UNCERTAINTY QUANTIFICATION IN MACHINE LEARNING BASED RETRIEVAL OF SOIL MOISTURE FROM GNSS-R OBSERVATIONS

G. Tsagkatakis[1,2], A. Melebari[3], R. Akbar[4], J. D. Campbell[3], E. Hodges[3], and M. Moghaddam[3]

[1]Institute of Computer Science, Foundation for Research and Technology - Hellas, Greece.
[2]Computer Science Department, University of Crete, Greece.
[3]Electrical and Computer Engineering, University of Southern California, USA.
[4]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, USA

ABSTRACT

While microwave imaging satellites, such as the NASA Soil Moisture Active Passive (SMAP), can provide reliable estimates of surface soil moisture at km resolution, the temporal frequency of observations is on the order of days. To increase the temporal frequency of observations, a new class of approaches considers global navigation satellite system (GNSS)-reflectometry (GNSS-R) signals. In this work, we consider observations from the NASA Cyclone GNSS (CYGNSS) constellation, as well as auxiliary observations, and seek to provide instantaneous soil moisture estimates. To achieve accurate retrievals, a novel machine learning approach for probabilistic regression is considered, namely the NGBoost. In addition to achieving an accuracy comparable to previous approaches employing state-of-the-art machine learning methods, the considered framework also provides prediction intervals to quantify prediction uncertainty. Using observations from the Yanco SMAP core validation site in southeast Australia over a period of three years, we quantify the performance in terms of both retrieval accuracy and associated uncertainty. Furthermore, using noisy observations, we experimentally demonstrate the impact of input noise on the prediction uncertainty.

CYGNSS, soil moisture, machine-learning, GNSS-R, NGBoost.

1. INTRODUCTION

Soil Moisture (SM) plays an important role in weather and climate by being part of both the water and the carbon cycles. Over the last decades, several remote sensing platforms have been deployed to monitor soil moisture. Current flagship missions include the National Aeronautics and Space Administration (NASA) Soil Moisture Active Passive (SMAP) [1] and European Space Agency (ESA) Soil Moisture and Ocean Salinity (SMOS) [2], which provide soil moisture on a global scale, through microwave imaging. The derived products are typically coarse in resolution, e.g., 36 km for SMAP, and are characterized by revisit frequency on the order of days. Unfortunately, the spatio-temporal resolution of existing products is coarser than required for applications like water management at agricultural field scale [3].

Recently, there have been multiple studies that employ global navigation satellite system (GNSS)-reflectometry (GNSS-R) to retrieve soil moisture [4]. The NASA Cyclone GNSS (CYGNSS) mission has been used recently for the estimation of soil moisture and other geophysical parameters [5]. CYGNSS consists of eight low-orbit receiver satellites called observatories. Each satellite has a zenith antenna for receiving direct Global Positioning System (GPS) signal and two nadir-looking antennas for receiving GPS signal reflected from the Earth's surface. The output of CYGNSS measurement is a delay-Doppler map (DDM). Unlike conventional microwave sensing platforms like SMAP and SMOS, which have a (global) revisit frequency of 2-3 days due to their sun-synchronous orbit, the CYGNSS constellation achieves an observation interval of 2.8 (median) and 7.2 (mean) hours between 38°S and 38°N [6].

Various geophysical parameters have been estimated from the recorded DDMs including surface soil moisture [7, 8]. Chew et al. [9] developed a soil moisture retrieval algorithm utilizing CYGNSS data within a regression model that was calibrated using SMAP soil moisture data. Regression models were also employed in [10]. In addition to regression, more powerful machine learning (ML) models have also been considered. In [11], Eroglu et al. considered incidence angles, derived reflectivity and trailing edge slope values, and ancillary data such as normalized difference vegetation index as inputs to an artificial neural network (ANN)-based regression model. In [12], Senyurek et al. retrieved soil moisture using ANN, random forests (RF), and support vector machine (SVM) with measurements from International Soil Moisture Network (ISMN) as training and validation data. This was later extended in [13] which considered ML for SM retrieval

This work was supported by the TITAN ERA Chair project (contract no. 101086741) within the Horizon Europe Framework Program of the European Commission, and by NASA grant number 80NSSC18K0704 with the University of Southern California.

from CYGNSS with a targeted spatial resolution of $9\,\mathrm{km}$.

More recently, a new class of ML methods based on gradient boost has been considered for problems in remote sensing, given the strong performance of such methods with tabular data. The XGBoost algorithm, a high-performance case of gradient boosting, was considered for SM retrieval from CYGNSS observations in [14, 15].

While providing an estimation of the SM is extremely valuable, providing an estimation of the prediction uncertainty is also critical. A key novelty of this work is that we explore how ML based retrieval can be considered from an uncertainty perspective. Furthermore, unlike state-of-the-art approaches that use the reported specular point of the DDM along with other parameters derived from the DDM as inputs, we consider an extended collection of 3×5 DDM cells. Unlike the published approaches, which retrieve SM at a daily scale, we perform instantaneous retrieval of SM. By doing so, we exploit the high temporal resolution of the CYGNSS platform in providing SM at sub-daily temporal scales. In the majority of cases, studies consider SMAP as reference and validation data with resolutions of 9 and 36 km [16]. Employing a potentially noisy satellite-derived estimate of soil moisture as the target for training the ML model can lead to inaccuracies and bias. To address this issue, we consider spatio-temporally neighboring measurements from in-situ sensors for the target variables.

2. PROPOSED APPROACH

Formally, we assume the availability of a set of N input-output pairs encoded in the dataset $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N}$. In our model, CYGNSS DDM measurements and ancillary observations act as the inputs $\mathbf{x} \in \mathbb{R}^k$, and the associated surface soil moisture values act as the target variables $y \in \mathbb{R}$, where k is the dimensionality of the input. To obtain probabilistic predictions, we consider the natural gradient boost (NGBoost) [17], a recently proposed ML approach for probabilistic regression.

The objective in NGBoost is to estimate the parameters of a specific conditional distribution $p(y|\mathbf{x}, \theta) = P_\theta(y|x)$, parameterized by θ. The NGBoost algorithm is defined based on three modules, namely the type and number of base learners $f^{(n)}$, the parametric form of distribution P_θ, and the scoring rule $\mathcal{S}(P_\theta, y)$. During the training of the NGBoost, the training set $\mathbf{x}$ is fit through a sequence of base learners

$$\{f^{(n)}(\mathbf{x})\}_{n=1}^{N} \tag{1}$$

in a sequential fashion following the boosting paradigm [18]. These base learners can be decision trees or other types of (typically) shallow classifiers. The cascade of base learners is employed for estimating the parameters of the distribution P_θ. In the case of a Gaussian distribution, the parameterization involves the mean and standard deviation of the distribution, i.e., $\theta = (\mu, \sigma)$, where each parameter is associated with a different set of base learners $f_\mu^{(n)}$ and $f_\sigma^{(n)}$.

While in typical point-wise predictions, training amounts to minimizing an appropriate loss function, for the case of probabilistic regression, this role is taken up by a *scoring rule*, which compares the estimated probability distribution to the observations. For the case of Gaussian distributions, a proper scoring rule $\mathcal{S}(\theta, y)$ is the negative log-likelihood that is given by

$$\mathcal{S}(\theta, y) = -\frac{1}{N}\sum_{i=1}^{N} \log P_\theta(y^{(i)}|x^{(i)}) \tag{2}$$

A key novelty of NGBoost is the introduction of the Natural Gradient during the optimization which more closely captures the gradient in the parametric space of the distribution, allowing the estimation of the steepest ascent direction. Increasing the scoring rule in this case is facilitated by estimating the natural gradient according to

$$\tilde{\nabla}\mathcal{S}(\theta, y) \;\propto\; \mathcal{I}_\mathcal{S}(\theta)^{-1}\nabla\mathcal{S}(\theta, y) \tag{3}$$

where $\mathcal{I}_\mathcal{S}(\theta)^{-1}$ is the Riemannian metric of the manifold defined through θ, induced by $\mathcal{S}$.

3. DATA ACQUISITION

The data consist of CYGNSS measurements and in-situ soil moisture values. In addition, there are ancillary data, which are the elevation and the ground slope estimated from Shuttle Radar Topography Mission (SRTM) digital elevation map (DEM) [19]. An overview of the CYGNSS system and the validation site is provided in Section 3.1 and Section 3.2, respectively. The method of compiling the dataset is given in Section 3.3.

3.1. CYGNSS System

CYGNSS is a GNSS-R NASA Earth Venture mission consisting of eight low-Earth orbit satellites at 35° inclination [6]. Each satellite has three antennas. There is one zenith antenna for navigation and estimating GPS effective isotropic radiated power (EIRP) and the other two antennas are nadir-looking antennas and for GNSS-R measurements. There are four receiving channels in each satellite, enabling the acquisition of four simultaneous measurements for each. The output of the GNSS-R measurement includes a a DDM. The processing integration time for each measurement was one second in the beginning. However, since 3 July 2019, the integration time has been changed to half a second (2 Hz). The estimated location of the specular point (SP) in the DDM is provided in the CYGNSS data. In this paper, the location of the SP will be used as the location of the DDM.

3.2. Validation Site

The SMAP Yanco site [20], located in southeast Australia, was selected for this study because: (i) it is within the CYGNSS coverage area; (ii) it is relatively flat; (iii) the vegetation does not play a significant role; and (iv) good quality data are available. The site includes 13 sensor stations. In each one, soil moisture, temperature, and precipitation are measured every 20 min. We consider soil moisture measured between zero and 5 cm depth. Data from 2019 and 2020 were used for training and validation. During this time period, soil moisture was below $0.5\,\mathrm{m}^3\,\mathrm{m}^{-3}$. In both training and validation, only the surface soil moisture values were used.

3.3. Generating Analysis-Ready Dataset

CYGNSS DDMs over the Yanco region were collected in the dataset. The bistatic radar cross section (BRCS) were converted to reflectivity using [21, eq. 24-26]. DDMs with quality flags or signal-to-noise ratio (SNR) less than 10 dB were discarded. DDMs with precipitation were discarded, since precipitation can cause soil moisture to change significantly between the time of a CYGNSS measurement and the time of the nearest in-situ measurement. These criteria for selecting CYGNSS data are similar to the criteria used in other soil moisture retrieval methods using CYGNSS data [9, 12]. To generate the training and validation datasets, an in-situ soil moisture value from the reference site was associated with a CYGNSS measurement if 1) the in-situ measurement was within six hours from CYGNSS measurement and 2) the location of the in-situ measurement was within 5 km of the CYGNSS DDM location.

4. EXPERIMENTAL RESULTS

To quantify the performance of different approaches, we consider both the point-wise retrieval accuracy and the quality of the prediction uncertainty. We employ three widely used metrics in this domain, namely, the root mean squared error (RSME) and the unbiased RMSE (ubRMSE).

4.1. Point-prediction accuracy

Table 1 presents the average prediction accuracy for five independent trials (including the standard deviation in parentheses) for the three approaches under investigation, namely retrieval using SMAP observations (baseline), as well as retrieval using XGBoost and NGBoost from CYGNSS measurements on the validation set. The results demonstrate that both ML-based retrieval approaches exhibit higher retrieval accuracy compared to SMAP when compared against the in-situ observations. Comparing the traditional XGBoost with the probabilistic NGBoost, we observe that both approaches provide equally good point predictions across all metrics.

Table 1. Retrieval accuracy from SMAP (baseline) and ML-based CYGNSS.

	RMSE $[m^3/m^3]$	ubRMSE $[m^3/m^3]$
SMAP (baseline)	0.103 (0.005)	0.151 (0.007)
CYGNSS (XGBoost)	0.055 (0.002)	0.059 (0.001)
CYGNSS (NGBoost)	0.058 (0.003)	0.060 (0.001)

Fig. 1 showcases the performance via scatter plots for the validation set, reporting the predictions of both the proposed NGBoost-based method, as well as the baseline predictions from SMAP. The scatter plot indicates that the prediction outliers for the case of SMAP are relatively equally distributed along the SM range, while for the case of the proposed approach, errors are concentrated along large values of SM where fewer observations are available.

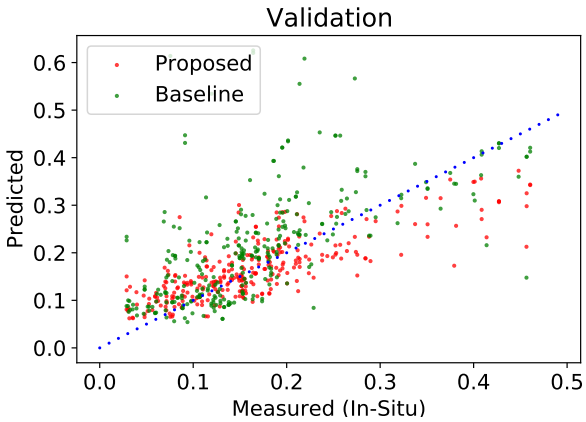

Fig. 1. Scatter plot of predicted vs measured SM for the validation.

4.2. Impact of signal noise

To demonstrate that the estimated uncertainty is closely coupled with the input signals, we investigate the prediction uncertainty of the NGBoost, given two noise conditions: keeping only DDMs with SNR above 10 dB and keeping all measurements above 0 dB. Fig. 2, and 3 present the characteristics of the retrieval uncertainty, i.e., standard deviation, as a function of (measured) SM for two cases of signal noise, as well as the mean value (red line).

Overall, these results confirm the expectation that higher noise levels should lead to higher prediction uncertainty. Furthermore, while in both cases higher values of SM are asso-

ciated with higher prediction uncertainty, the phenomenon is observed at lower SM values for the noisy case.

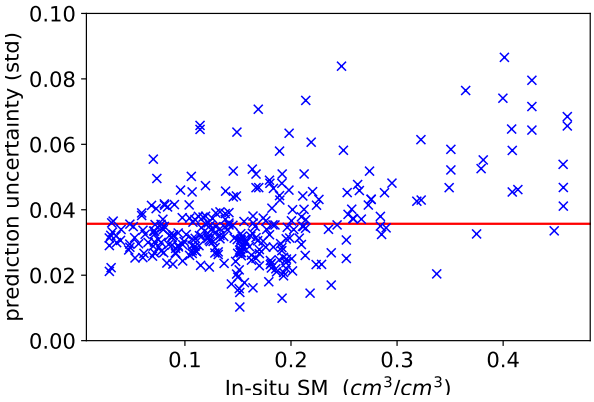

Fig. 2. Prediction uncertainty as a function of SM for high SNR case (DDM SNR threshold at 10 dB).

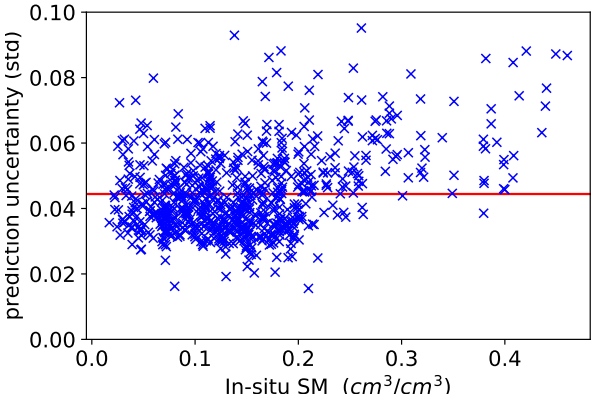

Fig. 3. Prediction uncertainty as a function of SM for low SNR case (DDM SNR threshold at 0 dB).

4.3. Quantifying uncertainty estimation

To quantify the reliability and accuracy of the prediction intervals generated by the model, we employ the coverage probability metric, which in the context of probabilistic regression with Gaussian predictions refers to the proportion of times that the measured value lies within the prediction intervals generated by the model.

Formally, let $\hat{y}_i$ be the predicted mean value and $\hat{\sigma}_i$ the standard deviation for observation i, and let y_i be the corresponding measured value. For a $(1-\alpha)$ prediction interval, we can use the quantile of the standard normal distribution, denoted as $z_{1-\frac{\alpha}{2}}$, where α is the significance level. The prediction interval is then denoted as $[\hat{y}_i - z_{1-\frac{\alpha}{2}}\hat{\sigma}_i, \hat{y}_i + z_{1-\frac{\alpha}{2}}\hat{\sigma}_i]$.

The coverage probability can be calculated as:

$$\text{CP} = \frac{1}{n}\sum_{i=1}^{n} I\left(y_i \in [\hat{y}_i - z_{1-\frac{\alpha}{2}}\hat{\sigma}_i, \hat{y}_i + z_{1-\frac{\alpha}{2}}\hat{\sigma}_i]\right), \quad (4)$$

where n is the number of observations, and $I(\cdot)$ is an indicator function that equals 1 if the condition is true and 0 otherwise.

Table 2 presents the theoretical and estimated coverage probabilities for both training and validation sets. Based on these results, two critical observations can be made. First, we observe that the coverage probability of the retrieval method closely follows the theoretical case. Second, we observe that there is a limited difference between the performance of the training and the validation sets. These results indicate that the model's prediction intervals are reliable and accurately capture the true values without suffering from overfitting.

Table 2. Coverage probability for theoretical (Gaussian), training, and validation data.

Significance level (α)	0.31	0.05	0.01
Gaussian	68.2	95.4	99.7
Training	63.7	89.6	94.2
Validation	63.1	89.6	93.8

5. CONCLUSION

In this work, we focus on the introduction of a probabilistic regression framework based on gradient boosting for estimating geophysical parameters from remote sensing observations. We specifically focus on surface soil moisture retrieval from GNSS-R measurements, although the general principles are applicable to other remote sensing-based variable predictions. Our analysis indicates that using algorithms such as the NGBoost can provide valuable insights that are aligned with expectations from the underlying physical processes.

Acknowledgment

The authors thank NASA PO.DAAC for making CYGNSS data available and Jeffrey Walker from Monash University for Yanco soil moisture data.

6. REFERENCES

[1] D. E. et al., "The Soil Moisture Active Passive (SMAP) Mission," *Proceedings of the IEEE*, vol. 98, no. 5, pp. 704–716, 2010.

[2] Y. H. Kerr, P. Waldteufel, J.-P. Wigneron, S. Delwart, F. Cabot, J. Boutin, M.-J. Escorihuela, J. Font, N. Reul, C. Gruhier *et al.*, "The SMOS mission: New tool for monitoring key elements of the global water cycle," *Proceedings of the IEEE*, vol. 98, no. 5, 2010.

[3] H. V. et al., "On the spatio-temporal dynamics of soil moisture at the field scale," *Journal of Hydrology*, vol. 516, pp. 76–96, 2014.

[4] X. Wu, W. Ma, J. Xia, W. Bai, S. Jin, and A. Calabia, "Spaceborne GNSS-R soil moisture retrieval: Status, development opportunities, and challenges," *Remote Sensing*, vol. 13, no. 1, p. 45, Dec 2020.

[5] H. Carreno-Luengo, J. A. Crespo, R. Akbar, A. Bringer, A. Warnock, M. Morris, and C. Ruf, "The CYGNSS Mission: On-Going Science Team Investigations," *Remote Sensing 2021, Vol. 13, Page 1814*, vol. 13, no. 9, p. 1814, 5 2021.

[6] C. S. Ruf, C. C. Chew, T. Lang, M. G. Morris, K. Nave, A. Ridley, and R. Balasubramaniam, "A New Paradigm in Earth Environmental Monitoring with the CYGNSS Small Satellite Constellation," *Scientific Reports*, vol. 8, no. 1, pp. 1–13, 12 2018.

[7] M. P. Clarizia, N. Pierdicca, F. Costantini, and N. Floury, "Analysis of CYGNSS data for soil moisture retrieval," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2227–2235, 2019.

[8] A. Azemati, A. Melebari, J. Campbell, J. Walker, and M. Moghaddam, "GNSS-R Soil Moisture Retrieval for Flat Vegetated Surfaces Using a Physics-Based Bistatic Scattering Model and Hybrid Global/Local Optimization," *Remote Sensing*, vol. 14, no. 13, 2022.

[9] C. Chew and E. Small, "Description of the UCAR/CU Soil Moisture Product," *Remote Sensing*, vol. 12, no. 10, p. 1558, 5 2020.

[10] Y. Jia, Q. Yan, S. Jin, and P. Savi, "CYGNSS soil moisture estimation using machine learning regression," in *IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2021.

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

[12] V. Senyurek, F. Lei, D. Boyd, M. Kurum, A. C. Gurbuz, and R. Moorhead, "Machine learning-based CYGNSS soil moisture estimates over ISMN sites in CONUS," *Remote Sensing*, vol. 12, no. 7, 4 2020.

[13] F. Lei, V. Senyurek, M. Kurum, A. Gurbuz, R. Moorhead, and D. Boyd, "Machine-learning based retrieval of soil moisture at high spatio-temporal scales using CYGNSS and SMAP observations," in *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2020, pp. 4470–4473.

[14] Y. Jia, S. Jin, P. Savi, Y. Gao, J. Tang, Y. Chen, and W. Li, "GNSS-R soil moisture retrieval based on a XGboost machine learning aided method: Performance and validation," *Remote Sensing*, vol. 11, no. 14, Jul 2019.

[15] K. Edokossi, A. Calabia, S. Jin, and I. Molina, "GNSS-reflectometry and remote sensing of soil moisture: A review of measurement techniques, methods, and applications," *Remote Sensing*, vol. 12, no. 4, Feb 2020.

[16] V. Senyurek, F. Lei, D. Boyd, A. C. Gurbuz, M. Kurum, and R. Moorhead, "Evaluations of Machine Learning-Based CYGNSS Soil Moisture Estimates against SMAP Observations," *Remote Sensing*, vol. 12, no. 21, p. 3503, 10 2020.

[17] T. Duan, A. Anand, D. Y. Ding, K. K. Thai, S. Basu, A. Ng, and A. Schuler, "NGBoost: Natural gradient boosting for probabilistic prediction," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 2690–2700.

[18] T. Chen and C. Guestrin, "XGboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794.

[19] Earth Resources Observation and Science (EROS) Center, "Shuttle Radar Topography Mission 1 Arc-Second Global," 2018. [Online]. Available: https://doi.org/10.5066/F7PR7TFT

[20] A. B. S. et al., "The Murrumbidgee Soil Moisture Monitoring Network Data Set," *Water Resources Research*, vol. 48, no. 7, 7 2012.

[21] S. Gleason, A. O'Brien, A. Russel, M. M. Al-Khaldi, and J. T. Johnson, "Geolocation, Calibration and Surface Resolution of CYGNSS GNSS-R Land Observations," *Remote Sensing 2020, Vol. 12, Page 1317*, vol. 12, no. 8, p. 1317, 4 2020.