A DEEP LEARNING-BASED SOIL MOISTURE ESTIMATION IN CONUS REGION USING CYGNSS DELAY DOPPLER MAPS

M M Nabi, Volkan Senyurek, Ali Cafer Gurbuz, Mehmet Kurum

Mississippi State University, Mississippi State, MS, USA

ABSTRACT

NASA Cyclone Global Navigation Satellite System (CYGNSS) mission has gained 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 (DDM) that contain valuable earth surface information from GNSS reflection measurements. Existing approaches use predefined features from DDMs to estimate SM. This paper presents a deep-learning framework to learn optimal features from DDMs for estimating SM. The proposed approach is applied over the Continental United States (CONUS) by leveraging CYGNSS DDM observations with ancillary remotely sensed geophysical data. The model is trained and evaluated using the Soil Moisture Active Passive (SMAP) mission's enhanced SM products at a 9km × 9km resolution with vegetation water content less than $5kg/m^2$. The mean unbiased root-mean-square difference (ubRMSD) between CYGNSS and SMAP SM retrievals from 2017 to 2020 is 0.0362 m^3/m^3 with a correlation coefficient of 0.9309 over 5-fold cross-validation.

Index Terms— CYGNSS, Soil Moisture, Deep-learning, CNN, GNSS-R, SMAP

1. INTRODUCTION

Soil Moisture (SM) is essential for crop harvesting, rain forecasting, hydrology, meteorology, and different earth science applications. High-resolution and accurate SM estimation is required for many applications such as flood forecasting and yield estimation [1]. There are several dedicated satellite missions used for SM retrieval with different spatial and temporal resolutions. The National Aeronautics and Space Administration's (NASA) Soil Moisture Active Passive (SMAP) [2] and the European Space Agency's (ESA) Soil Moisture and Ocean Salinity (SMOS) [3] are two conventional satellite missions which are operated with L-band passive radiometers and provide SM measurement approximately 36-km spatial resolution and 1-3 days temporal coverage. The spatio-temporal resolution of current SM products can be potentially improved by using constellations of small satellites enabled by "receive-only" GNSS-R techniques.

This work was supported by National Science Foundation under Grant No. 2047771 and the USDA-ARS Award NACA 58-6064-9-007.

NASA launched constellation of 8 micro-satellites called the CYGNSS mission in December 2016 to improve hurricane forecasting. It provides observations from 38^o North to 38^o South latitudes over both land and ocean. With 4 channels per satellite, CYGNSS observations have an improved spatial and temporal resolution under the assumption of coherent reflection. Many recent studies have been successfully able to retrieve surface SM using CYGNSS observations [4, 5]. The majority of the previous studies used effective reflectivity obtained from peak reflected power in a Delay-Doppler Maps (DDM) [6, 7]. 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 its peak power value. While ancillary information from other sources can provide additional information, the goal of this paper is to develop approaches that learn the optimal features directly from the entire DDM images for the SM estimation problem and by this way to increase SM estimation accuracy. CYGNSS provides three types of DDM images; Analog Power, effective scattering area, and bistatic radar cross-section (BRCS). Our proposed approach utilizes these three DDM images as inputs together with ancillary data within a deep learning (DL) architecture to map the SM value. A recent study showed a DDM could be used for SM estimation using the DL method [8], however the approach used only one type of DDM (Power Analog) and no quantitative performance metric is presented for their SM retrieval model. In this study, we have utilized three DDMs as input together with physical ancillary data relevant to SM estimation in a DL framework with convolutional and fully connected neural network layers for enhanced SM estimations. The proposed DL architecture is evaluated under different train/test scenarios using the SMAP mission's enhanced SM products at a 9km × 9km resolution over Continental United States (CONUS).

The rest of the paper is organized as follows: Section 2 summarizes the utilized datasets. Details of the proposed DL architecture and methodologies are provided in Section 3. Results are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. DATASET

2.1. Cyclone Global Navigation Satellite System

In this study, the CYGNSS Level-1 (L1) version 2.1 product is used, which is available at the NASA Physical Oceanography Distributed Active Archive Center. DDM is one of the key measurements in L1 dataset that represent the received surface power over a range of time delays and Doppler frequencies (bin-by-bin) for each observed specular reflection point. The DDM gives an 11×17 array. In addition to DDMs, geometric and instrumental variables are also incorporated to provide complete acquisition information for each specular point. For this study, we have considered the CONUS region from March 2017 to November 2020, where a total of more than 18 Million DDM samples exist.

2.2. SMAP Radiometer Soil Moisture Data

The SMAP Enhanced L3 Radiometer Global Daily 9-km EASE-Grid SM product is used as reference SM data to train and validate developed SM retrieval model. SMAP uses the L-band microwave radiometer to collect brightness temperature data and produces SM estimates. SMAP datasets contain SM product and the associated coordinates for the descending (A.M.) and ascending (P.M.) overpasses which are then combined to obtain daily results.

2.3. Ancillary Data

Different geophysical parameters play essential roles in accurately predicting SM. In order to characterize vegetation conditions, the 16-day composite Normalized Difference Vegetation Index (NDVI) is utilized from Moderate Resolution Imaging Spectroradiometer (MODIS) data. Vegetation Water Content (VWC) is calculated using the NDVI and Land Cover Type (MCD12Q1) products using the same lookup table method as the SMAP VWC product [9]. 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. Soil clay and silt ratios are obtained from the Global Gridded Soil Information (SoilGrids) [10]. A 30-m Global Surface Water Dataset from the Joint Research Centre (GSW-JRC) [11] is used to identify the presence of a surface inland water body. All ancillary data are spatially aggregated from their native resolutions to 3 km.

2.4. Quality Control Mechanisms

Several quality control criteria need to be applied to CYGNSS observations and ancillary data to maintain the data quality. CYGNSS metadata control flags such as S-band powered up, substantial spacecraft attitude error, black-body DDM, DDM test pattern, poor confidence GPS EIRP estimate are utilized [12]. Observations with an incidence angle higher than $\pm 65^0$ are eliminated [13]. If the surface water is sufficiently large within grid, SM estimate is highly effected. Thus, a CYGNSS observation is removed if more than 2% of the 3 km grid centered on a specular point is covered with permanent or seasonal water. Additionally, CYGNSS readings that fall over forested areas with VWC>5 kg/m^2 (dense vegetation canopy) are also eliminated. CYGNSS observations before December 2017 that are above 600m from the surface are masked out due to the altitude limitation of CYGNSS L1 data for the specified time period.

3. METHODS

In this study, we formulated SM retrieval as a complex function of CYGNSS DDMs and ancillary inputs. We consider Convolutional neural network (CNN) as our DL model where the primary inputs are coming from three types of CYGNSS DDM images. CNN learns a number of features directly from the DDM images which are later combined with ancillary features that are described in Section 2.3. Dense layers map the combined features to the final SM value. We consider a supervised learning approach where the model is trained over a labelled dataset of SMAP SM values (Sect. 2.2) minimizing the mean square error loss.

3.1. CNN Design

The developed DL model consists of three major parts; convolutional network, concatenation , and densely connected layers. The convolutional layers are used to extract features from the input DDM images. At the end of the layer, extracted features are flattened to concatenate with additional ancillary feature inputs. The concatenation layer is used for combining the features from ancillary features with the CNN's flattened output. As mentioned earlier, three DDMs are the primary input of the CNN network. Each type of image is given in a different channel (a total of 3 channels) after normalization. In total 3 convolutional layers are used followed by a maxpool layer at the end. Each convolutional layer consists of 3×3 kernel with 0 paddings and stride 1. The filter sizes of convolutional layers are 32, 64, and 128 respectively. A max-pooling layer (5×11 kernel size) is applied after the convolutional layers. A flattening layer is used to flatten the extracted features into a vector format. After each convolutional layer, a batch normalization layer followed by a Rectified Linear Unit (ReLU) activation is used. We extract 128 features from the DDM images. DDM based features combined with 9 ancillary features are input to 2 fully connected layer with 50 nodes where the final regression layer maps to a single SM value. Fig. 1 shows the overall structure of the developed DL model.

3.2. Training the Model

We trained the proposed DL model for three different scenarios. First, a single model is trained for the whole CONUS region. Second, clusters of 72km and 144km are formed, and a model is learned for each cluster separately. For clusters models are learned for smaller regions, however using less

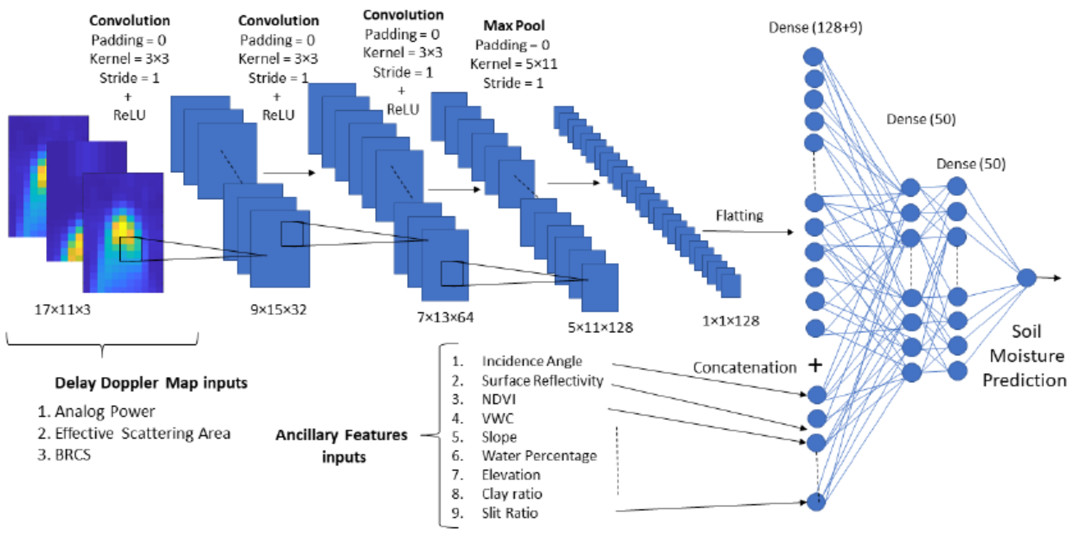

Fig. 1. Overall structure of CNN-based soil moisture estimation using DDMs and other ancillary input.

Table 1. Performance metrics for different models using 5-fold cross-validation

	Number of models	no. of samples	RMSD $(m^3m\ ^3)$	mean ubRMSD $(m^3m\ ^3)$	median ubRMSD $(m^3m\ ^3)$	R value $(m^3m\ ^3)$
72KM-cluster	5725	3247	0.0403	0.0362	0.0350	0.9309
144KM-cluster	1667	11148	0.0480	0.0417	0.0410	0.9009
One-cluster	1	1.86e+7	0.0580	0.0482	0.0470	0.8512

Table 2. Performance comparison of different SM product over CONUS regions (72km cluster model)

Models	RMSD $(m^3m\ ^3)$	ubRMSD $(m^3m\ ^3)$	R-value
Our DL SM	0.0407	0.0366	0.93
MSU-GRI product	0.0518	0.0434	0.91

number of training data. Root mean square propagation (RMSProp) is used as an optimizer. The network is trained using mini-batch gradient descent, where the training data is randomly split into subsets and the mini-batch is selected depending upon the types of model we used. The total number of epochs used is 250. The required computations are carried out using the DL toolbox of MATLAB R2021b software over a machine with Intel(R) Xeon(R) CPU E5-2643 and 128 GB memory.

4. RESULTS

In this section, the SM retrieval results from DL based approaches are presented. The overall performance of the DL model for SM retrieval is evaluated through 5-fold cross validation technique. Table 1 shows the overall SM prediction performance derived via the DL model for different approaches. The developed model reaches an overall RMSD of 0.0403 m^3/m^3 and a R value of 0.9309 using the 72km cluster models. For this model, total number of 72km cluster is 5725 and the mean ubRMSD 0.0362 m^3/m^3 and median 0.0350 m^3/m^3. The total number of model for 144km cluster is 1667 and it's mean ubRMSD 0.0417 m^3/m^3 and median 0.0410 m^3/m^3. We have also trained a single model for CONUS region. This single model reaches an overall RMSD of 0.0580 m^3/m^3 and a R value of 0.8512. Fig. 2 shows ubRMSD and correlation coefficient for each 9 km grid using the model for 72 km clusters. Fig. 2a shows that SM predictions are more accurate for the low vegetated regions of the CONUS. However, the error is higher for the relatively more vegetated eastern part of the CONUS. Fig. 3 depicts the scatter plot between label SM and predicted SM for the 72km cluster model with an R value of 0.93. We compare our approach with ML-based [14] SM retrieval. Table 2 shows the comparison of the 72-km cluster between the two approaches, and it is clearly observed that our DL approach outperformed the exiting ML-based model.

5. CONCLUSION

In this paper, DL based framework has been demonstrated for estimating SM using the CYGNSS DDMs along with ancillary geophysical data. Different models are trained using SMAP SM values as labels. Models are validated using the 5-fold cross-validation and the best ubRMSD and cor-

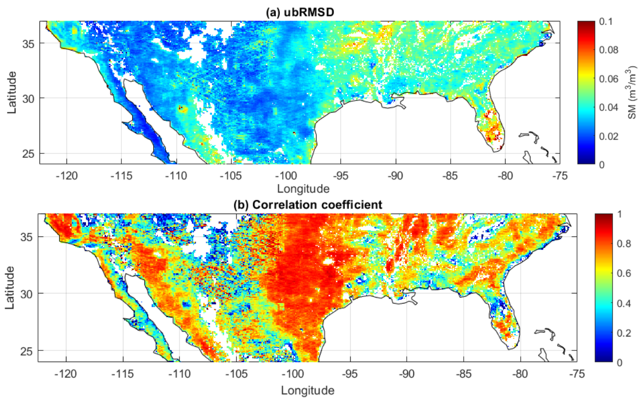

Fig. 2. (a) ubRMSD and (b) correlation coefficient map for 72 KM model test result (4-years averaged)

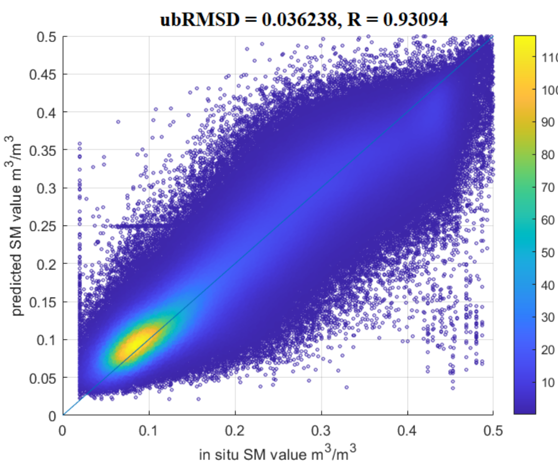

Fig. 3. The scatter plot of the predicted SM versus SMAP SM

relation coefficient is achieved for the 72km clusters with a mean ubRMSD of 0.0362 m^3/m^3 and R of 0.93. As a future study, the proposed model will be applied globally with varying train/test scenarios.

6. REFERENCES

[1] 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.

[2] D. Entekhabi, E. G. Njoku, P. E. ONeill, , et al., "The soil moisture active passive SMAP mission," *Proc. IEEE*, vol. 98, no. 5, pp. 704–716, 2010.

[3] Y. Kerr, A. Al-Yaari, N. Rodriguez-Fernandez, M. Parrens, B. Molero, et al., "Overview of SMOS performance in terms of global soil moisture monitoring after six years in operation," *Remote Sensing of Environment*, vol. 180, pp. 40–63, 2016.

[4] 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, pp. 2272, 2019.

[5] Qingyun Yan, Weimin Huang, Shuanggen Jin, and Yan Jia, "Pan-tropical soil moisture mapping based on a three-layer model from cygnss gnss-r data," *Remote Sensing of Environment*, vol. 247, pp. 111944, 2020.

[6] 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 Sensing*, vol. 12, no. 7, pp. 1168, 2020.

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

[8] T Maximillian Roberts, Ian Colwell, Rashmi Shah, Stephen Lowe, and Clara Chew, "Gnss-r soil moisture retrieval with a deep learning approach," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 147–150.

[9] Steven Chan, Rajat Bindlish, Raymond Hunt, Tom Jackson, and John Kimball, "Vegetation water content," *Jet Propulsion Laboratory, California Institute of Technology: Pasadena, CA, USA*, 2013.

[10] T. Hengl, J.M. de Jesus, G. B. Heuvelink, M. R. Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, and M. A. Guevara, "Soilgrids250m: Global gridded soil information based on machine learning," *PLoS one*, vol. 12, no. 2, 2017.

[11] Jean-François Pekel, Andrew Cottam, Noel Gorelick, and Alan S Belward, "High-resolution mapping of global surface water and its long-term changes," *Nature*, vol. 540, no. 7633, pp. 418, 2016.

[12] 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, pp. 1053, 2019.

[13] M. Al-Khaldi, J. Johnson, A. O'Brien, A. Balenzano, et al., "Time-series retrieval of soil moisture using CYGNSS," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4322–4331, 2019.

[14] Fangni Lei, Volkan Senyurek, Mehmet Kurum, Ali Cafer Gurbuz, Dylan Boyd, Robert Moorhead, Wade T Crow, and Orhan Eroglu, "Quasi-global machine learning-based soil moisture estimates at high spatio-temporal scales using cygnss and smap observations," *Remote Sensing of Environment*, vol. 276, pp. 113041, 2022.