth's surface with different spatial and temporal resolutions. The National Aeronautics and Space Administration's (NASA) Soil Moisture Active Passive (SMAP) [7], and the European Space Agency's (ESA) Soil Moisture and Ocean Salinity [8] are two satellite missions that are operated with *L*-band passive radiometers and provide SM approximately 36-km spatial resolution and 2–3 days temporal coverage. Another ESA mission called Sentinel-1 is a synthetic aperture radar operated at *C*-band, and it can generate SM with 1-km spatial resolution and 6–12 days global coverage [9].

Global Navigation Satellite System-Reflectometry (GNSS-R) has become popular with the scientific community as it has great potential by providing higher spatio-temporal coverage measurements over the traditional remote microwave remote sensing techniques. GNSS-R receives the reflected signals from the Earth's surface through bistatic scattering covering the space-time gap of conventional monostatic active or passive satellite missions. By cross-correlating a measured GNSS signal reflected from a scattering surface with either a received direct signal or a GNSS signal replica, the GNSS-R approach determines geophysical parameters of the observed surface area [10]. It becomes an effective approach for monitoring sea surface roughness and wind vectors using space-borne and airborne systems [11], [12], [13]. Extensive research is ongoing using GNSS-R for biomass retrieval [14], sea ice monitoring [15], ocean altimetry [16], and SM estimation [17], [18], [19], [20], [21], [22], [23].

NASA launched a mission in December 2016 called Cyclone Global Navigation Satellite System (CYGNSS). CYGNSS receives GNSS-R measurements from 32 channels with eight small microsatellites and four channels during the 95-min orbital period of each satellite. Its mean revisit time can be as small as seven hours with a 25-km spatial resolution across the ocean under dominantly diffuse scattering conditions. The mission's primary purpose is to enhance hurricane forecasting by better understanding the interactions between the air near the core of a storm and the sea. It covers from 38° north to 38° south latitudes over both land and ocean providing very useful observations over land as well.

Manuscript received 9 May 2022; revised 13 July 2022; accepted 2 August 2022. Date of publication 5 August 2022; date of current version 29 August 2022. This work was supported in part by National Science Foundation under Grant 2047771 and in part by the USDA-ARS Award under Grant NACA 58-6064-9-007. *(Corresponding author: Ali C. Gurbuz.)*

M M Nabi, Ali C. Gurbuz, and Mehmet Kurum are with the Department of Electrical and Computer Engineering, Information Processing and Sensing Lab, Mississippi State University, Mississippi State, MS 39672 USA (e-mail: mn918@msstate.edu; gurbuz@ece.msstate.edu; kurum@ece.msstate.edu).

Volkan Senyurek is with the Geosystems Research Institute, Mississippi State University, Mississippi State, MS 39579 USA (e-mail: volkan@gri.msstate.edu).

Digital Object Identifier 10.1109/JSTARS.2022.3196658

Many recent analyses show improved models and algorithms to estimate SM by taking advantage of the large amount of CYGNSS measurements at different spatial and temporal coverage [17], [20], [21], [23], [24], [25], [26], [27]. Despite being designed to estimate ocean wind vectors, CYGNSS has also shown a significant sensitivity to SM variation and a high correlation with SMAP SM data products [17], [18], [22], [24]. The majority of the previous studies used designed features such as effective reflectivity obtained from peak reflected power, leading edge slope, and trailing edge slope in a delay–Doppler map (DDM) [21], [28]. These approaches utilize their designed features computed from a DDM image as the main information DDM brings into the SM estimation problem. However, besides the SM content, vegetation and topographical properties also affect entire DDMs, and DDMs carry much more information than just their peak power value. While ancillary information from other sources can provide additional information, this article aims to develop approaches that learn the relevant features directly from the entire DDM images for the SM estimation problem and, by this way, to increase SM estimation accuracy.

Different processed DDM products are available from CYGNSS, including analog power, effective scattering area, and bistatic radar cross-section (BRCS). Our proposed approach utilizes these processed DDMs jointly as inputs, together with ancillary data, within a deep learning (DL) architecture to estimate the SM value. A recent study showed a DDM could be used for SM estimation using the DL method [29]. However, the approach used only one type of DDM (power analog) and no quantitative performance metrics for the SM retrieval model was presented. The contributions of this article are as follows.

1) A new DL framework with convolutional and fully connected neural network layers for enhanced SM estimation is developed that can utilize multiple DDMs jointly together with physical ancillary data relevant to SM estimation.
2) The proposed DL architecture is assessed under different train/test scenarios (spatial and year-based) using the SMAP mission's enhanced SM products at a 9 km × 9 km resolution over Continental United States (CONUS).
3) Training models for various size regions are studied and results for optimal region size and model complexity are presented.
4) Proposed DL-based SM approach using CYGNSS data and SMAP SM are compared to observations to International Soil Moisture Network (ISMN) locations and it is shown that proposed approach has a good dynamic range and produces similar characteristics to SMAP.
5) The mean unbiased root-mean-square difference (ubRMSD) between concurrent SM retrievals of the proposed approach and SMAP from 2017 to 2020 is 0.0333 $\mathrm{m}^3/\mathrm{m}^3$ with a correlation coefficient of 0.94 over fivefold cross validation and 0.0366 $\mathrm{m}^3/\mathrm{m}^3$ with a correlation coefficient of 0.93 over year-based cross validation. These results indicate an enhanced SM estimation performance compared to DL-based SM estimation techniques using CYGNSS data.

The rest of this article is organized as follows. Section II summarizes datasets used. Details on our approach and methodologies are provided in Section III. Results and discussions are presented in Sections IV and V. Finally, Section VI concludes this article.

II. Dataset

In order to effectively develop a DL-based retrieval algorithm for surface SM using CYGNSS observations, several datasets are utilized. The input selection for the retrieval process and each input's physical relationship to SM and GNSS-R sensitivity are described in the following sections. Different quality control approaches and multiresolution dataset combinations are explored to ensure consistent and accurate SM estimation.

A. *Cyclone Global Navigation Satellite System*

In this study, the CYGNSS Level-1 (L1) version 2.1 product is used, available at the NASA Physical Oceanography Distributed Active Archive Center (PO.DAAC).[1] The CYGNSS mission can record the reflected Global Positioning System (GPS) signals through a four-channel GNSS-R bistatic radar receiver using the eight microsatellites constellations. In CYGNSS L1 data, the DDM is one of the key measurements that represent the received surface power over a range of time delays and Doppler frequencies (bin-by-bin) for each observation frame [30]. DDMs are processed for nonsurface-related parameters through inverting the CYGNSS forward-scattering model in the L1 dataset and obtaining the surface's effective scattering area as well as BRCS images. The bin-by-bin measurements give 17 × 11 arrays of delay and Doppler spread in L1 data. In addition to DDMs, geometric and instrumental variables are also incorporated to provide complete acquisition information for each specular point, including features like incidence angle and distances between the GPS transmitter, CYGNSS receiver, and the specular point. The reflectivity can also be derived using L1 data through various methodologies based on some coherence and incoherence assumptions [18], [20], [31].

The approach of [28] is used for computing peak reflectivity Γ_{rl}, which is the peak value of each DDM corrected for gain, range, and incidence angle effects. P_r is called the uncorrected peak value of each DDM product, which is corrected for antenna gain, range, and GPS transmit power assuming a coherent reflection:

$$P_r = \frac{P_t G_t}{4\pi(R_{\mathrm{ts}}^2 + R_{\mathrm{sr}}^2)} \frac{G^r \lambda^2}{4\pi} \Gamma_{rl} \tag{1}$$

where P_t represents the transmitted right-hand circular polarized power, G_t is the gain of the transmitting antenna, R_{ts} is the distance between the transmitter and the specular reflection point, R_{sr} is the distance between the specular reflection point and the receiver, G^r is the gain of the receiving antenna, and λ is the GPS wavelength (0.19 m).

1) Delay–Doppler Maps (DDMs): One of the vital measurement of the CYGNSS mission is DDMs, which are mapping of

[1] [Online]. Available: https://podaac.jpl.nasa.gov/

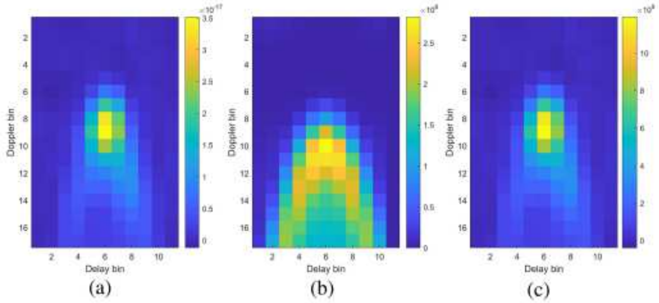


Fig. 1. Standard processed DDMs. (a) Analog power image. (b) Eff. scattering area image. (c) BRCS image.

received power caused from the observed surface to a delay–Doppler space. Delay is caused by the varied paths received signals taken from the scattering surface, whereas Doppler diversity is caused by the relative motions of the CYGNSS receiver, GPS transmitter, and Earth's surface. L1 DDMs are cropped from a larger "full" DDM using on board the CYGNSS delay–Doppler Mapping Instrument (DDMI) and have an 11-bin extent in Doppler at approximately 500-Hz reporting interval and a 17-bin extent in delay at approximately 0.25-μs reporting interval [32]. Thermal noise typically comes from the first three to four delay rows, resulting in an L1 DDM with a 3.5 chip delay extent on average, allowing each measurement to span an approximate surface area ranging from 70×70 km^2 to 100×100 km^2 depending on incidence angle [33]. Downlinked L1 DDMs are "raw counts" from uncalibrated instruments that go through two layers of calibration [34]. The raw counts are converted into power in watts and later processed to BRCS images by correcting the thermal noise effects, antenna patterns, instrument gain, and propagation losses [30]. Same as "raw counts" DDMs, the effective scattering area consist of a 17×11 element array of calibrated scattering area. Effective scattering area is an estimate of the actual surface scattering area that contributes power to each DDM bin after accounting for the GPS signal spreading function. It is calculated by convolving the GPS ambiguity function with the surface area that contributes power to a given DDM bin as determined by its delay–Doppler values, and the measurement geometry. The specular point bin location matches the specular point bin location in BRCS images. Each CYGNSS spacecraft generates DDMs' reflections simultaneously from four different transmitters at a 1-Hz rate. If the surface is rough (incoherent scattering), the power contributing to a given DDM can come from a location as large as 100 km or as small as a subkilometer body of water if the surface is smooth (coherent scattering) [33]. These processed DDMs are found as a form of NetCDF format along with "raw counts" DDMs on the CYGNSS website.

Fig. 1 shows processed DDMs product before normalizing each image. In addition, there could be some problematic DDMs in the dataset. We applied standard CYGNSS flags to remove these problematic DDMs. In addition, some images provide no value for effective scattering area. This is caused when specular point bin zero-based Doppler column is less than 4 or greater than 6. As part of data-quality control, such DDM images are eliminated from the dataset before the training and testing of models.

2) Spacecraft Ancillary Data: Besides DDMs, each CYGNSS L1 measurement contains an array of information regarding the spacecraft configurations, antenna factors, and geometry essential for reflection patterns [35]. The transmitter and receiver gain of the antenna is included as auxiliary input in each sample because the power gathered by the antenna is a function of the receiver/transmitter separation. The location of the measurements (latitude and longitude) is also included as ancillary spacecraft data. This feature allows a model to learn regionally particular SM behaviors, similar to how empirical relationships are developed when other SM products are generated. This feature is necessary for spatial comparison with other SM products, but it is not included in training our model. Some of the parameters from the spacecraft ancillary data such as transmitter and receiver ranges from specular point, receiver antenna gain, peak value of the DDM of the analog scattered power, and transmitter equivalent isotopically radiated power [36] are used to calculate the peak reflectivity via (1).

B. SMAP Radiometer SM Data

The SMAP Enhanced L3 Radiometer Global Daily 9-km EASE-Grid SM product is used to train and evaluate the proposed DL-based SM retrieval methodology. SMAP uses the *L*-band microwave radiometer to collect brightness temperature data and produces SM estimates. Though the SMAP SM product is generated at 36-km resolution, it has also a 9-km enhanced grid product by using Backus–Gilbert optimal interpolation techniques [37]. SMAP datasets containing the associated coordinates for the descending (A.M.) and ascending (P.M.) overpasses are combined to obtain daily SM results. With the help of a 1000-km swath width, a daily SMAP product can cover about 70% of all land areas within the CYGNSS coverage (±38° latitudes). The SMAP product also contains SM retrieval quality flags that indicate whether the SM retrieval is recommended or not. SMAP SM estimations can have an uncertain quality for several reasons: water body fraction, coastal proximity, urban area, precipitation, slope, and vegetation water content. The data are freely available through the National Snow and Ice Data Center website.[2]

In this study, SMAP SM with a 9-km EASE-Grid product is used where the network intends to predict single-valued SM. CYGNSS specular point location is used to obtain the SM value from the SMAP data over the CONUS regions. We consider the closest SMAP points from CYGNSS data within the 9-km grid and use the SMAP SM moisture as a label for those CYGNSS specular points. It is worth mentioning that we also consider the same day when the SMAP value is available.

C. International Soil Moisture Network

ISMN is used as ground truth information for additional evaluation of the developed DL model. Some previous studies showed

[2] [Online]. Available: https://nsidc.org/data/SPL3SMP_E/versions/3

different models that are trained using these SM stations [20], [21], [28]. Daily averaged in situ SM data of 170 sites selected from the ISMN dataset are used in order to compare our DL model performance. A uniform data format with preprocessing quality flags can be found in the ISMN global in situ SM database. A few sites in Asia, Australia, and Europe give both temporally and spatially collocated observations with CYGNSS data. The majority of these sites belong to North America. This study considers all available ISMN SM stations over CONUS within the CYGNSS spatial regions. ISMN sites that belong to the 2000-m altitude are not considered for comparison as CYGNSS measurements for high altitudes are unreliable. Detailed information about the ISMN is reported in [38] and [39]. The ISMN dataset is publicly accessible.[3]

D. Ancillary Data

Different geophysical parameters play essential roles in accurately predicting SM in conjunction with CYGNSS measurements. Vegetation density, surface roughness, soil topography, and soil texture are important geophysical parameters. We use some ancillary datasets as secondary features in our learning model:

1) Vegetation Water Content (VWC);
2) Normalized Difference Vegetation Index (NDVI);
3) elevation;
4) soil clay ratio;
5) water percentage;
6) slope;
7) soil silt ratio.

In order to characterize vegetation conditions, the 16-day composite NDVI is utilized from Moderate Resolution Imaging Spectroradiometer (MODIS) data. The NDVI data are spatially averaged to 3 km from its original 500-m resolution. This dataset is available in NASA Land Processes Distributed Active Archive Center.[4] The VWC is calculated using the NDVI and Land Cover Type (MCD12Q1) products using the same lookup table method as the SMAP VWC product [40].

The Digital Elevation Model GTOPO30 product (1-km resolution) is used to provide surface elevation information from the United States Geological Survey Earth Resources Observation and Science archive. The elevation data are also spatially averaged from 1 to 3 km. Topography is regridded spatially for each 3-km grid centered at the specular point, and averages of elevation and slope are utilized to reflect the underlying topographic complexities. Soil clay and silt ratios can be obtained from the Global Gridded Soil Information (SoilGrids) [41]. Soil profiles are discretized into several levels in the SoilGrids product, and the top layer data (5-cm depth from surface) is utilized for uniformity with the *L*-band signal penetration depth. For this investigation, the product is available at 250 m and is spatially aggregated onto 3 km.

A 30-m Global Surface Water Dataset from the Joint Research Centre [42] is used to identify the presence of a surface inland water body. The percentage of 30-m grids within each 3-km grid showing the existence of either permanent or seasonal water is calculated, and this number is employed throughout the retrieval algorithm's quality control phase. Table I shows all the auxiliary features that will be used for proposed model.

E. Quality Control Mechanisms

This analysis considers CYGNSS observations from March 2017 to November 2020 available in the CONUS region. Before performing SM retrieval, it is essential to conduct critical screening for the quality of CYGNSS data in underlying land surface conditions. Several quality control criteria need to be applied to CYGNSS observations and auxiliary data. The specific flags (*S*-band powered up, substantial spacecraft attitude error, blackbody DDM, DDM test pattern, poor confidence GPS EIRP estimate) are maintained in this study [17], [31]. Observations with an incidence angle higher than $\pm 65^\circ$ are eliminated in order to avoid noisy DDMs [24]. To prevent high-altitude measurements, observations with a DDM peak value outside of 5–11 delay bins are excluded from the dataset. For SM retrieval products, open water near the specular point is a critical source of inaccuracy. Due to the highly strong coherency over water surfaces, the power of a forward-scattered signal radiating from a water's surface is usually several orders of magnitude higher than a signal scattered from soil [43]. SM recovery close to the water bodies becomes infeasible if the surface water within the CYGNSS region is sufficiently large. As a result, a CYGNSS observation is removed if more than 2% of the 9-km grid centered on a specular point is covered with permanent or seasonal water. Additionally, CYGNSS readings that fall over forested areas with $\text{VWC} > 5\ \text{kg/m}^2$ (dense vegetation canopy) are also eliminated [44]. CYGNSS observations before December 2017 that are above 600 m from the surface are masked out due to the altitude limitation of CYGNSS L1 data for the specified time period [26]. After applying all the quality control masks, we have more than 18 million specular samples over CONUS regions. It is essential to mention that each specular points generate three different types of DDM. So, we will have a total of three times of 18 million DDM images for our DL model for training and testing.

III. Methodology

Convolutional neural networks (CNN) have been extensively utilized in computer vision applications and are shown to learn relevant features for the classification/regression tasks directly from the images. The application of CNN to DDMs is particularly fascinating since it provides ability to learn directly from the DDM itself. Currently, some existing machine learning (ML) models have been developed using designed features computed from DDM [20], [25] and estimate SM. Although these ML-based models provide promising results, in this study, we hypothesize that a CNN-based model can extract further features that will enhance SM estimation leading to a higher quality CYGNSS-based SM product. The complex information included in the entire 2-D DDM will be helpful under various

[3] [Online]. Available: http://ismn.geo.tuwien.ac.at

[4] [Online]. Available: https://lpdaac.usgs.gov/products/myd13a1v006/

TABLE I
PHYSICAL FEATURES CONSIDERED FOR THE DL-BASED SM RETRIEVAL MODEL

Input group	Feature name	Description	Category
CYGNSS	1. Peak reflectivity	Reflectivity calculated via 1	CYGNSS
	2. SP incidence angle	Incidence angle of specular point	
Topography	1. Elevation	Mean elevation for each specular point 3-km grid	Non-CYGNSS
	2. Slope	Mean Slope for each specular point 3-km grid	
	3. Water percentage	Mean water percentage for each specular point 3-km grid	
MODIS	1. NDVI	Mean normalized difference vegetation index	
	2. VWC	Mean vegetation water content	
Soil texture	1. Soil clay ratio	Mean clay proportion for each specular point 3-km grid	
	2. Soil silt ratio	Mean silt proportion for each specular point 3-km grid	

conditions, and the DL-techniques are the state-of-the-art approach to retrieve from DDMs. We explored the CNNs and fully connected neural layers to determine the complex relations between DDMs, ancillary data, surface attributes, and the SM.

We consider CNN as our core DL model, where the primary inputs come from multiple types of CYGNSS DDM images and ancillary data. We develop a supervised learning framework, where the model maps a set of input features to an SM value, which is the final output of the proposed architecture. The dataset utilized to train and test the developed model is constructed using the CYGNSS, SMAP, and ancillary data sources from CONUS during April 2017 to December 2020 as detailed in Section II.

A. Data Normalization

The analog power DDM image pixel values range from 10^{-16} to 10^{-18} W roughly, and then, the images are normalized for numeric stability. Normalization is performed by calculating the mean and standard deviation for the entire pixels and standardizing these values to achieve zero mean/unit variance. Effective scattering area, BRCS images, and the other ancillary inputs are normalized in the same way.

B. Design of the DL Architecture

The proposed DL architecture consists of three major parts; the convolutional layers, the concatenation layers, and the densely connected layers. Convolutional layers are used to extract the features from multiple DDM images. At the end of the convolutional layers, features are flattened and concatenated with other auxiliary feature inputs. A two-layer fully connected network is used to map the concatenated features to an SM value.

The three normalized processed DDMs are the primary input of the convolutional network. Each image type is given in a different channel. DDMs are of size 17×11 and combining three types of DDM images makes an input of $17 \times 11 \times 3$. Three convolutional layers are used, followed by a max pool layer. Each convolutional layer consists of 3×3 kernel with no paddings and stride of 1. The number of filters for the convolutional layers are 32, 64, and 128 respectively. After each convolutional layer, we have used a batch normalization layer followed by a ReLU activation layer. After the convolutional layers, a max-pooling layer with a kernel size of 5×11 is applied. A flattening layer is used to flatten the extracted image features into the vector format. The total number of extracted features from the DDM images is 128. After extracting the feature from the DDM images, we have concatenated the nine auxiliary features with that 128 extracted features. The combined vector of 137 total features are then fed into a neural network with two layered neural networks having two dense layers with 50 neurons with clipped ReLU activation in each layer. The final layer of this network is a regression layer with a sigmoid activation function before the final output. Fig. 2 illustrates the overall architecture of the proposed CNN network from the input DDMs to the output SM with details on different layers.

C. Training the CNN

This section will discuss how our model is trained based on the input and the label data. Our model has two main categories of input data. The primary inputs are the three different types of processed DDMs, being analog power, effective scattering area, and BRCS images, with a size of 17×11 each. The second category of inputs are the nine different types of ancillary data that are based on CYGNSS, topography, MODIS, and soil texture data. The list of these features are provided in Table I. After the dataset for DDM images and ancillary features are constructed as detailed in Section II and the quality control mechanisms are applied, the DL architecture is ready to be trained.

Fig. 3 shows the overall training process of the proposed DL model. The model maps the input DDMs and ancillary data to an SM value. In order to update the parameters of the DL model in the training phase, labeled SM values are needed for the corresponding inputs. SMAP SM data are used as the labels, as detailed in Section II-B. The model parameters are determined in order to minimize the root-mean-square loss between predicted and label SM values. During the training of the proposed model, a version of gradient descent based backpropagation approach, root-mean-square propagation (RMSprop), is used as the main model optimizer. RMSprop uses a decaying average of partial gradients in the adaptation of the step size for each parameter and it helps to accelerate the optimization process by decreasing the number of function evaluations required to reach the optimal point. A piece-wise learning rate schedule is used, which helps to decay the schedule constantly. We set the initial learning rate at 0.01 and gradually decreased the learning rate (10-times) after every 50 epochs. A total of 250 epochs are used for the training process. The epoch number is selected based on the convergence pattern of our model.

In order to speed up the training process, we have chosen a mini-batch size of 50 000. This big batch size is used to load

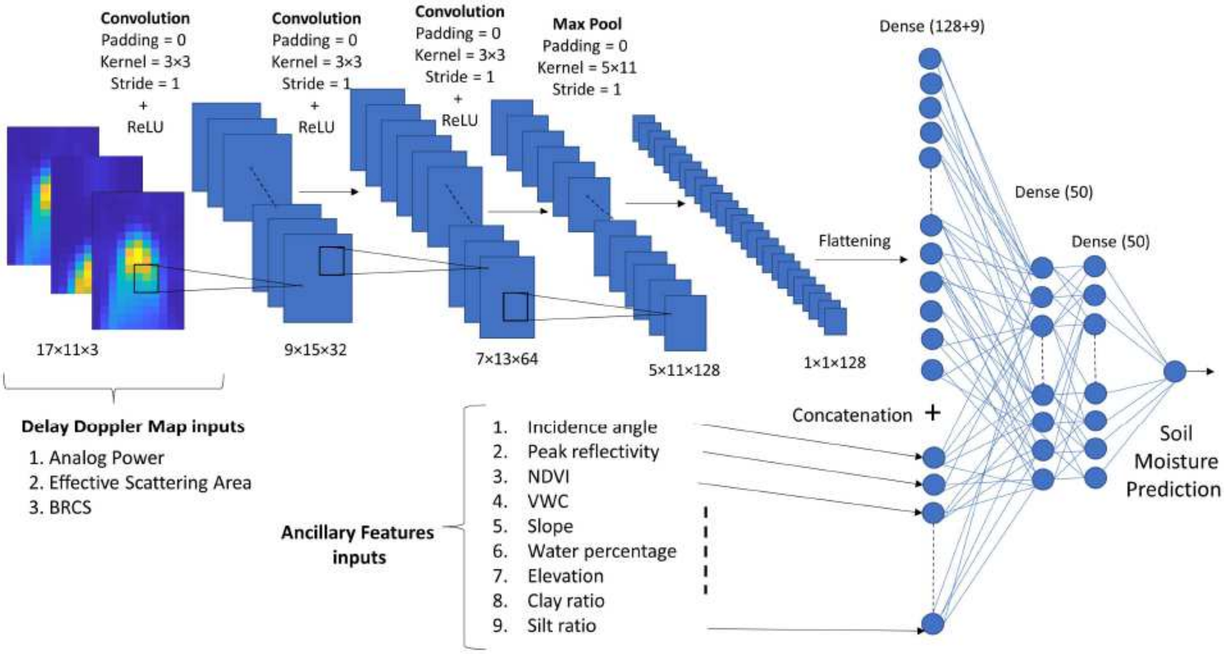


Fig. 2. Network overview of DL-CYGNSS SM estimation using DDMs and ancillary features.

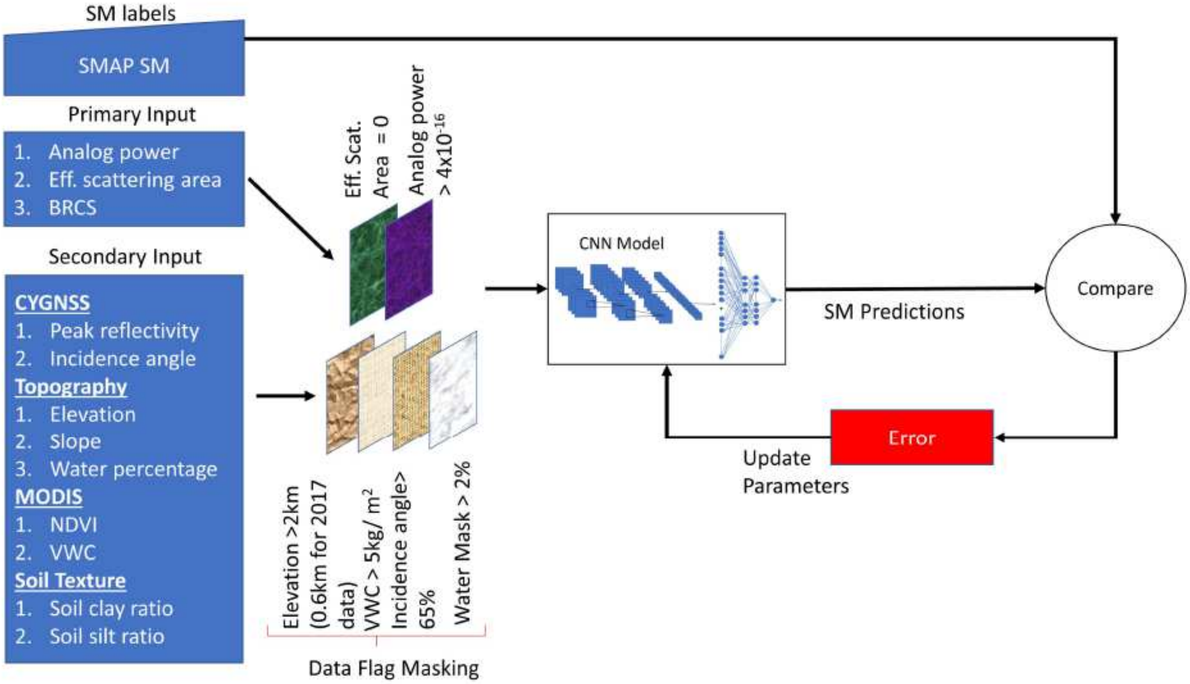


Fig. 3. Overall DL-CYGNSS training process.

the data in the memory and process it at a time. We have done different analysis depending on the training/testing scenarios and depending upon the model used, the mini-batch can be varied. More detail is discussed in Section III-D. All the required computations are carried out using the DL toolbox of MATLAB R2021a software over a machine with Intel(R) Xeon(R) CPU E5-2643 with three NVIDIA Titan RTX GPUs and 128-GB memory.

D. Training Models for Different Size Regions

The generic approach in a DL model could be to train a single model that will predict SM at any given location, given the related inputs. Such a model can be learned using all the available training data. Since a single model predicts SM everywhere, memory requirements are low. However, the model should be able to address the whole complexity and variations over the world and training a model with huge size of dataset is highly computationally complex. Another possibility we propose and test here is to divide the region into smaller clusters and learn a different DL model for each cluster using the data from that cluster region [22]. 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. To analyze this tradeoff, in addition to learning a single model for the CONUS region, we have constructed different clusters with 36, 72, and 144 km grid sizes. We divided the CONUS region into contiguous geographical grids with the specified grid sizes and used the CYGNSS observations belonging to that grid to train and evaluate the performance metrics. For example, for the 36-km cluster case, for each 36 × 36 km grid all 9-km cells falling within this grid are gathered, and one single model is learned with all the training data samples within that grid. When a prediction is made, the DL model for this grid is used to the prediction. One SM value is predicted for each 9-km cell and the same DL-model is used to predict SM for each cell in that cluster. In the extreme case, one can also learn DL models for each 9-km grid; however, the average number of data samples for each 9-km grid is not high enough to facilitate learning of a DL model with many model parameters currently. This is the reason we tested a minimum cluster size of 36 km. Even with the current cluster sizes, we set a threshold sample number of 300 and if any grid has less than 300 samples, we skip that grid for training and testing. In addition, a mini-batch size of 5000 is used in training of clustered models, while single cluster model uses mini-batch size of 50 000 samples.

IV. Results and Discussion

In this section, we provide and discuss quantitative performance of the proposed model under different validation strategies. The *K*-fold and year-based cross validation is used to evaluate the DL model performance against the SMAP SM within the CYGNSS coverage over CONUS. SMAP SM observations are also remotely sensed estimates like CYGNSS, and they have their own error uncertainties. Besides calculating the model performance, we also compare our predicted result with the SM station (ISMN sites) and present the temporal variation of different SM products. The error and correlation coefficient maps are generated that evaluates the model and its improvement. We further analyze the performance metrics for different land cover types to better understand learning performance of the proposed model for different land covers.

A. Evaluation Metrics and Validation Strategy

Several evaluation metrics are used to assess the model performance quantitatively. The trained DL model from Section III-B is tested within the CYGNSS coverage, and its performance are evaluated using SMAP SM predictions. The performance metrics used in this evaluation are the root-mean-square error (RMSE), unbiased RMSE (ubRMSE), and correlation coefficient (*R*-value). In addition, the root-mean-square difference (RMSD) is also computed for the proposed DL-CYGNSS and the SMAP SM product comparison as the label SMAP SM data might contain errors that cannot be considered as the "True" SM values [45]. The "RMSE" term is basically used for in situ evaluation as those measurements are considered ground truth data for SM. In our case, we compare our results also with the ISMN sites and provide RMSE and ubRMSE metrics in that case. We have evaluated our models using *K*-fold cross validation with $K = 5$ folds. The *K*-fold approach is a highly popular and common type of validation technique in order to evaluate the performance of a model. In *K*-fold cross validation, the total data are divided into *K* number of folds, and then, the model is trained using $(K - 1)$ folds and tested over the unused fold. This approach guarantees separation of training and test data and tests every fold. After the predictions of the DL-CYGNSS is obtained, different metrics are computed in order to evaluate the model performance.

B. Performance Analysis of Different Clusters

In this section, we discuss the performance of different clustering sizes. Table II shows the overall SM prediction performance derived via the proposed DL-CYGNSS approach for different cluster sizes together with the number of DL models and average number of samples for each model. We have evaluated DL models over the CONUS region, where the model is trained and validated using a fivefold cross validation. While for the one-cluster case, we learn a single model for the CONUS region with 18.6 million data samples, as the clusters get smaller the learned number of models increases and average number of data samples for each model decreases. For the 36-km cluster case, we learn a total of 3190 different DL models, where each model is trained/tested over an average of 5800 data samples.

It can be seen in Table II that the SM performance increases for smaller clusters leading to lower SM estimation errors and higher correlations. While the 36-km cluster case provides the best SM estimation performance of compared cases with a mean ubRMSD value of $0.0362\ \mathrm{m}^3/\mathrm{m}^3$ and a correlation coefficient of 0.93, its results are close to the results obtained from 72-km clusters. A significant increase in SM estimation performance is observed from one cluster to 144-km cluster case, where the mean ubRMSD is reduced to 0.0417 from $0.0482\ \mathrm{m}^3/\mathrm{m}^3$, and the *R*-value is increased to 0.90 from 0.85. A smaller but still important level of performance increase is also observed in transition from 144- to 72-km clusters. The performance change between 72- and 36-km clusters are minor, indicating a convergence of performance for tested cluster sizes. Our observation is that proposed DL model can learn the characteristics of smaller cluster regions better and clusters with 36- or 72-km grid sizes offer a good tradeoff between performance and number of models learned.

C. Year-Based Cross-Validation Performance

In addition to the presented fivefold cross validation, it is important to assess the performance of the proposed method

TABLE II
PERFORMANCE METRICS FOR DIFFERENT DL-CYGNSS MODELS USING FIVEFOLD CROSS VALIDATION

	Number of models	Average no. of samples for each model	RMSD ($m^3\,m^{-3}$)	mean ubRMSD ($m^3\,m^{-3}$)	median ubRMSD ($m^3\,m^{-3}$)	*R*-value [−]
36-km cluster	3190	5.8e+3	0.0390	0.0362	0.0352	0.93
72-km cluster	888	2.08e+4	0.0407	0.0366	0.0353	0.92
144-km cluster	242	7.64e+4	0.0480	0.0417	0.0410	0.90
One cluster	1	1.86e+7	0.0580	0.0482	0.0470	0.85

TABLE III
PERFORMANCE METRICS GRID-WISE DL-CYGNSS MODEL USING YEAR-BASED CROSS VALIDATION

		36-km cluster			72-km cluster			144-km cluster		
Validation year	Samples (millions)	RMSD (m^3m^{-3})	ubRMSD (m^3m^{-3})	*R*-value [−]	RMSD (m^3m^{-3})	ubRMSD (m^3m^{-3})	*R*-value [−]	RMSD (m^3m^{-3})	ubRMSD (m^3m^{-3})	*R*-value [−]
2017	1.80	0.0462	0.0415	0.91	0.0465	0.0419	0.89	0.0565	0.0419	0.87
2018	3.42	0.0410	0.0377	0.93	0.0457	0.0398	0.91	0.0528	0.0424	0.88
2019	5.54	0.0393	0.0359	0.93	0.0472	0.0405	0.90	0.0537	0.0433	0.87
2020	7.27	0.0378	0.0333	0.94	0.0438	0.0358	0.92	0.0473	0.0380	0.91

under a year-based cross-validation scenario. We have a total of four years of data from 2017 to 2020 and for the year-based cross validation, we have trained the model using three-year data (such as 2017, 2018, and 2019) and tested the model over all data from another year (such as 2020). This way, we have tested each year and calculated the performance metrics of RMSD, ubRMSD, and *R*-value. This analysis is also applied for 36-, 72-, and 144-km grid-wise models. The observed performance for each validation year and cluster size are presented in Table III. It can be observed that DL models trained over 2017–2019 and tested on all 2020 data can provide an SM estimation performance of mean ubRMSD 0.033 m^3/m^3 with a correlation coefficient of 0.94 under 36-km clusters. While this is the best obtained result under year-based cross validation, testing with other years of data show only slightly higher levels of estimation error with the exception of testing on 2017 data. Different cluster sizes provide a similar trend of performance as discussed in the previous section. We consider that the more refined CYGNSS data products in the recent years being one possible reason for increased performance for the 2020 performance. We think that year-based cross-validation approach is a more practical, since DL models can always be trained on previous years of data and those models can be used to provide SM predictions all over the next year. Obtaining an ubRMSE of 0.033 m^3/m^3 against SMAP under this practical training/testing scenario shows the high potential of CYGNSS measurements for SM estimation.

D. Performance Comparison via ISMN Sites

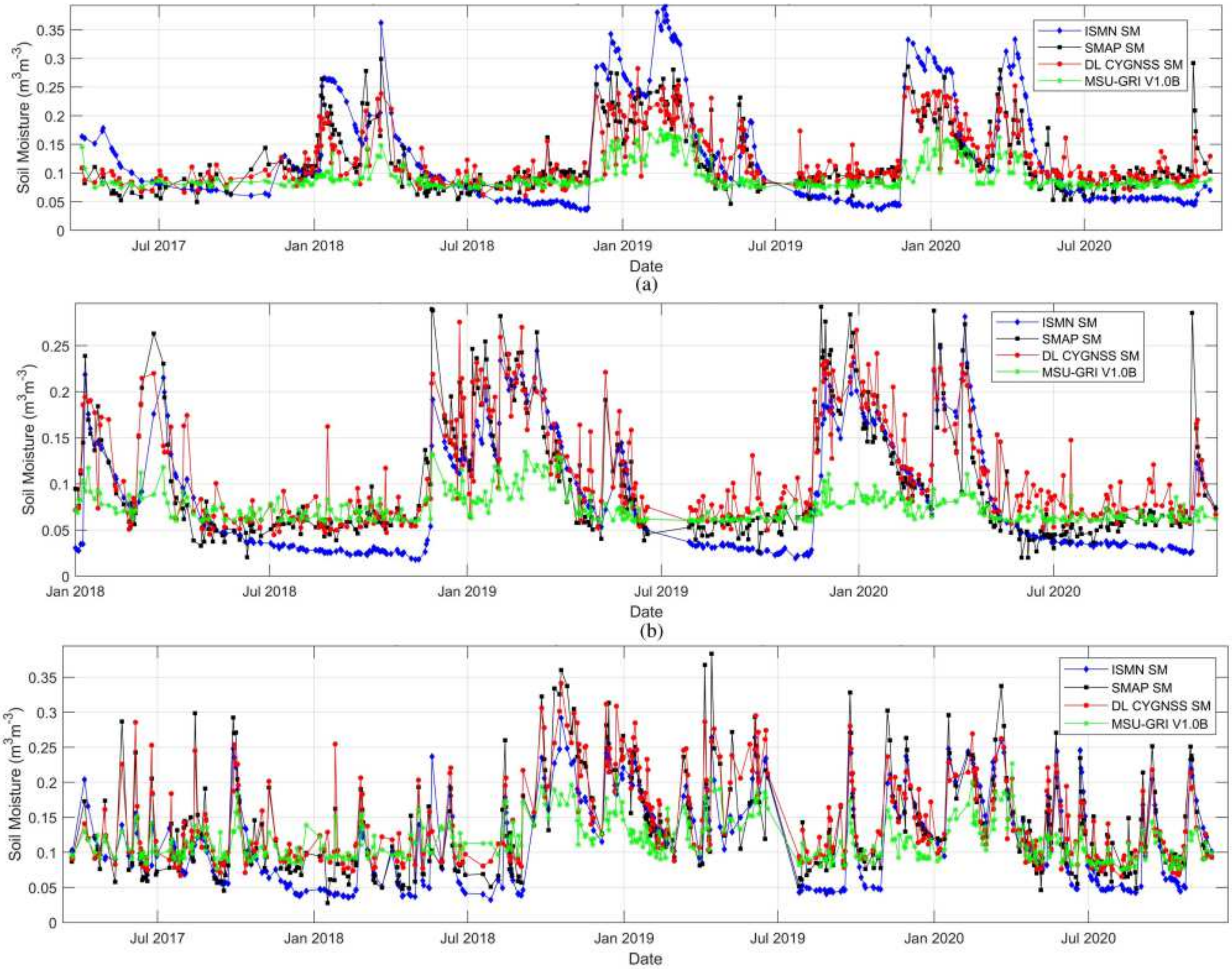
In order to additionally evaluate the performance of our proposed model, we have compared SM predictions of the proposed DL-CYGNSS and the SMAP with the SM observations at ISMN sites. The DL-CYGNSS model is trained using SMAP SM labels and evaluated against SMAP in previous sections under different cross-validation approaches. Here, we compare SMAP versus ISMN sites as well as DL-CYGNSS versus ISMN sites. This provides an evaluation of both SMAP and proposed DL-CYGNSS against an independent SM observation source, the ISMN sites' reference SM. Table IV shows the mean RMSE, ubRMSE and *R*-values when ISMN observations are compared to SMAP and DL-CYGNSS predictions, respectively. All the metrics are evaluated at each individual site and averages out of all sites are reported The comparisons were made with the ISMN sites that belong to a 9 × 9 km SMAP grid. Proposed DL-CYGNSS provides SM estimations on the same SMAP grid locations. We consider the SM predictions on the same day and same grid location to compare the results. In the CONUS region where we predict SM values, we have 89 ISMN sites to compare, and the average number of samples per site is approximately 207. It can be seen from Table IV that both SMAP and DL-CYGNSS predictions have similar RMSE and ubRMSE results when compared to ISMN observations, while SMAP is slightly outperforming DL-CYGNSS. This is expected considering that the DL-CYGNSS is trained using SMAP data. SMAP also produces a better *R*-value. Due to small number of samples per site, the bias is comparably higher and lower *R*-values for both SMAP and DL-CYGNSS are observed.

In addition to calculating the overall performance metrics, it is essential to understand the performance of DL-CYGNSS on following temporal SM variations. Here, three representative ISMN sites have been selected and SM observations recorded at these sites from 2017 to 2020 are shown in Fig. 4. All these sites belongs to Soil Climate Analysis Network (SCAN) and Hydraphobe Digital Sdi-12 (2.5 V) is used as an SM sensor, These sites generally provide soil temperature, precipitation, air temperature, and SM. The proposed DL-CYGNSS and SMAP SM predictions for the same time period are also illustrated on the same figure. In addition, the SM predictions of a recent estimation approach called Mississippi State University's Geosystems Research Institute (MSU-GRI) CYGNSS SM product [22] are also compared. The product uses ML with handcrafted features within a random forest framework to map quasi-global SM from CYGNSS measurements, and is publicly available.[5] Two versions of the product are available, i.e., v1.0a and v1.0b. For v1.0a, the ML model is trained using ISMN sites and for v1.0b, the model is trained and tested using SMAP. We consider the MSU-GRI v1.0b SM product for this temporal performance comparison as both algorithms are trained using SMAP data.

[5] [Online]. Available: https://www.gri.msstate.edu/research/ssm/

TABLE IV
DL-CYGNSS AND SMAP SM COMPARISON WITH ISMN SITES

ISMN vs SMAP SM			ISMN vs CYGNSS DL SM		
mean RMSE (m^3m^{-3})	mean ubRMSE (m^3m^{-3})	mean R-value [–]	mean RMSE (m^3m^{-3})	mean ubRMSE (m^3m^{-3})	mean R-value [–]
0.0960	0.0517	0.70	0.1029	0.0581	0.51



(a)

(b)

(c)

Fig. 4. Time-series examples of daily averaged SMAP, DL-CYGNSS, and MSU-GRI V1.0B SM predictions against selected ISMN sites with a moderate performance. (a) MonoclineRidge. (b) CochoraRanch. (c) KnoxCity. The summary table of results from all compared sites are provided for the same time period. (a) Site Name: MonoclineRidge, ubRMSE:0.0633, R: 0.80 (CYGNSS vs (ISMN). (b) Site Name: CochoraRanch, ubRMSE:0.0352, R: 0.83 (CYGNSS vs ISMN). (c) Site Name: KnoxCity, ubRMSE:0.0404, R: 0.80 (CYGNSS vs ISMN).

For these representative sites, the DL-CYGNSS predicted SM closely follows the temporal trend of the ISMN SM observations and closely describes the precipitation events and the dry-down process.

Fig. 4(a) shows the time-series analysis for site "MonoclineRidge" over the period from 2017 to 2020. As mentioned earlier, this site is in the SCAN network and located in the western part of the CONUS region. It started to give SM data since 2014 on a daily basis. This site gives us a high dynamic SM range, so we can compare the ability of our approach with a varied SM range. In the figure, the blue line shows the ISMN SM, the red line shows our CYGNSS DL SM and gives an ubRMSE $0.0632\ m^3/m^3$ and the correlation coefficient is 0.80 between these two SM. The black and green lines show the SMAP SM and MSU-GRI SM product, respectively. From this time-series analysis (site—MonoclineRidge), we can observe that the SM value is higher during the beginning of the year and stays high over May. For 2019 and 2020, we have seen that all the SM value increases starting from the beginning of December and staying high to the end of May. It shows a good agreement over the observation period. Though both our proposed DL model and publicly available MSU-GRI product are trained and tested using enhanced SMAP SM products, the proposed DL approach demonstrated improvements against the MSU-GRI

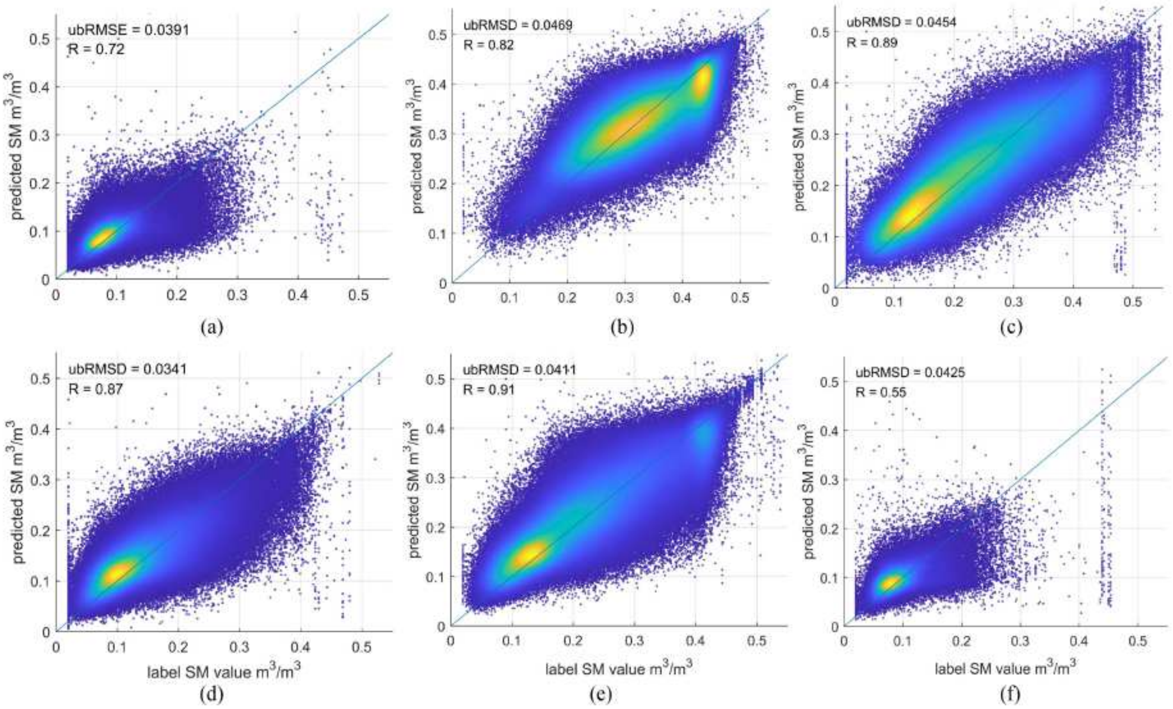


Fig. 5. Scatter plots of DL-CYGNSSS SM retrievals for different types of land cover. (a) Land Cover: Shrub. (b) Land Cover: Woody. (c) Land Cover: Savanna. (d) Land Cover: Grass. (e) Land Cover: Crop. (f) Land Cover: Barren.

product that uses designed features only (given in Table I) using the random forest model. The site "CochoraRanch" is also in the SCAN network and located in the western part of the CONUS. It has been producing SM data since 2012. In Fig. 4(b), the ubRMSE and R-value for the site "CochoraRanch" is 0.0352 $\mathrm{m}^3/\mathrm{m}^3$ and 0.83, respectively, comparing the CYGNSS SM with ISMN SM. We have seen similar SM patterns for this site during different periods. Though the SM value is below 0.1 $\mathrm{m}^3/\mathrm{m}^3$ most of the time, it gets as high as 0.3 $\mathrm{m}^3/\mathrm{m}^3$ at the beginning of each year. The soil is relatively dry (SM $<$ 0.3 $\mathrm{m}^3/\mathrm{m}^3$) for the growing season (from May to September). The "KnoxCity" [see Fig. 4(c)] provides a little different SM pattern during the whole observation period. The "KnoxCity" site is located in the northern part of the Texas area and it has been producing SM since 2013. Comparing with previous two sites, it is giving low dynamic SM value all the year around. Both SMAP and CYGNSS follow the similar pattern with the ISMN site. The ubRMSE and correlation coefficient between CYGNSS SM and ISMN SM are 0.0404 $\mathrm{m}^3/\mathrm{m}^3$ and 0.80, respectively. There are some number of sites that give relatively high ubRMSE and low correlation coefficient value. In order to check the both space-borne mission's performance with ISMN SM, we compare the average site results side by side, which is demonstrated in Table IV. As we have already mentioned, we have considered total 89 ISMN sites that belong to the specific SMAP grid and also we consider the same day and same grid SM predicted using the DL model.

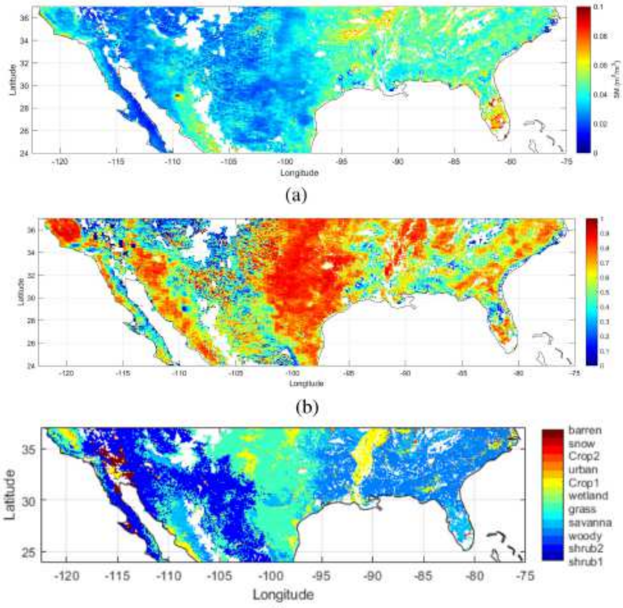
E. Performance Comparison for Different Land Covers

It is essential to quantify the impact of diverse land cover conditions on proposed DL-CYGNSS SM prediction model performance, because land cover type is a critical parameters affecting both GNSS-R observation and SM retrieval performance. The SM predictions are evaluated under various land cover categories. Fig. 5 shows the scatter plots between the predicted and labeled SM data for different land cover types.

Fig. 5(a) shows the scatter plot between predicted and labeled SMAP SM in shrub-land cover area. This land cover region provides ubRMSD 0.039 $\mathrm{m}^3/\mathrm{m}^3$, and correlation coefficient 0.72, where most of the points are below the 0.30 $\mathrm{m}^3/\mathrm{m}^3$ SM. Woody land plot shown in Fig. 5(b) gives slightly higher ubRMSD of 0.046 $\mathrm{m}^3/\mathrm{m}^3$ but provides a comparatively better R-value of 0.82. This land cover has a higher label SM value ($\mathrm{SM} > 0.25\,\mathrm{m}^3/\mathrm{m}^3$) indicating good prediction performance for a wide range of SM for the proposed approach. The Savanna land cover shown in Fig. 5(c) gives an overall ubRMSD and R-value 0.045 $\mathrm{m}^3/\mathrm{m}^3$ and 0.89, respectively. The lowest estimation error is achieved in the grassland cover with an ubRMSD of 0.034 $\mathrm{m}^3/\mathrm{m}^3$, as shown in Fig. 5(d). The highest R-value is achieved on crop land cover with 0.91, while barren results in the lowest R-value with 0.55. The number of data points for different land covers vary and comparably lower R-values in barren and shrub land covers are partially due to lower number of data for those types.

F. Results Over CONUS and Comparisons

In this section, the overall performance of our proposed model over CONUS is illustrated. From the presented results using fivefold and year-based cross validation, the 36-km cluster model outperforms the other models. Fig. 6 shows the maps of ubRMSD and correlation coefficient over CONUS on averaged 9-km grids using the DL-CYGNSS model using 36-km clusters.



(a)

(b)

(c)

Fig. 6. (a) ubRMSD and (b) correlation coefficient map for 36-km DL-CYGNSS model test result (4-years averaged) and (c) reference land cover map in the CONUS region.

From the ubRMSD map in Fig. 6(a), it can be seen that the SM predictions are more accurate for the low vegetated regions, but the error is higher for the relatively more vegetated eastern part of the CONUS. The correlation coefficient map in Fig. 6(b) shows that proposed approach provides highly correlated SM predictions with SMAP over important part of CONUS aligned with the land cover types shown in Fig. 6(c) and the analysis from Section IV-E. The scatter plot for all DL-CYGNSS predictions over CONUS using the fivefold cross validation and 36-km cluster model is shown in Fig. 7. It can be seen that the overall results show a very good correlation with SMAP with an *R*-value of 0.93.

We have already demonstrated the performance comparison with MSU-GRI product in terms of their temporal variation. The quantitative performance comparison between proposed DL-CYGNSS and the MSU-GRI SM [22] is given in Table V. This comparison is made over the CONUS region. The DL-CYGNSS model outperforms the compared ML model in all compared metrics including RMSD variants and *R*-value. MSU-GRI product uses handcrafted CYGNSS and ancillary features with a random forest ML model; however, our approach utilizes full DDM images and directly learns from them using a CNN structure. It is clear that learning from entire DDMs for SM estimation helps to reduce SM prediction errors.

G. Performance Using Different Inputs Strategies

In order to validate the idea of bringing the DDMs for estimating the SM, we have examined scenarios where we learn from different set of inputs. Table VI shows the performance of all tested input scenarios. First, we consider only using the nine features in our DL-CYGNSS model without the CYGNSS DDMs. This input set uses a classical neural network that contains two fully connected layers with 50 hidden units in each layer. The mean ubRMSE reached 0.0447 $\mathrm{m}^3/\mathrm{m}^3$, and correlation coefficients reached 0.73 for this input set. Compared to using all three DDMs jointly together with the nine features, which provide an ubRMSE of 0.0362 m^3/m^3, and correlation coefficient of 0.93, we observe that learning from DDMs provides a significant performance enhancement. Then, we test using a single DDM in our model instead of using all three DDMs jointly. Here, we would like to test the most effective single DDM and performance change between using a single DDM versus all DDMs jointly. All DL-based results are for models trained on 36-km cluster regions. We train and test our model using a single DDM image inputs of analog power, effective scattering area, and BRCS images. For the single DDM image cases, the analog power DDM case results the best overall performance with mean ubRMSE of 0.0372 $\mathrm{m}^3/\mathrm{m}^3$, and correlation coefficient reaching 0.93. While BRCS or effective scattering area DDMs provide closer and high *R*-values of 0.92, they provide slightly worse mean ubRMSE values of 0.039 $\mathrm{m}^3/\mathrm{m}^3$. While using all three DDMs provide the best overall results, using analog power DDM as a single image with DL provides slightly lower results, which could provide a tradeoff between performance and data storage/memory requirements. In addition, we see low bias if we consider DDMs instead of using only the features. Single DDM or multi-DDMs with ancillary features provide bias values very close to each other, but the nine features model provides a high bias value.

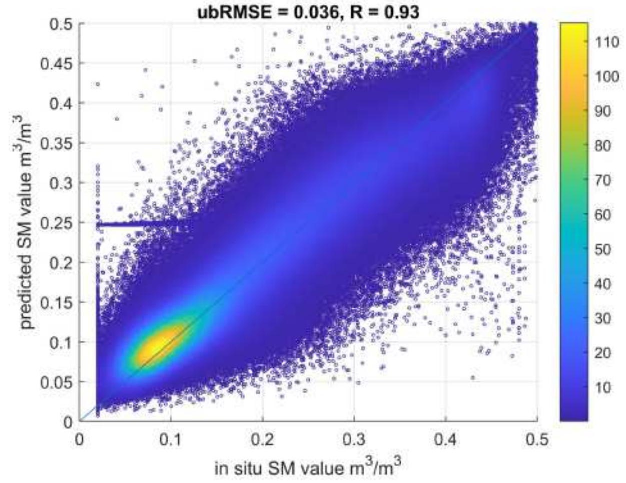


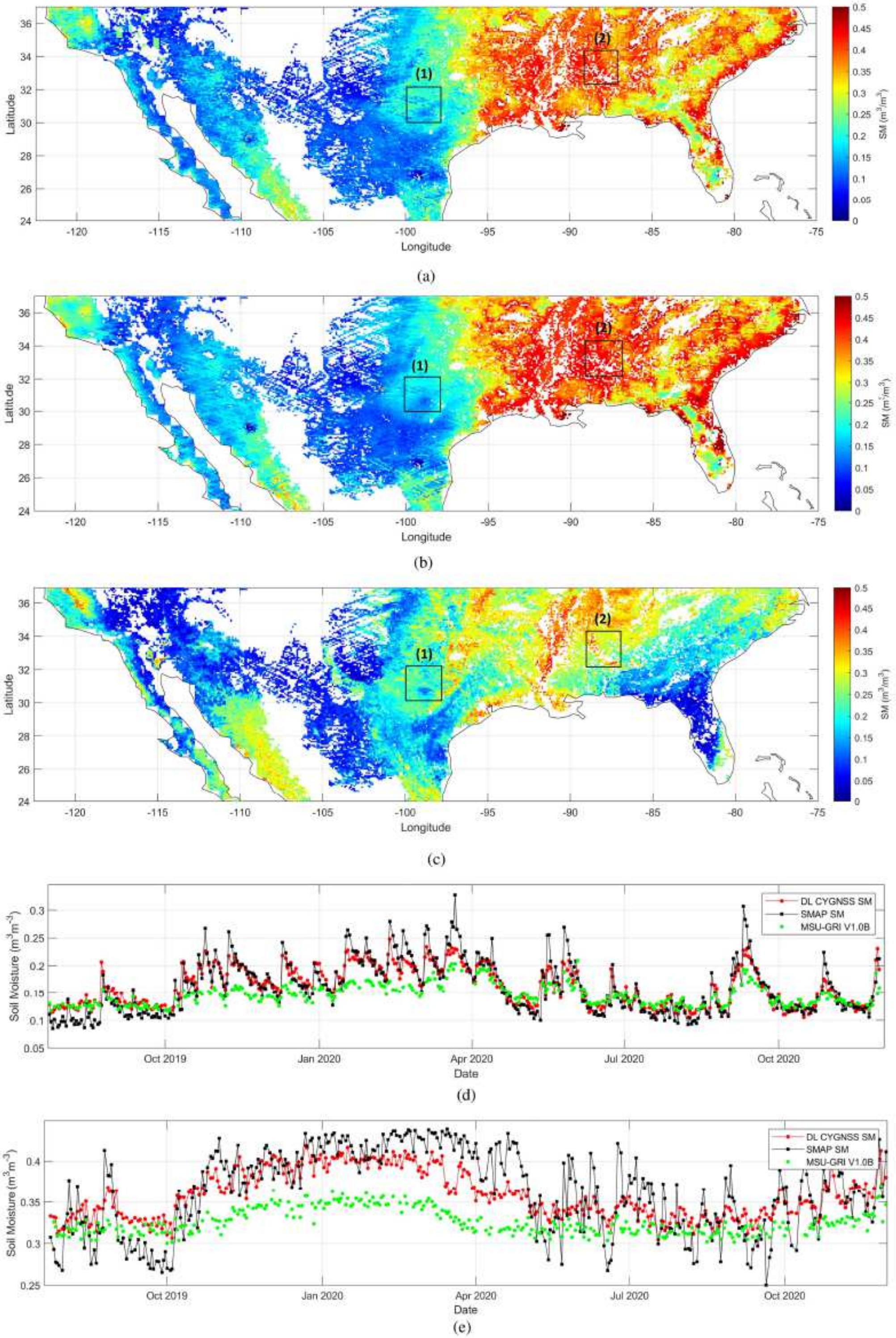
Fig. 7. Scatter plot of the predicted SM versus targeted SM for the 36-km cluster DL-CYGNSS model.

V. Discussion

The space-borne GNSS-R observations for SM retrievals have become popular with the hydrology community. This interest becomes particularly accelerated when the space-borne GNSS-R are available such as CYGNSS and TechDemoSat-1 (TDS 1) [46]. The main advantage of using this is the high spatial coverage over the Earth's surface, having sufficient measurement capabilities. Fig. 8 shows the spatial performance of DL CYGNSS SM compared to SMAP and MSU-GRI SM products. The results are averaged over the month of January 2020 in

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT SM PRODUCT OVER THE CONUS REGIONS (72-KM CLUSTER)

Models	RMSD ($m^3 m^{-3}$)	mean ubRMSD ($m^3 m^{-3}$)	median ubRMSD ($m^3 m^{-3}$)	R-value
Our proposed DL SM	0.0407	0.0366	0.0353	0.93
MSU-GRI SM product [22]	0.0518	0.0434	0.0433	0.91



(a)

(b)

(c)

(d)

(e)

Fig. 8. Monthly averaged SM predictions. (a) DL CYGNSS SM—9 km, (b) SMAP SM—9 km, and (c) MSU-GRI V1.0B—9 km for the month of January 2020 with the 9 × 9 km spatial resolution using 36-km cluster in CONUS. Temporal comparison of data products over different selected (225 × 225 km) regions (d) for location-(1) and (e) for location-(2).

TABLE VI
PERFORMANCE METRICS FOR DIFFERENT INPUTS STRATEGIES USING FIVEFOLD CROSS VALIDATION FOR 36-KM CLUSTER

Input selection	RMSD $(\mathrm{m}^3\mathrm{m}^{-3})$	mean bias $(\mathrm{m}^3\mathrm{m}^{-3})$	mean ubRMSD $(\mathrm{m}^3\mathrm{m}^{-3})$	median ubRMSD $(\mathrm{m}^3\mathrm{m}^{-3})$	R-value [–]
9 Features (Table. I)	0.0525	0.0110	0.0447	0.0444	0.85
DDMs (Analog Power) + 9 Features (Table. I)	0.0425	0.0015	0.0372	0.0359	0.93
DDMs (Eff. Scat. area) + 9 Features (Table. I)	0.0430	0.0013	0.0390	0.0386	0.92
DDMs (BRCS) + 9 Features (Table. I)	0.0420	0.0015	0.0399	0.0389	0.92
DDMs (Analog Power, Eff. Scat. area, BRCS) + 9 Features(Table. I)	0.0390	0.0014	0.0362	0.0352	0.93

TABLE VII
TIME COMPLEXITY ANALYSIS OF DIFFERENT CLUSTER MODELS CALCULATED OVER FIVEFOLD CROSS VALIDATION

Models	Number of models	Training time (h)	Testing time (h)	Total time (h)
36-km cluster	3190	154.01	0.76	154.68
72-km cluster	888	137.20	1.95	137.68
144-km cluster	242	116.00	4.6	116.31

order to compare spatially. Fig. 8(a) and (b) shows a better spatial correlation between DL-CYGNSS and SMAP predictions, while the MSU-GRI product in Fig. 8(c) shows under estimates specifically for high SM regions. It can be seen that the proposed DL-based approach can generate SM that is spatially more close to the SMAP SM. Besides the spatial assessment, we also present the temporal variations of compared products. Fig. 8(d) and (e) shows the temporal comparison among the three SM products from July 2019 to December 2020. For this analysis, we consider two different areas of size 225 × 225 km as shown in Fig. 8(a)–(c). Fig. 8(d) shows the temporal variation of the average SM in location-(1), and it shows our DL approach closely follows the SMAP SM. We see almost the same pattern for location-(2) in Fig. 8(e). The location-(1) is mostly arid, so it is easy for the model to learn the pattern easily. In both cases, the MSU-GRI product is sometimes unable to follow the label SM value.

The primary purpose of this study is to use CYGNSS data for high spatio-temporal resolution over the different heterogeneous areas utilizing the power of a DL model that has better function approximation as well as the capability to find complex nonlinearity. Proper utilization of a DL algorithm for SM retrievals needs a well-organized dataset that is reliable and labeled accurately before the training process. Our analysis shows that entire DDMs contribute more than only using generated features from DDMs. The main challenge to using the DL over the DDMs is computational power and processing time. In this study, we have used different clustering models for estimation SM. Table VII describes the model corresponding to their training and testing time for a fivefold cross-validation strategy. The 36-km clusters have more models for training and testing and take more time (154.68 h) than the other two cluster models. The 72-km cluster takes 137.68 h, where training time is 137.20 and 1.95 h for testing as the 144-km cluster has less number of the model, so it takes less training (116.00 h) and testing (4.6 h) time than the other two models.

It is imperative to note that the actual spatial resolution of CYGNSS data is subject to interpretation. The surface conditions dictate what the actual spatial resolutions are as the instrument DDM can spread nonuniformly for given observation. In our results, we refer posted resolution (9 km), which does not necessarily represent a CYNGSS native resolution as it can vary from a few kilometers to tens of kilometers depending on the degree of coherence. In fact, ancillary information around specular point (within 3 km) is used to train the model. The full DDM provides additional features that potentially learn about degree of the spread of the signal. The use of full DDM seems to help to replicate SMAP better than using designed features from DDM (e.g., peak power) as DDM can cover SMAP's native resolution (33 km). However, this assertion requires further investigation.

This study demonstrates the potential improvement over the existing ML-based approaches by utilizing DDMs for SM retrievals using the DL algorithms for over various land surface conditions at high spatio-temporal resolutions. The performance across different land covers yields promising results with higher accuracy. In addition, year-based cross validation is used to assess the generalized methodology over time. The DL-based technique captures temporal variation with varying biases, and the results show the importance of prior information on the DL-based model's prediction capabilities. The proposed method is limited to a reference label with its own uncertainty, and it can be further analyzed by using in situ data as labeled SM.

VI. CONCLUSION

In this article, a DL-based framework has been demonstrated for estimating SM using the CYGNSS DDMs along with ancillary geophysical data in the CONUS region. One of the most widely used DL methods (e.g., CNN) is utilized. Four different models are trained and validated using SMAP SM values. Models are validated using the fivefold cross validation as well as year-based cross validation using 4-years data between 2017 and 2020. We have developed and tested models learned from different size regions. Results show that errors are reduced, and correlations are improved with the increased number of DL models. Among different clustering approaches, the best ubRMSD and correlation coefficient is achieved using the 36-km clusters with a mean ubRMSD of 0.0362 $\mathrm{m}^3/\mathrm{m}^3$ and *R*-value of 0.93 using the fivefold cross-validation technique. More importantly, sufficient accuracy can be obtained via year-based cross validation as well. The year-wise cross validation shows a performance of mean ubRMSD of 0.033 $\mathrm{m}^3/\mathrm{m}^3$ and *R*-value of 0.94 for 36-km clusters, when models are learned using data from 2017 to 2019 and tested on all data from 2020. We have also compared our predictions result with ISMN SM stations, and it provides good agreement with SM station observations and SMAP data, which suggests that the proposed DL model can be generalized in space and time with promising confidence.

Meanwhile, the proposed DL-CYGNSS approach is analyzed regarding the temporal variation and different land cover conditions. Particularly, this model predicted SM with higher accuracy for grassland, croplands, and savanna compared to other land covers. The proposed DL-CYGNSS model can be extended to global scale with similar train/test scenarios as a future study.

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, 2008, Art. no. W00D06.

[2] DA 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] Mauro 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] J. Peng and A. Loew, "Recent advances in soil moisture estimation from remote sensing," *Water*, vol. 9, no. 7, 2017, Art. no. 530.

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

[8] 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.

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

[10] V. Zavorotny, S. Gleason, E. Cardellach, and A. Camps, "Tutorial on remote sensing using GNSS bistatic radar of opportunity," *IEEE Geosci. Remote Sens. Mag.*, vol. 2, no. 4, pp. 8–45, Dec. 2014.

[11] E. Valencia, V. U. Zavorotny, D. M. Akos, and A. Camps, "Using DDM asymmetry metrics for wind direction retrieval from GPS ocean-scattered signals in airborne experiments," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 3924–3936, Jul. 2014.

[12] A. Komjathy, M. Armatys, D. Masters, P. Axelrad, V. Zavorotny, and S. Katzberg, "Retrieval of ocean surface wind speed and wind direction using reflected GPS signals," *J. Atmospheric Ocean. Technol.*, vol. 21, no. 3, pp. 515–526, 2004.

[13] D. Guan et al., "Wind direction signatures in GNSS-R observables from space," *Remote Sens.*, vol. 10, no. 2, 2018, Art. no. 198.

[14] E. Santi et al., "Forest biomass estimate on local and global scales through GNSS reflectometry techniques," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 8680–8683.

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

[16] W. Li, E. Cardellach, F. Fabra, S. Ribó, and A. Rius, "Assessment of spaceborne GNSS-R ocean altimetry performance using CYGNSS mission raw data," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 238–250, Jan. 2019.

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

[18] M. P. Clarizia, N. Pierdicca, F. Costantini, and N. Floury, "Analysis of CYGNSS data for soil moisture retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2227–2235, Jul. 2019.

[19] 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.

[20] 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.

[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] 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.

[23] 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.

[24] M. Al-Khaldi et al., "Time-series retrieval of soil moisture using CYGNSS," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4322–4331, Jul. 2019.

[25] 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.

[26] 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.

[27] 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. IGARSS*, 2021, pp. 6303–6306.

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

[29] T. M. Roberts, I. Colwell, R. Shah, S. Lowe, and C. Chew, "GNSS-R soil moisture retrieval with a deep learning approach," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 147–150.

[30] S. Gleason, C. S. Ruf, A. J. O'Brien, and D. S. McKague, "The CYGNSS level 1 calibration algorithm and error analysis based on on-orbit measurements," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 1, pp. 37–49, Jan. 2019.

[31] 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.

[32] C P. Ruf Chang et al., *CYGNSS Handbook Cyclone Global Navigation Satellite System: Deriving Surface Wind Speeds in Tropical Cyclones*. Ann Arbor, MI, USA: National Aeronautics and Space Administration, 2016, p. 154.

[33] M. M. Al-Khaldi, J. T. Johnson, S. Gleason, E. Loria, A. J. O'Brien, and Y. Yi J O', "An algorithm for detecting coherence in cyclone global navigation satellite system mission level-1 delay-Doppler maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4454–4463, May 2020.

[34] S. Gleason, "Level 1B DDM calibration algorithm theoretical basis document," CYGNSS Project Document 148-0137, Rev 3, Oct. 20, 2020. [Online]. Available: https://cygnss.engin.umich.edu/data-products/

[35] S. Gleason and C. Ruf, "Overview of the delay Doppler mapping instrument (DDMI) for the cyclone global navigation satellite systems mission (CYGNSS)," in *Proc. IEEE MTT-S Int. Microw. Symp.*, 2015, pp. 1–4.

[36] M. P. Clarizia, N. Pierdicca, F. Costantini, and N. Floury, "Analysis of CYGNSS data for soil moisture retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2227–2235, Jul. 2019.

[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. Dorigo, S. Zwieback, A. 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, pp. 1–16, 2013.

[40] S. Chan, R. Bindlish, R. Hunt, T. Jackson, and J. Kimball, "Vegetation water content," Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA, Tech. Rep. JPL D-53061, 2013.

[41] T. Hengl et al., "SoilGrids250m: Global gridded soil information based on machine learning," *PLoS One*, vol. 12, no. 2, 2017, Art. no. e0169748.

[42] 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, pp. 418–422, 2016.

[43] 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.

[44] P. E. O'Neill, E. G. Njoku, T. J. Jackson, S. Chan, and R. Bindlish, "SMAP algorithm theoretical basis document: Level 2 & 3 soil moisture (passive) data products," Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA, Tech. Rep. JPL D–66480, 2015.

[45] 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.

[46] A. Camps, M. Vall llossera, H. Park, G. Portal, and L. R. Spatafora, "Sensitivity of TDS-1 GNSS-R reflectivity to soil moisture: Global and regional differences and impact of different spatial scales," *Remote Sens.*, vol. 10, no. 11, 2018, Art. no. 1856.

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

He is currently a Graduate Teaching Assistant with 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.

After he received the Ph.D. degree, he was an Assistant Professor with Marmara University until 2015. Between 2015 and 2017, he was a Postdoctoral Researcher with the Department of Mechanical and Materials Engineering, Florida International University. Between 2017 and 2019, he was with the Department of Electric and Computer Engineering, University of Alabama, as a Postdoctoral Researcher. He is currently an Assistant Research Professor with Geosystems Research Institute, Mississippi State University, Starkville, MS, USA. His research interests include remote sensing, biomedical signal processing, wearable sensors, pattern recognition, fiber optic sensors, and structural health monitoring.

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

From 2003 to 2009, he researched compressive sensing-based computational imaging problems at Georgia Tech. Between 2009 and 2017, he held faculty positions with TOBB University and University of Alabama, 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, Starkville, MS, USA, where he is the Co-Director of Information Processing and Sensing (IMPRESS) Laboratory.

Dr. Gurbuz was the recipient of The Best Paper Award for *Signal Processing Journal* in 2013, the Turkish Academy of Sciences Best Young Scholar Award in Electrical Engineering in 2014, and the 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*.

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, Starkville, MS, USA, where he is also the Co-Director of Information Processing and Sensing (IMPRESS) Laboratory. His current research focuses on 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 is a Senior Member of IEEE Geoscience and Remote Sensing Society (GRSS) and a Member of U.S. National Committee for the International Union of Radio Science (USNC-URSI). He has been an 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. He was the recipient of the Leopold B. Felsen Award for excellence in electromagnetic in 2013, the URSI Young Scientist Award in 2014, and the NSF CAREER Award in 2022. He was an Early Career Representative for the International URSI Commission F (Wave Propagation and Remote Sensing) from 2014 to 2021.