



Quantifying Uncertainty in Machine Learning Based Soil Moisture Retrieval From GNSS-R Measurements

G. Tsagkatakis^{*(1,2)}, A. Melebari⁽³⁾, J. D. Campbell⁽³⁾, E. Hodges⁽³⁾ and M. Moghaddam⁽³⁾

(1) Institute of Computer Science, FORTH, Greece e-mail: greg@ics.forth.gr

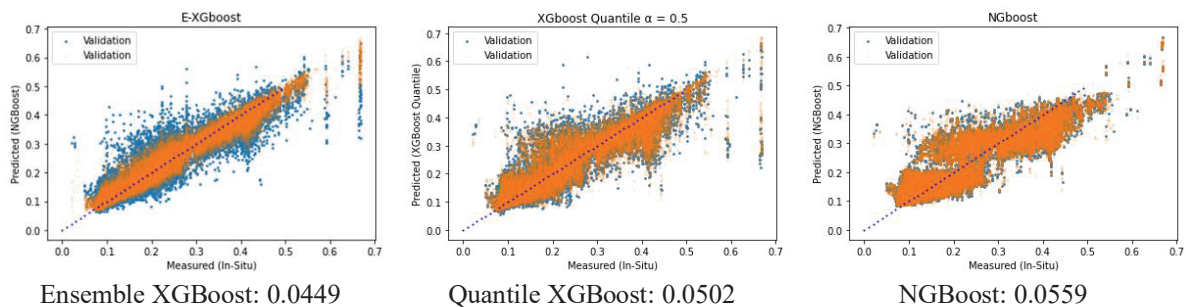
(2) Computer Science Department, University of Crete, 73100, Greece,

(3) Department of Electrical and Computer Engineering, University of Southern California, CA 90007 USA (e-mail: amelebar@usc.edu, jamesc91@usc.edu, ehodges@usc.edu, mahta@usc.edu).

Spaceborne platforms like the NASA SMAP and the ESA SMOS offer global-scale coarse-resolution observations that can be used to estimate surface soil moisture. However, they acquire observations at a moderate revisit frequency, reducing the impact on applications that require higher temporal resolutions. Global navigation satellite system reflectometry (GNSS-R) is a prime example of a signal-of-opportunity (SoOP) which has been shown to be highly sensitive to variations in soil moisture. In this work, we consider measurements from the NASA Cyclone GNSS (CYGNSS) mission which consists of eight low-orbit observatories, that are equipped with two nadir-looking antennas to receive reflected GPS signals. To extract soil moisture content, the traditional approach involves inverting parameterized forward models which account for aspects like dielectric properties of the soil, topography, and incidence angle information among others. Despite their potential, these models typically involve simplified assumptions and demonstrate increased sensitivity to the values of different parameters. To address this challenge, a new line of research tries to address this challenge by utilizing machine learning models and treating the problem as that of supervised regression.

In this work, we focus on the use of machine learning models for the retrieval of soil moisture. We focus on models based on gradient boosting trees algorithm, specifically the XGBoost algorithm, from CYGNSS derived delay-Doppler maps and auxiliary measurements, which has shown exceptional performance in retrieval. The novelty of this work is that in addition to retrieval accuracy, we also explore the ability to quantify prediction uncertainty. To that end, we consider three distinct scenarios for uncertainty quantification, centered around the gradient boosting trees algorithm. In the first case, we utilize the XGBoost in an ensemble learning scenario where multiple models are trained on different subsets of the training set data. During inference/run-time, all the models are collectively applied to the new measurements, producing a distribution of soil moisture values. In the second case, we consider the utilization of XGBoost under a quantile regression framework, where three models are trained to produce the 25th, 50th and 75th quantiles. In the third case, we replace XGBoost with NGBoost, a recently proposed method that outputs an estimate of the mean and variance from a Gaussian distribution for each retrieval.

To train and validate the proposed framework, we consider observations CYGNSS acquired over California, USA in 2023, and soil moisture values derived from the SMAP mission. To quantify performance, 80% of the observations were utilized for training the model, and the rest for validation. The following figure presents scatter plots for validation data in all three cases as well as the retrieval accuracy, measured in unbiased RMSE. The results demonstrate that the introduction of ensemble learning achieves the best retrieval accuracy while introducing quantiles is also accompanied by a reduction in prediction accuracy. The uncertainty-aware NGBoost appears to be the least favorable in terms of retrieval error.



To obtain a better understanding of the quality of the uncertainty quantification, we consider the Prediction Interval Coverage Probability (PICP) estimated at a 95% confidence interval and report the results in the side table. These results demonstrate that although the ensemble approach achieves the best accuracy, the estimated values are the least reliable while the other two approaches are able to capture the uncertainty to a much greater extend.

Ensemble XGBoost	0.74
Quantile XGBoost	0.99
NGBoost	0.97