Investigating CYGNSS-derived surface reflectivity for estimating soil moisture in Texas

Undergraduate Thesis

Submitted in partial fulfillment of the requirements of
BITS F421T Thesis

By

Shray Mathur
ID No. 2017A7TS1180P

Under the supervision of:

On-campus supervisor: Prof. Mukesh K. Rohil
&
Off-campus cosupervisor: Dr. Michael H. Young

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI, PILANI CAMPUS
June 2021

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI, PILANI CAMPUS

Abstract

Bachelor of Engineering (Hons.)

Investigating CYGNSS-derived surface reflectivity for estimating soil moisture in Texas

by Shray MATHUR

This dissertation presents a Machine Learning based soil moisture retrieval method for NASA's Cyclone Global Navigation Satellite System (CYGNSS). The CYGNSS observations are compared to the Soil Moisture Active Passive (SMAP) satellite and in-situ Texas Soil Observation Network (TxSON) soil moisture (SM) measurements for the months of January, April and July 2019. An initial grid-wise sensitivity analysis of CYGNSS reflectivity ($P_{r,eff}$) to Soil Moisture (SM) is conducted at a 9x9 km^2 grid resolution over the 36x36 km^2 TxSON region to assess the spatio-temporal relationships between $P_{r,eff}$ and SM. Variability among grid cells and seasonal shifts in correlations motivated inclusion of land physical parameters and CYGNSS observation geometry in the analysis. Specifically, we include the Specular Point (SP) incidence angle (θ), Elevation, Clay Fraction, Normalized Difference Vegetation Index (NDVI), Depth to Restrictive Layer (DepRes), and surface roughness. The individual effects of these variables on $P_{r,eff}$ are assessed through a correlation and regression analysis. Finally, an Artificial Neural Network (ANN) model is trained for different combinations of input features to attain SM estimates at 9x9 km^2 and 3x3 km^2 grid resolutions. The model structure is tuned to attain optimal results for different combinations and a 5-fold cross validation approach is employed to train the models. SM predictions with a root mean squared error of 0.0409 (0.0497) cm^3/cm^3 and Pearson correlation coefficient of 0.7024 (0.6794) are reported at 9x9 (3x3) km^2 grid resolution.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Till date, three spaceborne GNSS-R programs have been used to repurpose GNSS data for estimating soil moisture (SM). The programs include UK-DMC, TDS-1 and CYGNSS, out of which the UK-DMC was equipped with the first spaceborne GNSS-R receiver and was initially intended to study ocean surface windspeed and roughness. The potential of the land reflected signals was found in later research [11]. TDS-1 satellite was launched in 2014, with the SGR-ReSI (Space GNSS Receiver Remote Sensing Instrument). Data collected by the TDS-1 satellite was analyzed for sensitivity to SM [4, 8] and spatial and temporal variations in GNSS-R reflections were found to be driven by similar variations in SM. While the data collected by TDS-1 is similar to data collected by CYGNSS, the volume of data using TDS-1 is orders of magnitude less than CYGNSS, because the revisit time of TDS-1 is greater than 6 months, while that of CYGNSS about 1 day. For this reason, and because the CYGNSS data availability through NASA data portals was relatively simple, we chose CYGNSS platform for this study.

1.1 Previous Investigations

Several methodologies have been proposed to retrieve SM using CYGNSS [2, 9, 6, 7, 13, 1, 12, 16, 15, 10]. Majority of these works assume that reflections from land surface are dominated by the coherent reflections. In [6] the SNR data from CYGNSS is compared to SMAP SM for a 1-year time period from March 2017 to March 2018. A strong positive linear correlation between the temporal deviation of CYGNSS reflectivity and SMAP SM is demonstrated, and a linear regression model is established with an RMSE of 0.045 cm^3/cm^3. Their work indicates that CYGNSS can be used to develop global SM products at a high temporal scale (possibly every 6 hours) but poor results are achieved for regions with drought conditions or dense vegetation coverage. Kim and Lakshmi's work [13] also ignore the calculations of non-coherent scattering and assume land surface reflected signals are coherent. They develop a relative SNR index

(rSNR) by normalizing the SNR values from April 2017 to April 2018 using the incidence angle. A linear correlation between rSNR and SMAP is shown (r = 0.68) and they are coupled to obtain daily change in soil moisture. In this study, the angle factor is considered through simple normalization process; however, the scattering characteristics of SM under different observation geometries have not been explored and fully used. Pan-tropical SM values are retrieved using CYGNSS in [9, 17]. Both studies assume that the reflected signal of CYGNSS is mainly coherent and use additional ancillary data sources along with CYGNSS-derived surface reflectivity for estimating SM. [9] proposed a trilinear regression-based reflectivity-vegetation-roughness (R-V-R) algorithm wherein the daily SM estimates were derived based on the CYGNSS reflectivity along with SMAP vegetation opacity and roughness coefficient. In [16], vegetation opacity was adopted as an auxiliary variable. For surface roughness they calculate statistical moments from the 2D Delay Doppler Map (DDM). The surface-reflected GNSS signals are recorded by the CYGNSS receivers in the form a delay-Doppler map (DDM). The statistical moments calculated include the mean, variance, skew and kurtosis. This step was employed in an attempt to reduce the number of ancillary data sources used. The approach was supported by the fact that the shape of the 2D DDM can be used to interpret surface roughness. [9, 17] both conduct their studies at a 36x36 km2 grid resolution and both report an RMSE of 0.07 cm^3/cm^3. These studies all assume a linear relationship between CYGNSS-derived surface reflectivity (and its combinations with ancillary data) and SM. However, the SM retrieval process is known to involve high complexity and nonlinearity. To address this, [10, 16] investigated the potential use of non-parametric, non-linear Machine Learning Algorithms for this task. [10] demonstrated the potential of an artificial neural network (ANN) to learn nonlinear relations of SM and other land physical parameters to CYGNSS observables. They use a 10-fold validation method to train the model over a limited set of data collected from reference sites from the International Soil Monitoring Network (ISMN) sites between 2017 and 2018. [15] look to extend ML-based studies to the global scale by including more reference stations and using an independent validation strategy. Their model used in-situ SM data from 170 ISMN sites over a nearly-3 year period and achieved a performance better than 0.05 cm^3/cm^3 mean unbiased root-mean-square difference (ubRMSD) at a 9 km grid resolution.

The broad goal of our work here is to retrieve accurate soil moisture estimates over Texas at fine grid resolutions (9km and 3km). To achieve this goal we divide the work into the following objectives: 1) Download and harmonize the CYGNSS, SMAP and in-situ SM data at the required spatial and temporal scales. 2) Carry out a set of statistical steps to study the relationships between these data sources and also study the effects of a set of ancillary variables on CYGNSS-derived surface reflectivty. 3) Finally build an Artifical Neural Network model to estimate SM over the TxSON region at the required spatial resolutions.

1.2 Contributions

Extending the research described above, the following contributions are made in this study:

- We investigate whether a linear relationship can sufficiently describe the relationship between CYGNSS reflectivity and SM, and derive insights into the spatio-temporal complexities of these relationships.
- Because reflectivity is known to be effected by CYGNSS observation geometries and land parameters, in addition to SM, we look to quantify these relationships between $P_{r,eff}$ and the selected ancillary data sources through a correlation and regression analysis.
- Using a Grid-Search Cross Validation Approach, we investigate individual contributions of several CYGNSS-derived variables and land surface parameters in the SM estimate model.
- An optimal set of input features for SM estimates is established, which achieves satisfactory results at both 9km and 3km grid resolutions.

Chapter 2

Materials and Methods

2.1 Study Region

We study a 36x36 km^2 region, which is home to the Texas Soil Observation Network (TxSON). The site is located in the central Texas Hill Country near Fredericksburg, Texas (Figure 2.1).

TxSON is located within the Edwards Plateau, an uplifted area formed from marine deposits of limestone, shales, sandstones, and dolomites of Cretaceous age. This region is representative of the terrain of the semiarid rangelands of Texas Hill Country. Vegetation over this region includes oak trees (red, live, and post), woody plants (ashe juniper and honey mesquite), and a mixture of short and mid-height grasses (grama, switchgrass, bluestem, curly mesquite). The

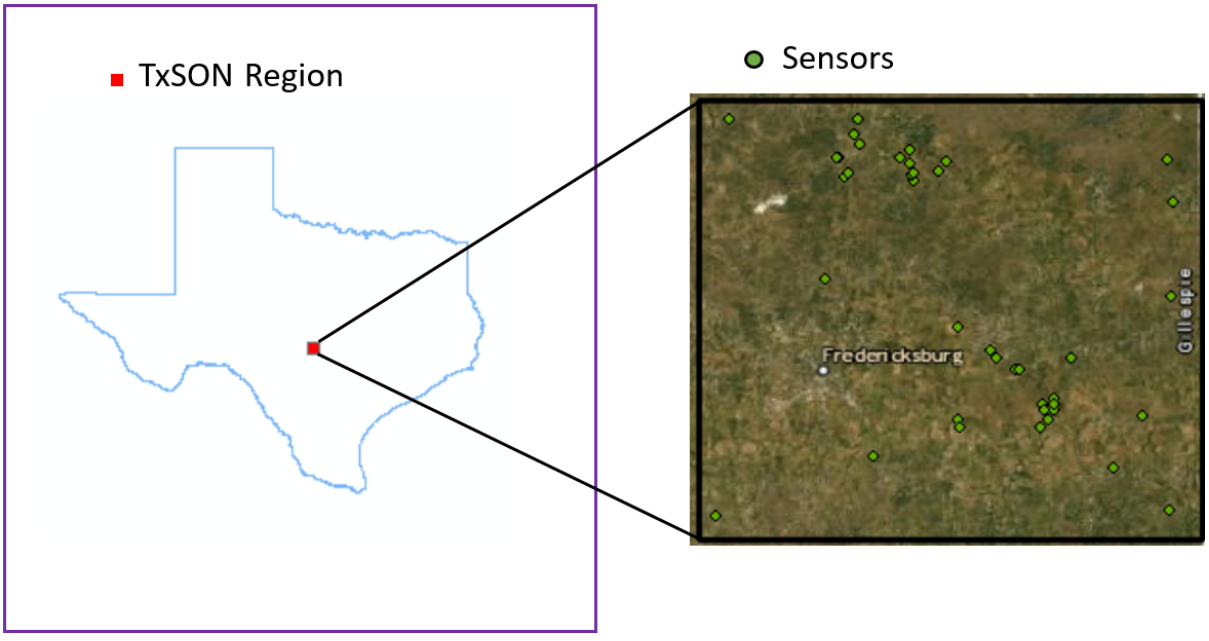

FIGURE 2.1: Site Map

soils are generally not appropriate for small grain or row crop production due to high erosion rates, shallow depths, and low water retention capacity, but they are well suited for grazing and viticulture. The 30-yr mean annual precipitation is 807 mm and air temperature is 18.4°C (PRISM Climate Group, Oregon State University, http://prism.oregonstate.edu, accessed 14 Dec. 2018).

2.2 Datasets

2.2.1 CYGNSS

The data used for this analysis are the Level 1 (L1) CYGNSS data, version 3.0, available at the Physical Oceanography Distributed Active Archive Center (PODAAC; https://podaac.jpl.nasa.gov). For each day, PODAAC processes a set of 8 NetCDF files—one for each CYGNSS satellite. Since the CYGNSS satellites are capable of recording 4 simultaneous reflections per second, each file contains 4 Delay Doppler Maps (DDMs) of analog scatter power for each second it records data on a given day. In addition, useful metadata about geometry of the acquisitions (e.g., incidence angle, azimuth angle, etc.) for each specular point, as well as information about the transmitting GPS satellite and the CYGNSS observatory receiving the reflection. In this study, we use 3 months of CYGNSS acquisitions: January, April and July 2019, to develop and test the retrieval algorithms.

Typically the CYGNSS data points are upscaled to an Equal-Area Scalable Earth 2 (EASE2) grid resolution (36km x 36km), which is coarser than its theoretical spatial footprint over land (3.5 x 0.5 km). This can effectively degrade the CYGNSS data and does not use its full potential. In addition, several CYGNSS observations are likely to be taken on each EASE2 grid, though not enough to completely sample the grid due to its smaller spatial footprint. Pixels with heterogeneous physical land parameters will result in variations in $P_{r,eff}$, which will not be associated with similar variations in SM values. In this study, we look to investigate the CYGNSS data at finer scale grid resolutions of 9km and 3km.

To remove outlier data points we use standard quality flags available in the CYGNSS metadata. Specifically, we use flags 2 (S-band transmitter powered up), 4 (spacecraft attitude error), 5 (black body DDM), 8 (DDM is a test pattern), 16 (direct signal in DDM), and 17 (low confidence in the GPS EIRP estimate). Additional quality control is applied by removing any observations with a signal-to-noise (SNR) value < 2.0 dB and any data points with incidence angle > 65 degrees.

The total scattered power of the forward scattered L band signals is a sum of both coherent and incoherent returns. However, it has been reported that reflections coming from the land surface

are dominated by the coherent component [14]. Similar to previous studies [6, 9, 10], we also assume that much of the reflections recorded by CYGNSS originate from coherent reflections and the non-coherent component is ignored. Under this assumption assumption the total scatter power is given by:

$$P_{coh}^{T} = \frac{P^t G^t G^r}{(r_{ts} + r_{sr})^2} (\frac{\lambda}{4\pi})^2 \Gamma \tag{2.1}$$

where: P_{coh}^{T} is the coherently received 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, λ is the GPS wavelength (0.19 m), and Γ is the reflectivity of the surface.

To solve 2.1, the coherently received power is replaced with a DDM observable. Previous studies have investigated the use of the DDM Signal-to-Noise-Ratio (SNR), DDM peak value or the Bistatic Radar Cross Section (BRCS) [6, 9, 17]. In this study we employ the SNR value which is directly available in the CYGNSS data for each specular point as the *ddm_snr* variable. The following relationship is used to convert the SNR to effective surface reflectivity ($P_{r,eff}$)

$$P_{r,eff} = SNR - 10\log(P^t G^t) - 10\log(P^r) + 20\log(r_{ts} + r_{sr}) + 20\log(4\pi) - 20log(\lambda) \tag{2.2}$$

To carry out this conversion, the CYGNSS variables required are : sp_rx_gain (G^r), rx_to_sp_range (r_{rs}), tx_to_sp_range (r_{ts}) and gps_eirp ($P^t G^t$). These variables are first converted to a dB scale.

2.2.2 TxSON

Here, hourly in-situ SM data are collected and used as ground truth from 40 monitoring sites, which constitute the Texas Soil Observation Network (TxSON) region. The spatial distribution of SM sensors within the 36 km TxSON region are shown in Figure 2.1. TxSON stations are nested within an Equal-Area Scalable Earth Grid at 3, 9 and 36 km and provides mean hourly SM values at 4 separate depths (5, 10, 20 and 50 cm). In addition, precipitation is also measured at all locations with six meteorological stations also providing air temperature and humidity, wind speed and direction, and solar radiation. These data are publicly available through the Texas Data Repository (https://dataverse.tdl.org/). In this study, we only use 5cm SM readings provided because the CYGNSS-derived reflectivity originates from the topsoil layer (0–5 cm).

2.2.3 SMAP

In this work the Half-orbit SMAP Enhanced L3 Radiometer Global Daily 9-km Equal Area Scalable Earth-Grid (EASE Grid) SM data are employed. They contain SM, quality flag, and other auxiliary information gridded over the EASE-Grid v2.0, with ascending and descending passes averaged together to form a single daily pass.

2.2.4 Downloading and Harmonizing Data

The CYGNSS, SMAP and TxSON data were retrieved for the months of January, April and July from 2019. These months were selected for this study with the aim of studying the effects of leaf-on and leaf-off conditions.

To conduct this analysis, the three data sources need to be at the same spatial and temporal scale, which is a daily 9km grid resolution. While SMAP data were available at the required resolution, the raw CYGNSS and TxSON data were upscaled and aggregated appropriately. In both cases, we employ the Voronoi upscaling method (as done in [3])to bring both these data sources to a 9km grid resolution. For the CYGNSS data, individual observations at sub-daily scale were ignored and daily observations were directly gridded to a daily 9km resolution. Because the TxSON data provides SM estimates every hour, the Voronoi method was first applied to obtain hourly gridded in-situ SM measurements. Then, within each grid, hourly estimates were aggregated to daily SM estimates for each 9km grid over the TxSON region.

2.3 Data Analysis

2.3.1 Relationship Between CYGNSS Reflectivity, TxSON SWC and SMAP SM

Studies analyzing the land returns of the CYGNSS constellation have investigated the sensitivity of the SNR value to SMAP measured SM and patterns of complementary change thereof. Strong positive correlations between these have been reported and have motivated the use of linear regression models for estimating SM [5, 6]. Grid resolution at which the correlation analysis is conducted has been a source of uncertainty in these studies

To address these uncertainties, we investigate correlations between $P_{r,eff}$ and SMAP SM at a finer resolution (9km x 9km) over the TxSON grid, a region with complex terrain and diverse topography. We go one step further than previous studies and include in-situ measured SM values from the TxSON sensors, thereby allowing a a three-way spatial sensitivity analysis between CYGNSS, TxSON and SMAP. This analysis is conducted separately for the months of

January, April and July from 2019 to account for the effect of leaf-on and leaf-off conditions and to study seasonal shifts and trends. We also consider the spatial variability of correlations in the absolute values of these data sources, and we investigate the temporal relationship by calculating the correlation between daily changes in values of $P_{r,eff}$, TxSON SM, and SMAP SM. Correlations from these studies are quantitively analysed and compared.

After upscaling the CYGNSS data and TxSON data to a 9km grid resolution at a daily scale, each grid was assigned one value for effective surface reflectivity and one value for 5cm SM using TxSON and SMAP, thereby yielding 30 data points each from each data source for each month. Pair-wise correlation coefficients were computed between the data points assigned to a grid, to get a monthly correlation value for each 9km cell.

While previous studies have focused on month-to-month variations, here we study the correlations between daily changes in the CYGNSS and TxSON data sources.

2.3.2 Variable Selection for Linear Models

The behavior of the forward scattered signals of GNSS-R data depends on many parameters other than SM. We look to quantitatively investigate the effect of land physical parameters that have been commonly used in previous CYGNSS-based SM retrieval methods [10, 17] on Pr,eff . These include surface roughness, elevation, clay fraction and NDVI. For surface roughness, we follow an approach similar to [17] by calculating statistical moments from the 2D DDM including mean (Γ_M) variance (Γ_V), skew (Γ_S) and kurtosis (Γ_K). In addition, we also study the effect of Depth to Restrictive (DepRes) Layer, a variable whose influence on Pr,eff has not been reported in the literature.

2.3.2.1 Correlation Analysis

We conduct a monthly correlation analysis between the aforementioned ancillary variables and CYGNSS-derived variables. We look to study the correlations significant at the 0.001 level and also look into seasonal changes in correlations, if any.

2.3.2.2 Regression Analysis

Elevation, Clay and DepRes are well correlated and therefore the regression analysis is conducted for each variable individually to avoid the problem of multicollinearity. The linear regression problem is set up as follows:

$$P_{r,eff} = a * variable + b \tag{2.3}$$

2.4 Machine Learning Model to Estimate Soil Moisture

Because SM can take up continuous values, estimates of SM with a predictive model is a regression problem. In the previous sections , we assess the linear relations between $P_{r,eff}$ and other ancillary variables; however, SM can be related to these variables through complex non-linear functions. One of the most common Machine Learning algorithms for nonlinear regression problems is the Artificial Neural Network (ANN). ANNs are a class of deep neural networks that are capable of approximating arbitrary and complex mappings between the input and output without much apriori knowledge of the underlying relationships in the data. This is achieved by building a hierarchical representation of the data through a layered architecture wherein the model weights are tuned through a sophisticated two-way training process. The non-parametric nature of ANNs also provides added benefits compared to commonly used linear regression models, in that they are flexible in the number of input features they can handle and make fewer assumptions about data distribution, allowing the potential use of several data sources together to achieve a common task.

2.4.1 Model Architecture

In this study, we employ a type of ANN known as the Multi-Layer Perceptron (MLP). Input features to the MLP, are the effective surface relectivity ($P_{r,eff}$), SP incidence angle from CYGNSS observation (Θ), statistical moments computed from the 2D DDM as a proxy for surface roughness (Γ_M, Γ_V, Γ_S, Γ_K), Clay Ratio, Elevation, Depth to restrictive layer (DepRes) and NDVI. We investigate results from different combinations of these variables and infer their individual contributions. To make a fair comparison between results from different combinations, we find the optimal MLP structure for each combination of variables through hyperparameter tuning. The MLP structure typically comprises three sections: Input Layer, Hidden Layers, and Output Layer. Each layer is composed of a series of neurons, which are connected to neurons of the previous layer via a set of weights. The essential part of the ANN model is the learning of the model weights through an iterative two-way propagation mechanism.

More specifically, each layer has a trainable set of parameters in the form of an array of weights, the size of which is a product of the number of neurons in the current layer and previous layer. In each iteration, the input features are fed into the model through the input layer. Through matrix multiplication of input features and the weight array of the first layer, linear activation values for each neuron in the first layer are computed. These values are passed through a non-linear activation function (ReLU, tanh, sigmoid, etc.) to introduce non-linearity in the system. This non-linear value acts as input for the next layer and the same process is repeated at each layer of the model until the output layer is reached (Forward Pass). The model then assesses performance by comparing the model predictions and ground truth values using a loss

function and computing an error value. This error information is then used to update the weights of the previous layers through the backward pass. This requires estimating the partial derivative of the loss function with respect to each parameter in the network. The parameters are then adjusted along the gradient descent direction by a predefined stride, which is named 'learning rate'. Several iterations of the forward and backward pass together are required to tune the model weights to achieve satisfactory performance of the required task. The trained network can then be used to make predictions (in our case SM estimates) on any new and unseen datapoint through a single forward pass

2.4.2 Training and Validation

For each CYGNSS observation, the corresponding ancillary variable values are extracted based on the SP latitude and longitude. On a given day, CYGNSS observations that lie within a grid are assumed to have the same ground truth (in-situ) SM value. As shown in [10], this is considered feasible as physical parameters (elevation, NDVI, DepRes) corresponding to each CYGNSS observation would differ from each other due to the spatial variation, which in turn could explain variations in the CYGNSS observations despite uniform SM values. The original dataset is binned into a train-and-test set using an 80:20 ratio. The test set is left untouched during the training process and is used to finally evaluate and compare performance of different combination of variables. We optimize the model performance for each combination through careful hyperparameter tuning using the Grid Search Cross Validation approach, which is described as follows. For each possible combination of input features, we first define a set of possible MLP model structures to be evaluated. We investigate the performance of 3 and 4 layered networks and try all possible combinations of either 10 or 100 neurons in any of the layers. For each MLP network, training and validation are performed together through a 5-fold Cross Validation approach in which the entire training dataset is split into 5 separate groups and the ANN model is trained over 5 iterations. In each iteration, the model is trained from scratch and 4 out the 5 groups of the data are used to train the model, while 1 group is kept as the validation set and is used to evaluate the model. After each iteration, the validation set is fed into the trained model and the Mean Squared Error for the model is calculated on this set. The group of data used for the validation set must be unique for each iteration. This allows to model to be evaluated on unseen data for each iteration and it allows each instance in the dataset the opportunity to be a part of the validation set. At the end of the 5 iterations, an average is taken of the 5 MSE scores to give one cross-validated score to this model type. For a given combination of input features, the model with the lowest cross-validated score is selected. This model is then run on the original test set and its performance measured in terms of Pearson' Correlation R and Root Mean Squared Error (RMSE). Apart from the model structure, we use

the same learning rate (0.001) activation function (ReLU) and optimization algorithm (Adam) for all input combinations.

Chapter 3

Results and Discussions

3.1 Comparison between CYGNSS, TxSON and SMAP

As expected, correlation between TxSON SM and SMAP SM is high ($r \geq 0.5$) (Figure 3.1 c, f, i) for nearly all 9km grids across the three months. However, spatial variation and monthly shifts in correlation trends in the CYGNSS-SMAP and CYGNSS-TxSON comparisons seem to exist.

We first analyse the correlations from a spatial perspective. Using the CYGNSS vs TxSON comparisons, for example, at a monthly scale, grid cells in the eastern region in January are strongly negatively correlated (grids 7 and 11 significant at 0.05 level). For April the central and eastern regions (grid cells 5, 6, 7, and 11 significant at the 0.001 level), and for July the south and south east regions (grids 1 and 3 significant at the 0.01 level) are positively correlated. At a sub-monthly (approximately daily) scale, correlation values between the 9km grids are not stable and can change drastically. The difference between the maximum and minimum correlations for the three months is 0.498, 0.600 and 0.638. We also noted substantial changes in absolute values of R for grid cells and changes in significance levels over the three months. For example, the R values for cell 7 (highlighted in red) are -0.39 (significant at the 0.05 level), 0.628 (significant at the 0.01 level) and 0.279 (not significant at the 0.1 level) for January, April, and July, respectively.

To further analyze the daily variability we plot scatter plots in Fig. 3.2 for grids with the highest correlation value in each heatmap (outlined in black in Fig. 3.1)

Results for January and July SM values show a decrease over the month, whereas for April there is a steady increase in soil moisture. These trends are expected given that April is one of the wettest months in central Texas, after which the region undergoes seasonal drying. As expected from Figures 3.1 a) and b, this variability in SM is not captured by Pr,eff for the month of January, but the trends are more visible for April and July. Another thing to note is that outlier

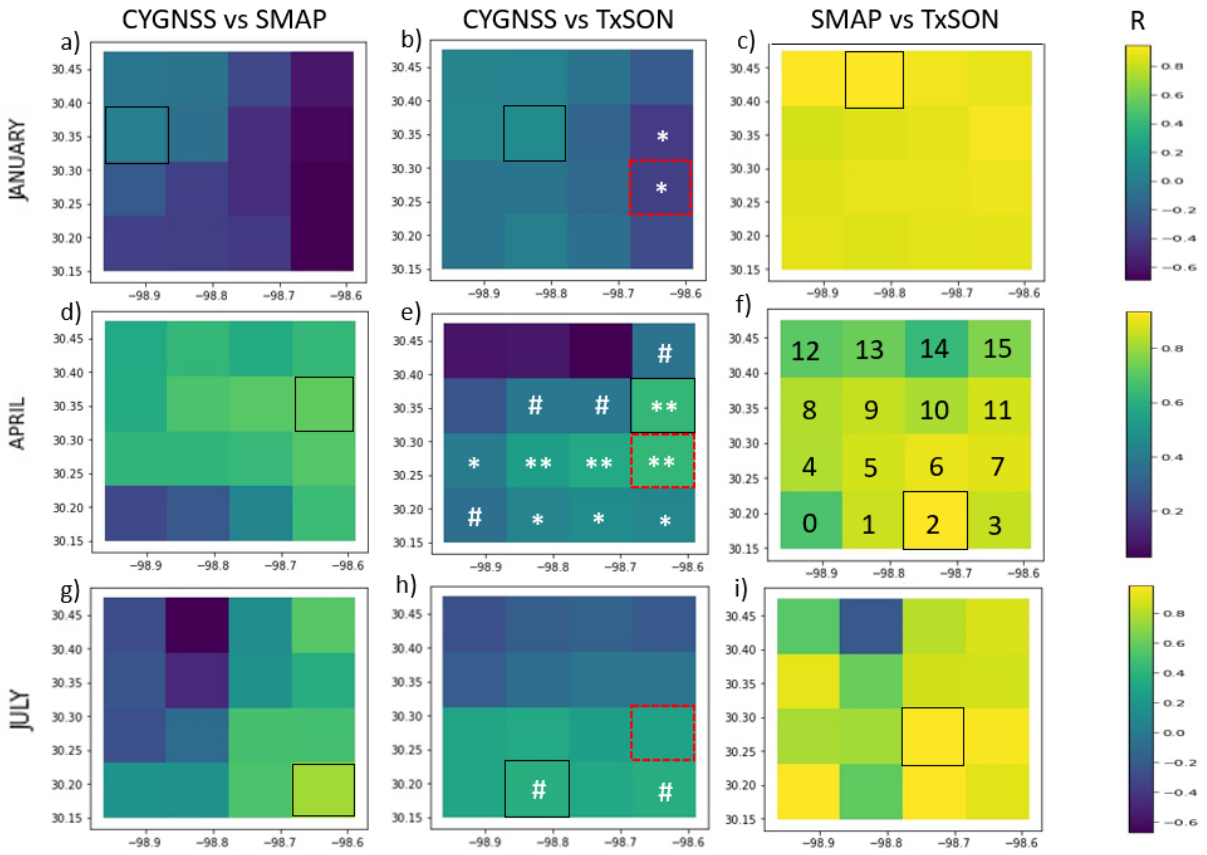

FIGURE 3.1: Spatial correlation heatmaps for pair-wise comparisons between CYGNSS, SMAP and TxSON. Grid outlined in black represents grid with highest R value in that heatmap. In the second column (CYGNSS vs TxSON) the level of significance for each grid is shown (# - significant at 0.1 level, $*$ - significant at 0.05 level, $**$ - significant at 0.001 level). f) shows the grid numbering followed for all heatmaps.

points can have a negative impact on the correlations. For example in Figure 3.2 h), the data point in the box can be considered as an outlier and on removing this datapoint the correlation value for grid cell 1 in Figure 3.1 h) increases from 0.37 to 0.585. Whether such outlier points can be removed through better data preprocessing, or are affected by land parameters needs to be further investigated. Another source of uncertainty is the lack of SMAP data for the month of July. Only 5 days of data are available for this month, which surely effects the correlation analysis and does not necessarily yield a clear picture of how well SMAP compares with CYGNSS and/or TxSON.

We also carry out a similar analysis between the daily change in values of CYGNSS, SMAP and TxSON (Figure 3.3).

While there is an increase in absolute values of grid correlations across the TxSON region for the CYGNSS vs TxSON comparison for January and CYGNSS vs SMAP comparison for April, the number of grid cells that are significantly correlated at 0.1, 0.05 and 0.001 levels has decreased.

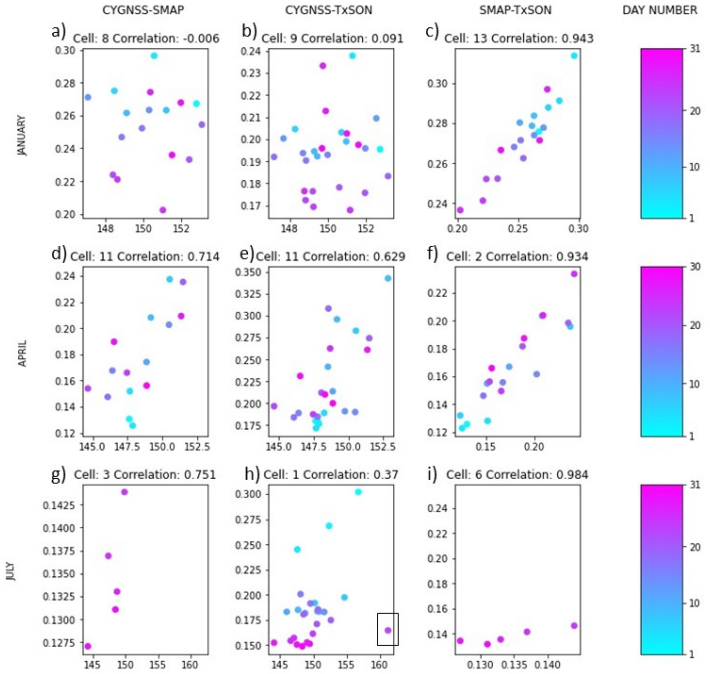

FIGURE 3.2: Scatter plots for the grid (outlined in black in Figure 3.1) with the highest R value in each heatmap in Figure 3.1. Data point within box in h) represents an outlier (discussed in text)

The above analysis demonstrates that the linear relation between SM and $P_{r,eff}$ changes with space and time. When we study the effect of physical land parameters (surface roughness, clay fraction, elevation, NDVI and DepRes) on Pr,eff and their combined potential in estimating soil moisture, we yield Tables 3.1 and 3.2

3.2 Statistical Analysis of Additional Variables with $P_{r,eff}$

Table 3.1 shows the Pearson's R values for each variable with $P_{r,eff}$. For all three months elevation is significantly negatively correlated with $P_{r,eff}$. However, it is interesting to note that the magnitude of correlation for elevation, Clay and DepRes decreases as we move from winter to the summer. However only elevation is significantly correlated with $P_{r,eff}$ for all three months. We further investigate the relations between variables that are significantly correlated with $P_{r,eff}$, through a regression analysis

We study the significance of the coefficient of the variables using the p-value approach. From Table 3.2 it can be seen that the coefficients of elevation, DepRes are significant and therefore have a significant linear correlation with $P_{r,eff}$. This also means that Elevation and DepRes may not add any additional information when used along with $P_{r,eff}$ in a linear model. However, the possibility of non-linear relationships between these variables can not be ruled out. We

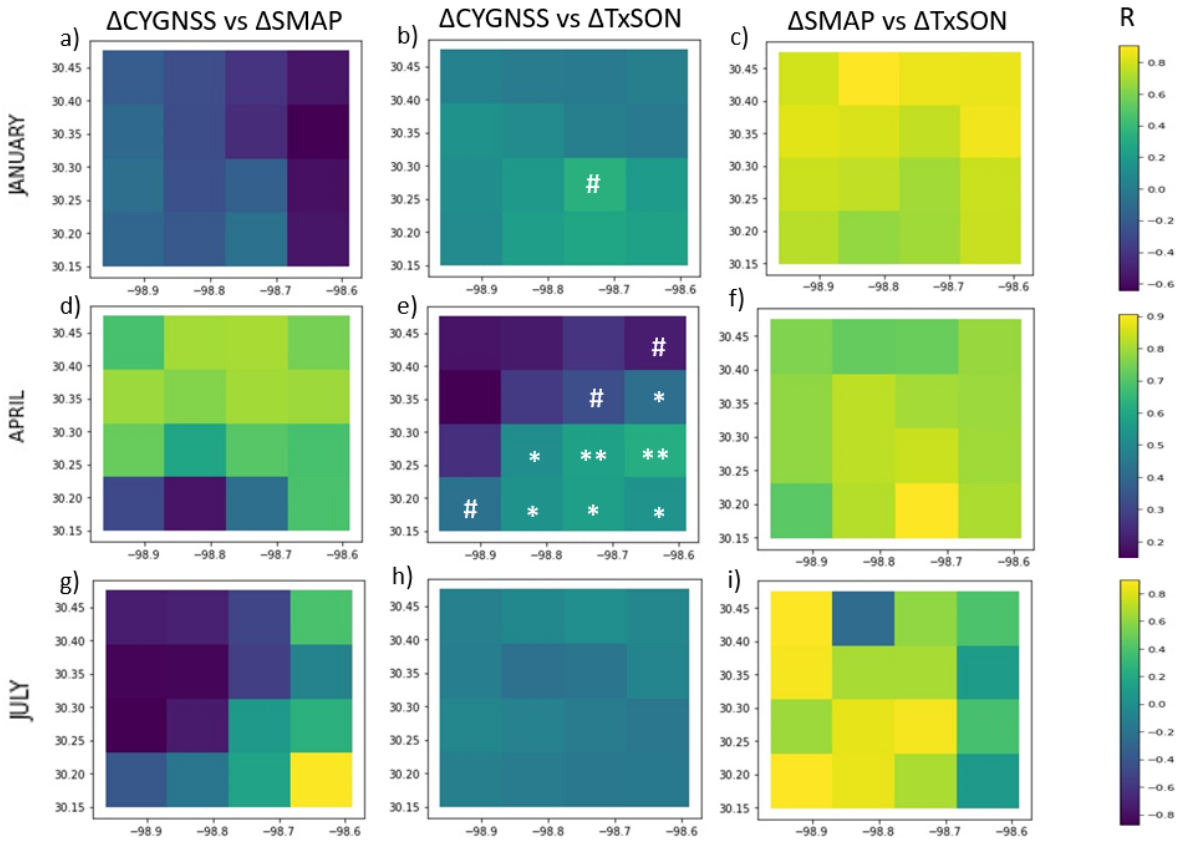

FIGURE 3.3: Temporal correlation heatmaps for pair-wise comparisons between CYGNSS, SMAP and TxSON. Similar to Figure 3.1, significance levels are shown for grids in the second column (CYGNSS vs TxSON comparison). Grid numbering is same as Figure 3.1

		Θ	Elevation	Clay	DepRes	NDVI	Γ_M	Γ_V	Γ_S	Γ_K
January	$P_{r,eff}$	-.058	**-0.542**,	**-0.240**	**0.315**	-0.148	-0.57	0.057	0.056	-0.056
April	$P_{r,eff}$	-.006	**-0.421**	-.103	0.183	0.023	-.074	0.072	0.07	-0.071
July	$P_{r,eff}$	-.103	**-0.305**	-.006	**0.123**	-0.105	-0.093	0.094	0.101	-0.104

TABLE 3.1: Correlations of $P_{r,eff}$ and ancillary data variables. The bold font indicates that correlations are significant at the 0.001 level

investigate more complex relations between these variables in the next section with the help of an Artificial Neural Network.

Variable	Coefficient	P-value
Elevation	-0.893	8.1 x 10^{-28}
DepRes	0.413	7.9 x 10^{-7}
Clay	-0.126	**0.134**

TABLE 3.2: Coefficients for Eq. 2.3. Bold font indicates NOT significant at 0.001 level

3.3 ANN Model Results

In this section we present the results of our SM retrieval results for the three studied months in 2019, for spatial resolutions of 9x9 km^2 and 3x3 km^2. For both these spatial resolutions we first assess the individual contributions of input features, and then combinations of input features from only CYGNSS-derived variables for our baseline, and then investigate the contributions of individual ancillary variables by adding them to this baseline combination. Tables 3.3 and 3.3 show the results on the test set for the 9x9 km^2 and 3x3 km^2 grids respectively. In addition to individual ancillary variables, we also examine combinations of ancillary variables along with the CYGNSS-only variables.

Input Combination	R	RMSE
$P_{r,eff} + \Theta + \Gamma_M + \Gamma_V + \Gamma_S + \Gamma_K$	0.6642	0.0429
$P_{r,eff} + \Theta + \Gamma_M + \Gamma_V + \Gamma_S + \Gamma_K$ + Elevation	0.6713	0.0425
$P_{r,eff} + \Theta + \Gamma_M + \Gamma_V + \Gamma_S + \Gamma_K$ + NDVI	0.6801	0.0423
$P_{r,eff} + \Theta + \Gamma_M + \Gamma_V + \Gamma_S + \Gamma_K$ + DepRes	0.6428	0.0442
$P_{r,eff} + \Theta + \Gamma_M + \Gamma_V + \Gamma_S + \Gamma_K$ + Clay	0.5809	0.0467
$P_{r,eff} + \Theta + \Gamma_M + \Gamma_V + \Gamma_S + \Gamma_K$ + Elevation + NDVI + DepRes	**0.7024**	**0.0409**
$P_{r,eff} + \Theta + \Gamma_M + \Gamma_V + \Gamma_S + \Gamma_K$ + Elevation + NDVI + DepRes + Clay	0.6656	0.0431

TABLE 3.3: Results for 9km gridded data

Input Combination	R	RMSE
$P_{r,eff} + \Theta + \Gamma_M + \Gamma_V + \Gamma_S + \Gamma_K$	0.5801	0.0553
$P_{r,eff} + \Theta + \Gamma_M + \Gamma_V + \Gamma_S + \Gamma_K$ + Elevation	0.65	0.0523
$P_{r,eff} + \Theta + \Gamma_M + \Gamma_V + \Gamma_S + \Gamma_K$ + NDVI	0.6104	0.0538
$P_{r,eff} + \Theta + \Gamma_M + \Gamma_V + \Gamma_S + \Gamma_K$ + DepRes	0.5697	0.0563
$P_{r,eff} + \Theta + \Gamma_M + \Gamma_V + \Gamma_S + \Gamma_K$ + Clay	0.521	0.0579
$P_{r,eff} + \Theta + \Gamma_M + \Gamma_V + \Gamma_S + \Gamma_K$ + Elevation + NDVI + DepRes	**0.6794**	**0.0497**
$P_{r,eff} + \Theta + \Gamma_M + \Gamma_V + \Gamma_S + \Gamma_K$ + Elevation + NDVI + DepRes + Clay	0.6201	0.0532

TABLE 3.4: Results for 3km gridded data

Similar observations can be made from both the Tables 3.3 and 3.4 and we therefore summarize the key results for both the 9x9 and 3x3 grids in the following points:

1. Adding the clay or the DepRes variables to the baseline combination of only CYGNSS-derived variables results in lower Pearson' s R and higher RMSE values with the change in values for Clay being substantial.

2. Individually Elevation and NDVI information added to the baseline result result in improved performance on the test set.

3. For both the 9x9 km^2 and 3x3 km^2 spatial resolutions the highest R and lowest RMSE are obtained for the combination of CYGNSS-only variables, Elevation, NDVI and DepRes

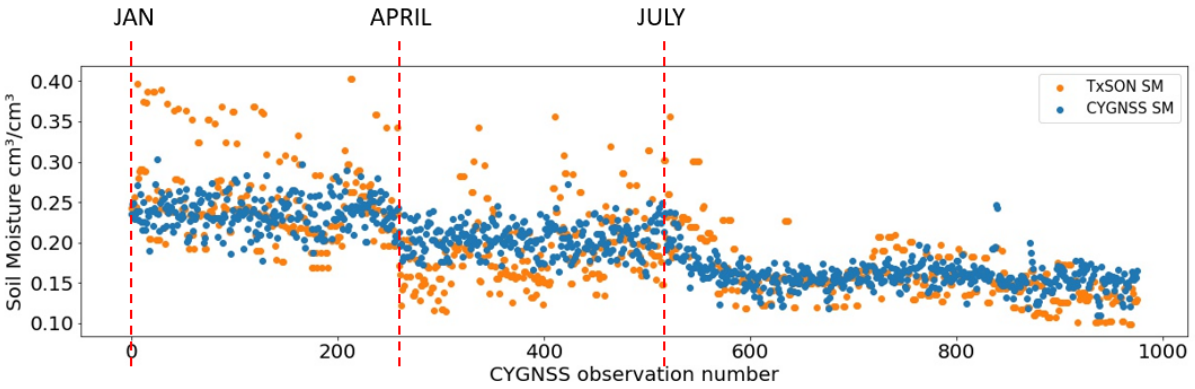

FIGURE 3.4: 9km grid results

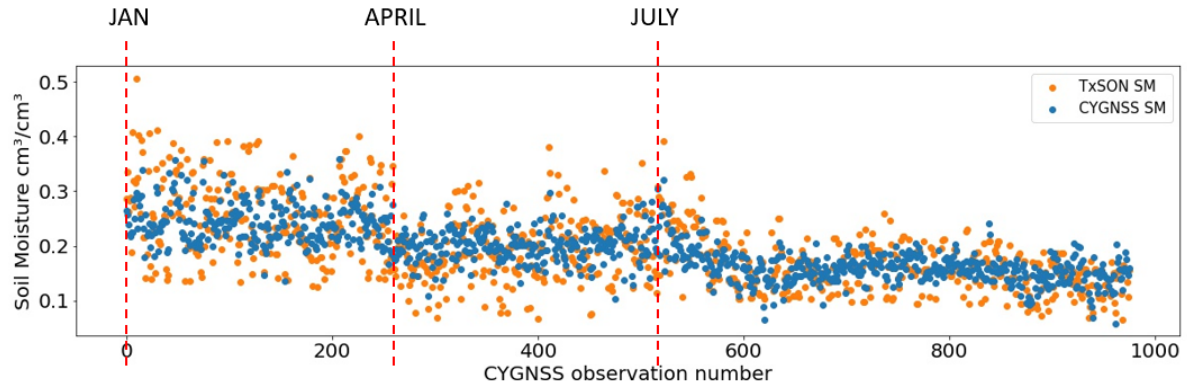

FIGURE 3.5: 3km grid results

After selecting the optimal input features we look to assess model performance on the entire dataset. We run the trained model on the entire dataset to get CYGNSS-derived SM predictions corresponding to each CYGNSS observation. Figure 3.4 and Figure 3.5 show the time series of CYGNSS-derived SM and in-situ (TxSON) SM values . For the month of July results show a strong agreement between the CYGNSS-derived SM and in-situ (training) values. Also the decrease in SM values from April to July is well captured by the model at both grid resolutions. While the overall trend seems to be captured by CYGNSS-derived SM values for the months of January and April, the model seems to struggle in capturing large variations in daily SM values which is more apparent in the 9x9 km^2 resolution.

Chapter 4

Conclusions

We first study the spatial and temporal variability of CYGNSS-derived surface reflectivity and two reference SM datasets, SMAP and TxSON. This is done through a grid-wise correlation analysis. It is shown that the correlations are not stable across different months and correaltions fluctuate due to spatial and temporal changes. This further motivates the usage of land geophysical parameters such as elevation, surface roughness, NDVI, Depth to Restrictive Layer and Clay ratio which either vary spatially, temporally or both. We then develop an Artificial Neural Network model for SM retrieval using CYGNSS-derived variables and ancillary variables. The approach is applied using insitu data from TxSON sensors as the reference SM data . The model is trained using a Grid Search Cross-Validation approach to find the optimal model for different combinations of input features. Satisfactory agreement between the prediction and ground truth showed the efficiency of this proposed model that was demonstrated by a correlation coefficient of 0.7024 (0.6794) and an RMSE of 0.0409 (0.0497) cm^3/cm^3 on the test set at the 9x9 (3x3) km^2 grid resolution over the TxSON region.

The key takeaways of this work are as follows:

1. The linear correlations between CYGNSS-derived surface reflectivtity and SM vary significantly both spatially and temporally. This demonstrtaes that a simple linear model to estimate SM from $P_{r,eff}$ may not be sufficient. Either other land physical parameters need to be taken into account or complex non-linear relations need to be captured using more sophisticated methods.

2. In case of a multi-variate linear model for SM estimation, we demonstrate that Elevation and DepRes need not be considered in the model along with $P_{r,eff}$. This is because Elevation and DepRes have a significant linear relationship with $P_{r,eff}$ and therefore do not provide any additional information to the model.

3. Artificial Neural Networks can be used to achieve satisfactory results and an optimal set of input features is established for the task. The features are: surface reflectivity, Specular Point incidence angle (Θ), surface roughness (Γ_M, Γ_V, Γ_S, Γ_K), Elevation, Depth to Restrictive Layer and NDVI.

Bibliography

[1] Mohammad M Al-Khaldi et al. "Time-series retrieval of soil moisture using CYGNSS". In: *IEEE Transactions on Geoscience and Remote Sensing* 57.7 (2019), pp. 4322–4331.

[2] Andres Calabia, Iñigo Molina, and Shuanggen Jin. "Soil Moisture Content from GNSS Reflectometry Using Dielectric Permittivity from Fresnel Reflection Coefficients". In: *Remote Sensing* 12.1 (2020), p. 122.

[3] Todd G Caldwell et al. "The Texas soil observation network: A comprehensive soil moisture dataset for remote sensing and land surface model validation". In: *Vadose Zone Journal* 18.1 (2019), pp. 1–20.

[4] Adriano Camps et al. "Sensitivity of GNSS-R spaceborne observations to soil moisture and vegetation". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9.10 (2016), pp. 4730–4742.

[5] C Chew et al. "The sensitivity of ground-reflected GNSS signals to near-surface soil moisture, as recorded by spaceborne receivers". In: *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2017, pp. 2661–2663.

[6] CC Chew and EE Small. "Soil moisture sensing using spaceborne GNSS reflections: Comparison of CYGNSS reflectivity to SMAP soil moisture". In: *Geophysical Research Letters* 45.9 (2018), pp. 4049–4057.

[7] Clara Chew and Eric Small. "Description of the ucar/cu soil moisture product". In: *Remote Sensing* 12.10 (2020), p. 1558.

[8] Clara Chew et al. "Demonstrating soil moisture remote sensing with observations from the UK TechDemoSat-1 satellite mission". In: *Geophysical Research Letters* 43.7 (2016), pp. 3317–3324.

[9] Maria Paola Clarizia et al. "Analysis of CYGNSS data for soil moisture retrieval". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12.7 (2019), pp. 2227–2235.

[10] Orhan Eroglu et al. "High spatio-temporal resolution CYGNSS soil moisture estimates using artificial neural networks". In: *Remote Sensing* 11.19 (2019), p. 2272.

[11] Scott Gleason, Mounir Adjrad, and Martin Unwin. "Sensing ocean, ice and land reflected signals from space: results from the UK-DMC GPS reflectometry experiment". In: *Proceedings of the 18th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS 2005)*. 2005, pp. 1679–1685.

[12] Yan Jia et al. "GNSS-R soil moisture retrieval based on a XGboost machine learning aided method: Performance and validation". In: *Remote sensing* 11.14 (2019), p. 1655.

[13] Hyunglok Kim and Venkat Lakshmi. "Use of Cyclone Global Navigation Satellite System (CYGNSS) observations for estimation of soil moisture". In: *Geophysical Research Letters* 45.16 (2018), pp. 8272–8282.

[14] Son V Nghiem et al. "Wetland monitoring with global navigation satellite system reflectometry". In: *Earth and Space Science* 4.1 (2017), pp. 16–39.

[15] Volkan Senyurek et al. "Evaluations of a Machine Learning-Based CYGNSS Soil Moisture Estimates against SMAP Observations". In: *Remote Sensing* 12.21 (2020), p. 3503.

[16] Volkan Senyurek et al. "Machine learning-based CYGNSS soil moisture estimates over ISMN sites in CONUS". In: *Remote Sensing* 12.7 (2020), p. 1168.

[17] Qingyun Yan et al. "Pan-tropical soil moisture mapping based on a three-layer model from CYGNSS GNSS-R data". In: *Remote Sensing of Environment* 247 (2020), p. 111944.