Refinement of NOAA AMSR-2 Soil Moisture Data Product: 1. Inter-comparisons of the Commonly-used Machine Learning Models

Jifu Yin, Xiwu Zhan, Michael Barlage, Jicheng Liu, Huan Meng, Ralph R. Ferraro

***Abstract*—Soil Moisture is an important variable for hydrological, meteorological and agricultural studies and applications. The Soil Moisture Operational Products System (SMOPS) was developed by the National Oceanic and Atmospheric Administration (NOAA)-National Environmental Satellite, Data, and Information Service (NESDIS) to operationally provide an integrated satellite soil moisture data product. The Advanced Microwave Scanning Radiometer 2 (AMSR2) soil moisture retrieval is an important component of the currently operational SMOPS. This study is proposed to refine the AMSR2 data product using an optimal machine learning model, and this first paper of the two-part series is to intercompare the six commonly-used machine learning models including multiple linear regression (MLR), Regression Tree (RRT), Random Forest (RFT), Gradient Boosting (GBR), Extreme Gradient Boosting (XGB) and Artificial Neural Network (ANN). Results indicate that all of the six approaches can preserve the reference data information beyond the training time period, which ensures them to predict past and future satellite retrievals without a new training procedure. Relative to other models, the XGB method is more successful to respect to the reference data Soil Moisture Active Passive (SMAP) and the in-situ observations from the U. S. Department of Agriculture Soil Climate Analysis Network (SCAN). It has a good implication on the implementation of the XGB model to refine the AMSR2 soil moisture retrievals in the second paper.**

***Index Terms*— AMSR-2 Soil Moisture, Soil Moisture Operational Products System (SMOPS), Machine Learning**

I. INTRODUCTION

Soil moisture (SM) plays a central role in the terrestrial water, biogeochemical and energy cycles through constraining the partitioning of incoming radiation into latent and sensible heat-fluxes [1-2]. A wetter soil moisture condition generally makes plant growth better through altering soil physical properties, physiological hierarchy and soil biogeochemistry, whereas low SM value reduces evaporative cooling and in turn rising temperature [3-5]. Given the positive SM-precipitation feedback at a continental scale, precipitation is generally decreased under a dry SM condition [6]. Those changes in evapotranspiration and precipitation trends could promote the uncertainties of monitoring and predicting rainfall, drought and flood events. Low SM also contributes to fire occurrence, development and propagation through influencing plant productivity and live fuel moisture [7-9]. Accurate understanding the SM status would thus benefit the meteorological, hydrological, climatological and environmental studies and applications.

Unlike the ground in situ measurements, microwave remote sensing offers SM observations with sufficient coverage and consistency at regional and global scales. Links between soil dielectric constant and soil emissivity provide a direct manner to retrieve SM using passive microwave satellite observations [10-11]. The active microwave radar uses an electromagnetic pulse to sense the land surface backscatter that is primarily affected by surface roughness and geometric [12], while the passive microwave radiometer receives the land surface emission impacted by the physical temperature and emissivity of the Earth [10-11]. The traditional microwave satellite SM retrieval algorithms typically include a radiative transfer model linking brightness temperature (Tb) and soil dielectric constant and a dielectric mixing model calculating soil moisture [13].

The first-generation retrieval algorithms estimate soil moisture through minimizing the difference between modeled and observed Tb in either the vertical (V-pol) or the horizontal (H-pol) polarization. However, such SM retrieval models including the Single Channel Algorithm (SCA) and the Land Surface Microwave Emission Model are significantly impacted by surface temperature, vegetation optical depth

This work was jointly supported by NOAA Joint Polar Satellite System (JPSS)-Proving Ground and Risk Reduction (PGRR) Program, NOAA Infrastructure Investment and Jobs Act (IIJA) program, NOAA Climate Program Office (CPO)-Modeling, Analysis, Predictions and Projections (MAPP) Program and NOAA grant (NA19NES4320002). We thank the anonymous reviewers for helping improve the manuscript quality. The authors declare that there is no conflict of interest regarding the publication of this paper. The manuscript contents are solely the opinions of the authors and do not constitute a statement of policy, decision, or position on behalf of NOAA or the U. S. Government. *(Corresponding author: Jifu Yin).*

J. Yin is with the Earth System Science Interdisciplinary Center, Cooperative Institute for Climate and Satellites, University of Maryland at College Park, College Park, MD 20740 USA (e-mail: jifu.yin@noaa.gov).

X. Zhan is with the National Oceanic and Atmospheric Administration/National Environmental Satellite, Data, and Information Service Center for Satellite Applications and Research, College Park, MD 20740 USA.

M. Barlage is with the National Oceanic and Atmospheric Administration/National Centers for Environmental Prediction-Environmental Modeling Center (EMC), College Park, MD 20740 USA.

J. Liu is with the Earth System Science Interdisciplinary Center, Cooperative Institute for Climate and Satellites, University of Maryland at College Park, College Park, MD 20740 USA.

H. Meng is with the National Oceanic and Atmospheric Administration/National Environmental Satellite, Data, and Information Service Center for Satellite Applications and Research, College Park, MD 20740 USA.

R. R. Ferraro is with the Earth System Science Interdisciplinary Center, Cooperative Institute for Climate and Satellites, University of Maryland at College Park, College Park, MD 20740 USA.

(VOD) and land surface roughness generally assigned as constant on the basis of ancillary data [11, 14]. Given the passive microwave satellite sensors can provide multi-frequency Tb measurements, additional observations were used to represent the parameters relevant to surface SM condition. This kind of evolution has been reflected in the Dual Channel Algorithm and the Land Parameter Retrieval Model [15-16]. These advanced algorithms have reduced uncertainties in soil moisture retrievals by obtaining surface temperature from either high frequency microwave Tb observations or thermal infrared ancillary data. However, these physical models still suffer from inadequate representativeness of retrieving parameters [17-18].

Machine learning models have been broadly employed in satellite SM estimation [19-21] in the past decade. These artificial intelligence techniques use high-quality soil moisture datasets as the base reference to train models for satellite SM retrievals. Based on experiment results from test sites and watershed experiment, the linear regression method was initially involved to build the relationships between Tb and SM without considering ancillary information [22-23]. Apart from the traditionally physical SM retrieval model, these attempts offered another way to estimate soil moisture. The earlier-generation machine learning algorithms, such as Random Forest and Regression Tree, were not designed with the relevant architectural elements to enable automatic extraction of features, but they can still produce the fairly accurate satellite SM retrievals [21, 24-26]. Relatively, tree boosting has shown the state-of-the-art results [27]. It considers the errors of the previous trees and then uses sequential learning to make weak trees become as strong learners. Gradient Boosting and Extreme Gradient Boosting approaches have presented the advantages including high speed, excellent efficiency and great accuracy in satellite SM retrievals [7, 28]. The new generation of machine learning with the term deep learning is making major advances in the ability of neural networks to automatically capture data distributions, allowing to empower satellite SM retrieval capabilities [29-30]. Artificial Neural Network is a typical deep learning model to create collection functions of connecting neurons. Its advantages of storing information and having a distributed memory could benefit the development of long-term SM data products [32].

The Advanced Microwave Scanning Radiometer 2 (AMSR2) onboard the Global Change Observation Mission 1st-Water (GCOM-W) satellite was launched by the Japan Aerospace Exploration Agency (JAXA) in 2012 [32]. AMSR2 soil moisture data product is an important component of Soil Moisture Operational Product System (SMOPS) that is developed by National Oceanic and Atmospheric Administration (NOAA) to offer a one-stop microwave satellite SM data product [33-35]. Refinement of the currently operational NOAA AMSR2 soil moisture datasets can not only further improve the SMOPS quality, but also potentially benefit the third generation of AMSR mission series (AMSR3). This study aims to address the open scientific questions including 1) there are many machine learning models used to retrieve satellite soil moisture in the past decade [29-30], which is the optimal approach for AMSR2 SM retrieval? 2) can machine learning reasonably estimate satellite SM without utilizing the retrieving parameters subjectively used in the physical models, such as VOD and surface roughness? 3) does the refined AMSR2 SM have a better performance than the currently operational AMSR2 data product? 4) is the refined AMSR2 comparable to the latest version (V8.0) Soil Moisture Active Passive (SMAP) SM product? This first paper of the two-part series is to address the first two questions, while the latter two will be tackled in the second part.

The reminder of this paper proceeds as follows: Section 2 contains descriptions of the primary data sets used here. The strategies of training and comparing machine learning models are introduced in Section 3. The results focused on model evaluations and differences in model performance are provided in Section 4. The discussions relevant to interpret the validation results and model comparisons are shown in Section 5. And brief summary is finally given in Section 6.

II. DATASETS

Datasets leveraged for building machine learning models include multifrequency microwave Tb observations from AMSR2, the Version 8.0 SMAP SM retrievals, Normalized Difference Vegetation Index (NDVI) of the Moderate Resolution Imaging Spectroradiometer (MODIS), and ancillary maps for characterizing land surface conditions. In situ SM observations from the Soil Climate Analysis Network (SCAN) network are used to validate the trained machine learning models.

2.1 AMSR2 Brightness Temperature

The AMSR2 onboard the GCOM-W1 satellite is a total power microwave radiometer that can detect the microwave energy emitted from the Earth surface at 6.925, 7.3, 10.65, 18.7, 23.8 and 36.5 GHz in vertical and horizontal polarizations. The corresponding footprint ground resolutions with Cross-track (km) × Along-track (km) are 35×62, 34×58, 24×42, 14×22, 15×26 and 7×12 [32]. The AMSR2 frequency set is identical to that of AMSR for Earth-Observation System (AMSR-E) except the 7.3GHz channel for Radio Frequency Interference (RFI) mitigation purpose. The 7.3 GHz Tb observations are thus excluded to enable this study eventually benefit the long-term AMSR mission series including the decommissioned AMSR-E, the currently operational AMSR2 and the upcoming AMSR3. Recent study has revealed that the Tb measurements at 23.8 GHz provide meaningful and valuable information on total precipitable water estimation over the land areas [36]. The 23.8 GHz Tb observations are thus excluded to make the refined AMSR2 SM retrievals independently offer the valuable information without involving satellite precipitation retrieval information.

The GCOM-W1 satellite equatorial crossing time is 13:30 and 1:30 ±15 mins in the ascending and descending tracks, respectively. Given AMSR2 rotation takes 1.5s and its ground speed is about 6.76 km/s, the intervals at the swath center are about 10 km [32]. The ascending and descending AMSR2 Tb observations at 6.925 GHz, 10.65 GHz, 18.7 GHz and 36.5 GHz in dual polarizations over the 3 July 2012-31 December 2021 time period are used in this paper. The original footprint Tb data are gridded to 25 km spatial resolution over the global domain using the nearest neighbor method. The corresponding RFI flags were applied to quality control AMSR2 Tb observations before they were used to train machine learning models and produce SM retrievals.

2.2 SMAP Soil Moisture Data Product

The National Aeronautics and Space Administration (NASA) SMAP was specifically designed to acquire surface soil moisture conditions at high accuracy on the basis of L-band signals [37]. It was successfully launched in January 2015 and began to provide the global SM science data in April 2015. Based on the L-band radiometer and radar sensors, the SMAP mission was targeted to measure SM for the top 5 cm surface layer with retrieval errors smaller than 0.04 m^3/m^3 [37]. After loss of the L-band radar, the SMAP has been continuously providing L-band Tb measurements on the global domain, allowing to generate high quality SMAP passive SM data product as of 2015 [38]. Benefiting from the Dual Channel Algorithm, the most recent version (V8.0) of Level-3 SMAP soil moisture data showed better data quality than previous versions retrieved by the Single Channel Algorithm. The original SMAP Level-3 SM retrievals from 2016 to 2021 were re-gridded to 25 km spatial resolution using nearest neighbor approach. SMAP V8.0 data are accessible from National Snow and Ice Data Center (https://nsidc.org/data/spl3smp/versions/8).

2.3 Ancillary data

The Tb observations are dependent on target variables including land surface roughness, vegetation condition, soil moisture status and physical temperature of the soil and vegetation [39]. It thus needs to adequately represent those parameters while training machine learning models. Specifically, global soil texture map from the Food and Agriculture Organization of the United Nations (FAO) is used to characterize the soil retention capacity for water and the spatial variations of soil type. According to sand, silt and clay proportions, global soils are classified into 9 soil types in the FAO texture map [40]. The 1 km FAO texture map was derived from FAO/United Nations Educational, Scientific and Cultural Organization Soil Map of the World at 1: 5,000,000 scale (https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/faounesco-soil-map-of-the-world/en/), and then upscaled to 25 km spatial resolution.

The annual Visible Infrared Imaging Radiometer Suite (VIIRS) land cover type maps and the MODIS NDVI datasets were used to represent the underlying vegetation conditions. The land surface type team of NOAA-Center for Satellite Applications and Research (STAR) has developed global annual surface type classification maps using previous one or more years of VIIRS surface reflectance, brightness temperature and vegetation index data [41]. Following the International Geosphere-Biosphere Program (IGBP) classification scheme, the annual VIIRS global land cover maps at 1 km spatial resolution offer 17 surface type classes. The composite 16-daily MODIS NDVI observations at 500 m resolution from either Terra or Aqua satellite were combined to develop 8-daily NDVI datasets. MODIS NDVI data are accessible from the NASA EARTHDATA (https://www.earthdata.nasa.gov/), while VIIRS land cover maps can be obtained from the NOAA Comprehensive Large Array-data Stewardship System (https://www.avl.class.noaa.gov/saa/products/welcome).

2.4 SCAN Soil Moisture

The U. S. Department of Agriculture Soil Climate Analysis Network (SCAN) provides ground SM observations over agricultural areas of the United States. Hourly SM measurements are automatically screened for the limits of the sensors through measuring the dielectric constant of the soil [42]. In this paper, the SCAN SM observations were quality controlled by detecting problematic datasets. SCAN offers soil temperature measurements, allowing to exclude SM measurements at corresponding sites and soil layers under frozen conditions [34-35]. We also excluded SM observations outside of the physically possible range of 0~0.6 m^3/m^3. Finally, the 0~5 cm SM measurements over the 2012-2021 time period from 52 SCAN sites within the contiguous United States (CONUS) were chosen by excluding that with fewer than 3-year of observations.

III. METHODOLOGY

3.1 Machine Learning Models

This study is aimed to assess the AMSR2 SM datasets retrieved by the six commonly-used machine learning models, including Multiple Linear Regression (MLR), Regression Tree (RRT), Random Forest (RFT), Gradient Boosting (GBR), Extreme Gradient Boosting (XGB) and Artificial Neural Network (ANN). The MLR is the simplest regression approach to evaluate the relationships between more than 2 independent variables and one dependent variable through fitting a line. MLR method was primarily involved in SM retrieval studies from late 1970s to the end of 20th century [22-23].

RRT uses a tree structure to represent attribute judgments [43]. Each leaf node characterizes a prediction result, while each branch corresponds to a fitting model. By contrast, RFT is an ensemble machine learning method assembling a large number of estimators [44]. The Law of Large Numbers ensure that the RFT does not overfit, while the input feature space is divided into many regression trees. The RFT uses randomized, adaptive and decorrelated features to build better relationships under highly nonlinear conditions [44]. Both RRT and RFT have been widely used in satellite

SM retrievals, as they are easily implemented and take more advantageous for addressing large scale problems [21, 24-26].

Tree boosting is capable of capturing complex patterns and thus has shown state-of-the-art features [27]. GBR uses either learning rate or the number of components to control the degree of fit. The two regularization parameters affect each other with the smaller component size resulting in learning rate increase [45]. The ideal solution is to optimize both by jointly minimizing a model selection criterion. However, it is worth noting that a greater component size produces a proportionate increase in computation [45].

Relatively, XGB is a salable machine learning approach for tree boosting. Benefiting from utilizing a cost function to control model complexity, the model variance could be reduced to make the trained model simpler and in turn to avoid overfitting. The XGB model can scale billions of examples on a memory-limited platform since a novel tree learning algorithm and a theoretically justified weighted quantile sketch procedure are used to handle spare data and instance weights [27].

Unlike the above five models, ANN has an unsupervised feature, though it is also a supervised learning method to connect neurons. It is a nonlinear mathematical computing system that is capable of identifying complex nonlinear relationships between train- and reference-data [46]. The ANN finds the weight for each network through minimizing a loss function representing the difference between data and predictions [30]. It is thus more efficient and more widely-used when the process characteristics are difficult to be physically described [46].

3.2 Machine Learning Framework

The input ancillary datasets of land cover map and NDVI data are relevant to VOD, vegetation water content and surface roughness, while soil type map represents the soil retention capacity and the ratios of soil clay and silt. Those are the critical parameters for the traditionally physical SM retrieval models [11,14,37]. Figure 1 shows the specific input data collection, as well as training and testing procedures for each of the six commonly-used machine learning models. A descending and an ascending model are separately trained using the corresponding 6.925, 10.65, 18.7 and 36.5 GHz AMSR2 Tb observations in dual-polarization, while the corresponding pass-set SMAP SM data products are used as base references. The microwave emission is primarily relying on soil dielectric constant that links soil moisture, allowing to use X-band (8.0-12.0 GHz) and C-band (4.0-8.0 GHz) satellite measurements to retrieve SM in a direct manner [11, 14]. The 6.925 GHz and 10.65 GHz Tb were thus included to respect the physical retrieval theory. The 36.5 GHz is the most appropriate microwave frequency for representing land surface temperature conditions, while the 18.7 GHz is used to characterize water surfaces and land surface with high soil moisture [47-48]. All Tb data used in this study were quality controlled by the corresponding RFI flags.

Based on the assembled ascending and descending datasets, AMSR2 SM retrieval models are developed through XGB, ANN, RRT, MLR, RFT and GBR respectively, which were optimized separately before training. Based on the training data from 2019 to 2021, the daily AMSR2 SM data were developed by combining the ascending and descending retrievals (Figure 1). The daily SMAP combined SM observations over the 2019-2021 period and the 2016-2018 period were used to evaluate model performances during the training and testing time period, respectively. The year-based cross-validation method ensures that the established machine learning models are evaluated with the reference data from different time period [21]. A model with a good "memory" can be easily backward implemented to generate AMSR2 soil moisture in earlier time period. In addition, an independent validation against SCAN measurements was also conducted to further evaluate the performance of each model. Considering the speed and efficiency, the study area in this paper is focused on the region domain from -130°E, 20°N to -60°E, 60°N.

Finally, six kinds of AMSR2 SM datasets are generated by the corresponding machine learning models, including XGB, ANN, RRT, MLR, RFT and GBR. The same input variables, reference data and training strategy could highlight the model differences in terms of the agreements with the reference SMAP data during the training and testing period and the validations with in situ observations.

IV. RESULTS

4.1 Evaluation with SMAP Data

The root-mean-square difference (RMSD) is an extensively used metric to evaluate the predicted soil moisture retrievals versus the reference data [21, 49]. Based on the daily SMAP observations, Figure 2 shows the temporal RMSD distributions for the daily AMSR2 soil moisture retrievals based on each machine learning model over the 2016-2021 time period. Areas shading in red color indicate that model has a modest performance with higher RMSD values, while those in blue color highlight that the machine learning models are successful with respect to the reference data. Relatively, greater RMSD values (>0.1 m^3/m^3) are found for the RFT not only in the eastern densely-vegetated areas, but also in the western sparsely-vegetated areas. Compared to the RFT, the MLR, ANN and RRT show little improvements with the significant SMAP-based differences primarily distributed in the eastern and northern areas. Both tree-boosting models exhibit much better performance with the higher RMSDs mainly in the northern areas, while the XGB has the most successful behavior in the 6 machine learning methods. Specifically, the study area domain-averaged RMSDs for XGB, ANN, RRT, MLR, RFT and GBR are 0.064 m^3/m^3, 0.093 m^3/m^3, 0.087 m^3/m^3, 0.093 m^3/m^3, 0.097 m^3/m^3, 0.078 m^3/m^3, respectively.

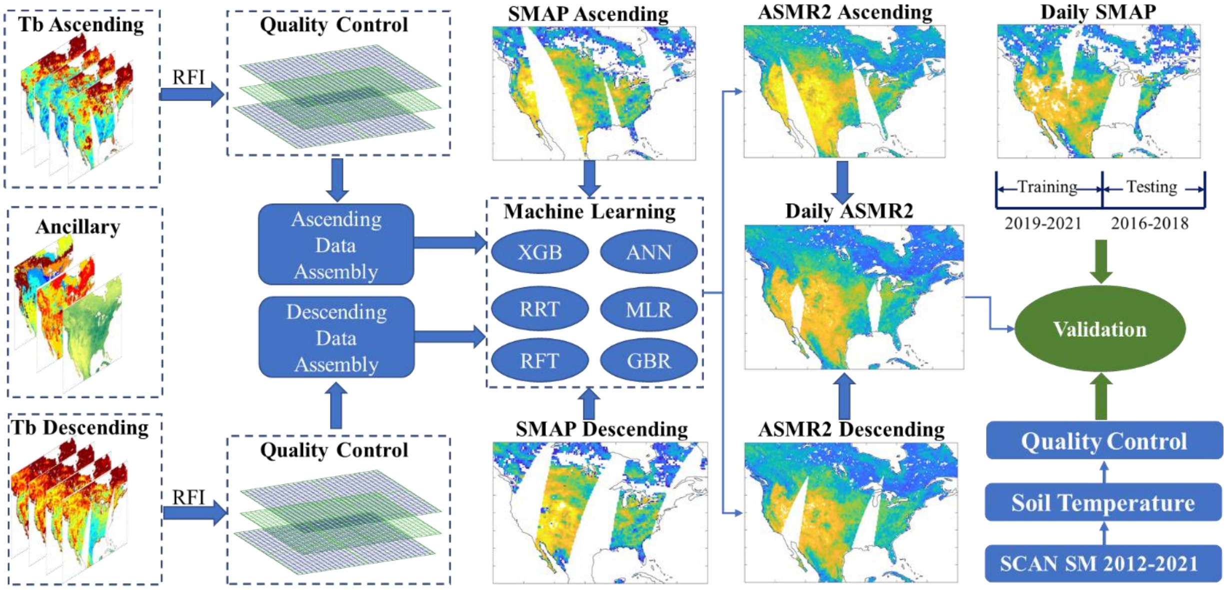

Figure 1. A schematic framework of comparing the commonly-used machine learning models. The abbreviation RFI indicates Radio Frequency Interference, while the abbreviations XGB, ANN, RRT, MLR, RFT and GBR are Extreme Gradient Boosting, Artificial Neural Network, Regression Tree, Multiple Linear Regression, Random Forest and Gradient Boosting machine learning models, respectively.

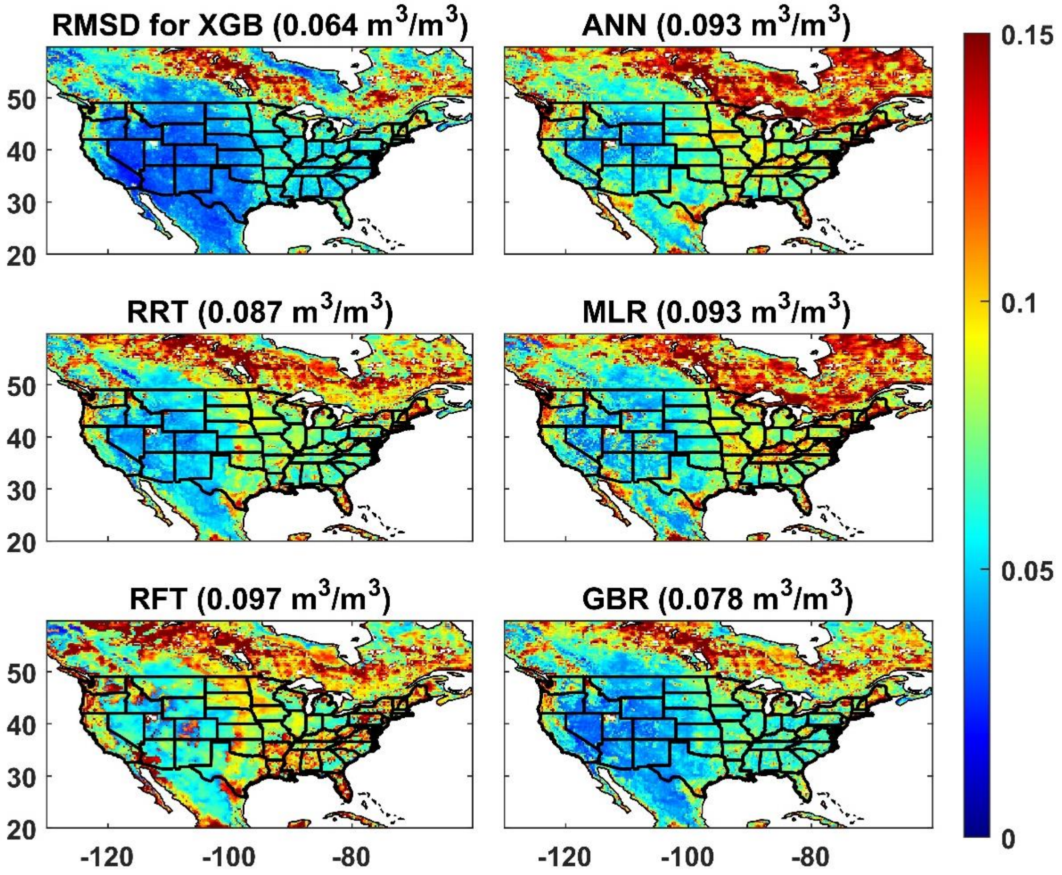

Figure 2. With respect to the daily SMAP soil moisture datasets, the RMSD spatial patterns for the 6 machine learning models over the 2016-2021 time period. The abbreviations XGB, ANN, RRT, MLR, RFT and GBR indicate Extreme Gradient Boosting, Artificial Neural Network, Regression Tree, Multiple Linear Regression, Random Forest and Gradient Boosting machine learning models, respectively.

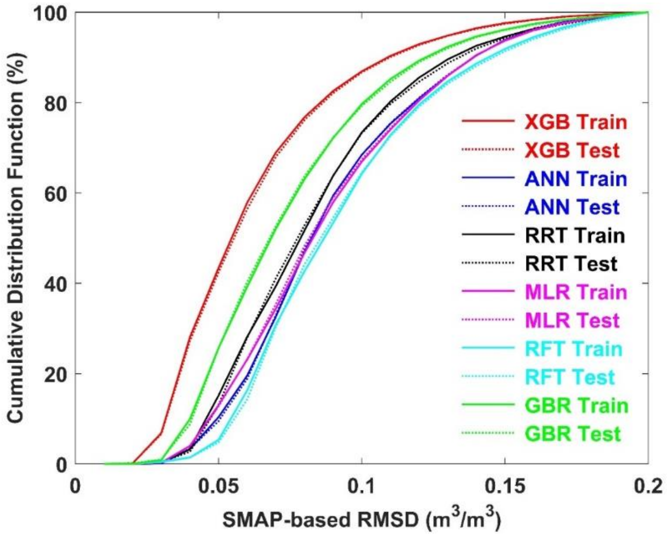

Figure 3. The Cumulative Distribution Functions (CDFs) of the SMAP-based root-mean-square difference (RMSD) values for the 6 machine learning models. CDF for the training time period (2019-2021) is highlighted in solid line, while the dotted line highlights the CDF for the testing time period (2016-2018).

In probability theory and statistics, the Cumulative Distribution Function (CDF) is an important method to describe the variable distribution with a given probability less than or equal to a particular threshold. Based on the daily SMAP data, Figure 3 shows the CDFs of RMSDs over the study area for each of the six machine learning models. The solid and dotted lines in the same color indicate the CDFs for the same approach during the training time period from 2019 to 2021 and the testing time period between 2016 and 2018, respectively. Curves shifting toward left mean better performance in reducing the probability of larger RMSD values, whereas shifting toward right indicate high probability yielding modest behavior. It is consistent with Figure 2 that the XGB model shows the best performance, following by GBR, RRT, ANN, MLR, whereas the RFT has a higher probability to show larger RMSD values. It is worth to note that the dotted lines are basically overlapped with the corresponding solid lines in the same color. This suggests that all of the six models can be implemented to predict the AMSR2 soil moisture beyond the training time period without retraining the new models. This characteristic is very important to demonstrate their feasibility and generalizability for the operational satellite SM retrievals, which can ensure the machine learning-based soil moisture data product to meet the latency requirements of the operational users.

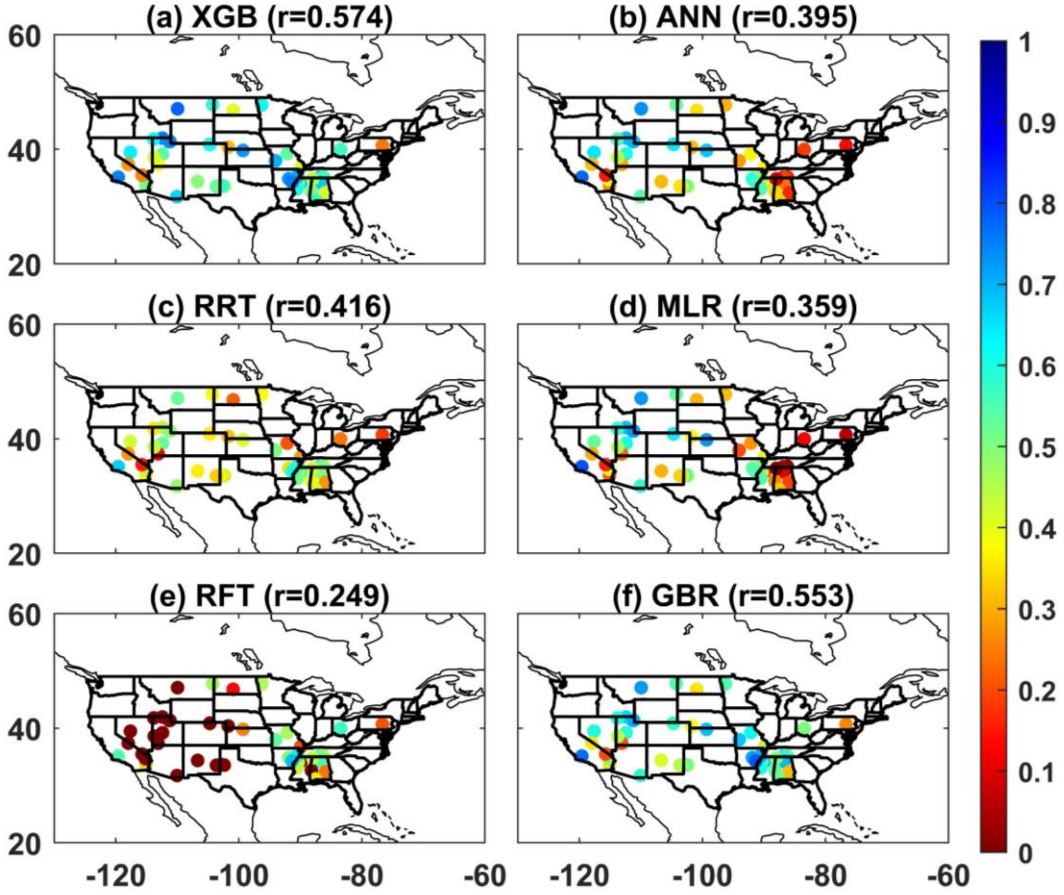

Figure 4. With respect to the quality-controlled daily SCAN soil moisture observations, correlation coefficients for (a) XGB, (b) ANN, (c) RRT, (d) MLR, (e) RFT and (f) GBR over the July 2012-December 2021 period.

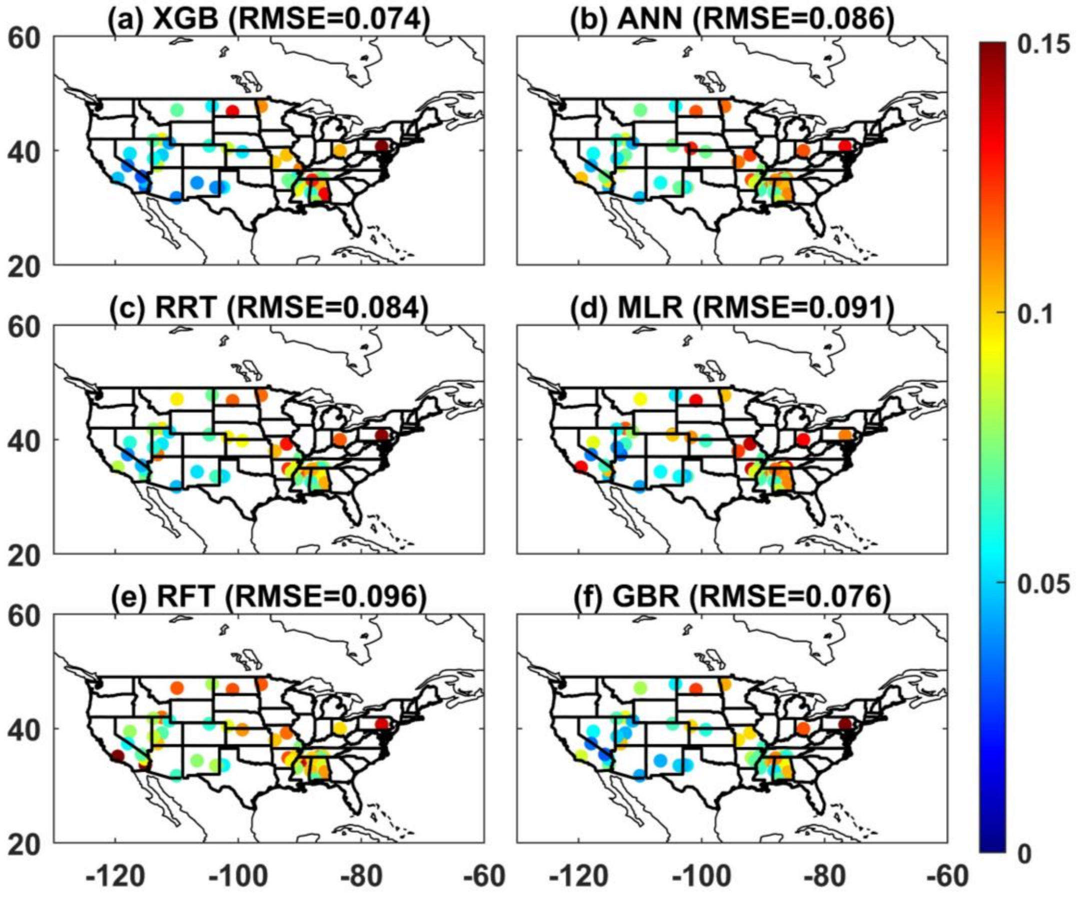

Figure 5. With respect to the quality-controlled daily SCAN soil moisture observations, RMSEs (m^3/m^3) for (a) XGB, (b) ANN, (c) RRT, (d) MLR, (e) RFT and (f) GBR over the July 2012-December 2021 period.

4.2 Validation with In Situ Observations

Selecting performance metrics is typically relying on the nature, variable and characteristics. A single metric can not comprehensively assess all the variable attributes, as each metric is only sensitive to partial features of environmental variables [50]. Based on the quality-controlled SCAN soil moisture measurements, the machine learning approaches were evaluated by three metrics including correlation coefficient (*r*), root mean square error (RMSE) and unbiased RMSE (ubRMSE). The correlation coefficient measures the dynamic trend agreements between AMSR2 SM retrievals and the quality-controlled SCAN measurements from July 2012 to December 2021 (Figure 4). Sites in warm color indicate weak correlations, while in cold color highlight strong positive correlation relationships. Compared to the other five models, the RFT shows the modest performance with the lowest *r* value (0.249) over the CONUS domain (Table 1). This situation can be improved by the MLR (0.359), ANN (0.365) and RRT (0.416), whereas they still present unexceptional behaviors primarily in the eastern areas. Relatively, the GBR performs much better with the CONUS domain-averaged *r* value reaching to 0.553, while the XGB (0.574) yields the strongest agreement with SCAN observations in the 6 cases (Table 1).

The RMSE is a frequently-used metric that measures the differences between model estimates and the observed values. Here, the RMSE represents the prediction errors for each machine learning models, while applying them to retrieve AMSR2 SM data. With respect to the quality-controlled daily SCAN soil moisture observations, Figure 5 shows the temporal RMSEs for each of the six approaches over the July 2012-December 2021 time period. Sites in blue and red colors indicate smaller and greater errors, respectively. In the 6 cases, the RFT yields the highest CONUS domain-averaged RMSE value, reaching to 0.096 m^3/m^3 (Table 1). Unexceptional behaviors are also found for MLR (0.091m^3/m^3), ANN (0.086 m^3/m^3) and RRT (0.084 m^3/m^3) with higher RMSE values in the eastern CONUS areas. This situation can be significantly improved by GBR (0.076 m^3/m^3) and XGB (0.074 m^3/m^3). Compared to the other five models, the XGB more successfully respects to the quality-controlled SCAN SM observations with showing the lowest CONUS domain-averaged RMSE values (Table 1).

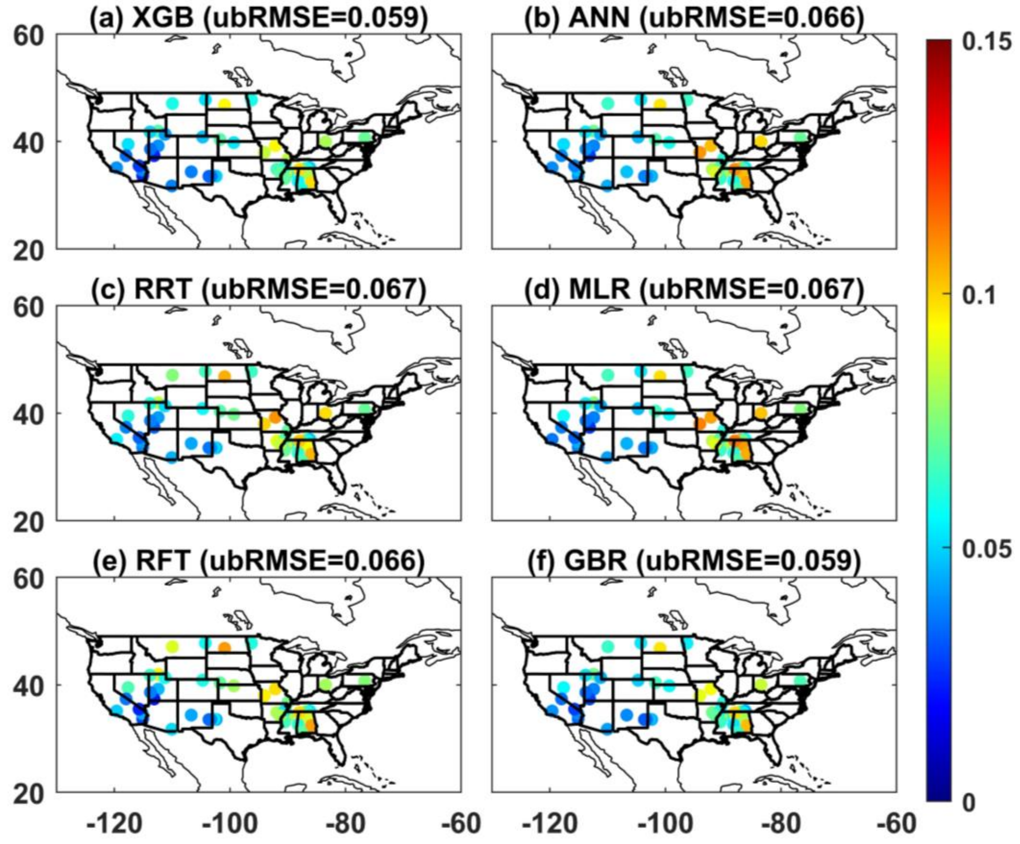

Figure 6. With respect to the quality-controlled daily SCAN soil moisture observations, ubRMSEs (m^3/m^3) for (a) XGB, (b) ANN, (c) RRT, (d) MLR, (e) RFT and (f) GBR over the July 2012-December 2021 period.

In general, satellite SM retrievals have considerable mean and seasonal biases compared to the ground stationary observations. The ubRMSE is a good metric for unbiased evaluations through removing the seasonal and climatological biases. Based on the quality-controlled SCAN SM measurements, further validations on the six machine learning models are thus conducted using the ubRMSE metric (Figure 6). The ANN, RRT, MLR and RFT present similar performance with showing higher ubRMSE values (>0.1 m^3/m^3) in the Mississippi River areas. The CONUS domain-averaged ubRMSEs for those 4 cases are spanning from 0.066 m^3/m^3 to 0.067 m^3/m^3. Relatively, the GBR and XGB models are more successful to retrieve the AMSR2 soil moisture with significantly reducing the unbiased errors. On average, the ubRMSEs for ANN, RRT, MLR and RFT are reduced by 0.07 m^3/m^3 (11.8% reduction) by the GBR and XGB methods (Table 1).

V. DISCUSSION

The goal of this paper is to select an optimal machine learning model to refine the NOAA AMSR2 soil moisture data product. Results in section 4 clearly indicate that the performances of the six commonly-used machine learning models vary significantly from each other. The evaluations are conducted by the reference data SMAP and the quality-controlled SCAN SM measurements. However, the considerable differences between the SMAP and the AMSR2 SM retrievals on the basis of the six approaches are found in the northern areas (Figure 2). Further considerations relevant to interpret the validation results are discussed in this section associated with data availability.

Table 1. With respect to the SCAN SM observations, CONUS domain-averaged correlation coefficients, RMSE (m^3/m^3) and ubRMSE (m^3/m^3) for the six kinds of AMSR2 SM data products over the July 2012-December 2021 period.

AMSR2 SM	r	RMSE(m^3/m^3)	ubRMSE(m^3/m^3)
XGB	0.574	0.074	0.059
ANN	0.395	0.086	0.066
RRT	0.416	0.084	0.067
MLR	0.359	0.091	0.067
RFT	0.249	0.096	0.066
GBR	0.553	0.076	0.059

The NASA SMAP offers land surface SM measurements with near global revisit coverage in 2-3 days [37]. Given the time period from 2016 to 2021, the available day numbers for each SMAP pixel should be greater than 730 days ($365\ days \times 6\ years/3 = 730$). However, there are a number of pixels in the northern study areas filled by missing values, resulting in low data availability that even fewer than 50 days over the 6 years (Figure 7). For instance, compare to Figure 2, the higher RMSD values for the XGB model are basically distributed in the low spatial coverage areas. It is thus expected that the XGB-based AMSR2 can yield much smaller RMSD values in those areas if the SMAP pixels with the available day numbers fewer than 100 are masked out.

The low data availability in sub-regions also affects the training strategy. A straightforward way of establishing machine leaning model is to construct one single model for the entire study domain, which was conducted in this study. Another general training strategy is to build a machine learning model for a grid box (i.e., 100 km by 100 km) and even for each pixel [21, 31]. The low data availability of reference data will increase the uncertainties for the later clustering approaches, and will eventually result in unexcepted results as too few samples. The goal of this first paper of the two-part series is to intercompare the commonly-used machine learning models. Given the same inputs data, training strategy and reference data, the differences among the six cases are good metrics to investigate the model performances. Therefore, the inter-comparison results do not depend on the particular clustering technique or training strategy.

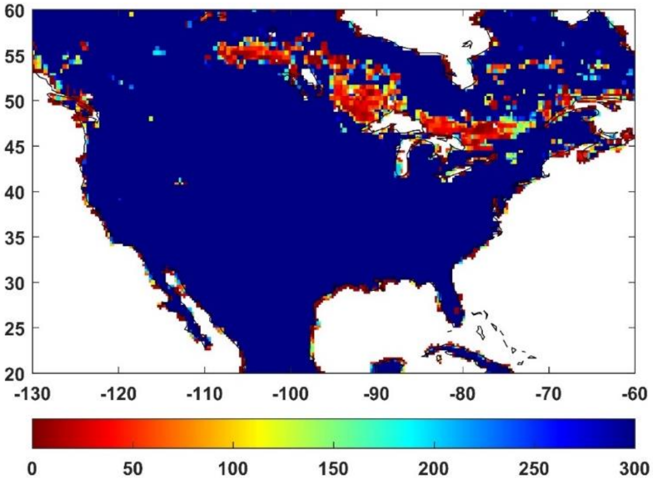

Figure 7. Available day numbers for the daily NASA SMAP soil moisture datasets from 2016 to 2021.

VI. SUMMARY

This study is proposed to refine the currently operational NOAA AMSR2 soil moisture data products. Inter-comparisons of the currently commonly-used machine learning approaches are conducted in this first paper of the two-part series. It offers a solid foundation of selecting an optimal model, which will be eventually used to operationally produce AMSR2 datasets in the NOAA. Results show that all machine learning models can preserve the reference data information during both of the training and testing time periods. This feature ensures to predict past and future satellite retrievals without a new training procedure. The same training technique, input variables and reference data highlight the differences among the six models as good metrics of measuring model behavior. The inter-comparison results thus do not rely on either the training strategy or the clustering technique. Compared to the other five models, the Extreme Gradient Boosting (XGB) shows a more successful performance with respect to the reference data SMAP and the quality-controlled in situ observations. This conclusion has a good implication on implementing the XGB machine learning model to develop the refined AMSR2 datasets in the second paper.

REFERENCES

[1] S.I. Seneviratne, T. Corti, E.L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, A.J. Teuling, 2010. “Investigating soil moisture–climate interactions in a changing climate: a review,” *Earth Sci. Rev.*, vol. 99, no.3, pp. 125–161, 2016

[2] D. Entekhabi, E. G. Njoku, P. E. O'Neill, K. H. Kellogg, W. T. Crow, W. N. Edelstein, et al., “The Soil Moisture Active Passive (SMAP) mission,” *Proceedings of the IEEE*, 98(5), vol. 98, no.5, pp. 704–716, 2010

[3] S. Schwinning, O. E. Sala, M. E. Loik, J. R. Ehleringer, “Thresholds, memory, and seasonality: Understanding pulse dynamics in arid/semi-arid ecosystems,” *Oecologia*, vol. 141, no.2, pp. 191–193, 2004

[4] V. E. Turcu, S. B. Jones, D. Or, “Continuous soil carbon dioxide and oxygen measurements and estimation of gradient-based gaseous flux,” *Vadose Zone Journal*, vol. 4, pp. 1161–1169, 2005

[5] J. Yin, X. Zhan, J. Liu, H. Moradkhani, L. Fang, J. P. Walker, “Near-real-time one-kilometre Soil Moisture Active Passive soil moisture data product,” *Hydrological Processes*, pp.1–14, 2020, DOI: 10.1002/hyp.13857

[6] R. D. Koster, and co-authors. Regions of strong coupling between soil moisture and precipitation. Science, vol. 305, pp. 1138–1140

[7] M. A. Krawchuk, M. A. Moritz, “Constraints on global fire activity vary across a resource gradient,” *Ecology*, vol. 91, no.1, pp. 121–132, 2011

[8] M. A. Moritz, M.-A. Parisien, E. Batllori, M. A. Krawchuk, J. Van Dorn, D. J. Ganz, K. Hayhoe, “Climate change and disruptions to global fire activity,” *Ecosphere*, vol. 3, no.6, pp. 1–22, 2012

[9] E. S. Krueger, T. E. Ochsner, S. M. Quiring, D. M. Engle, J. D. Carlson, D. Twidwell, S. D. Fuhlendorf, “Measured soil moisture is a better predictor of large growing-season wildfires than the Keetch–Byram drought index,” *Soil Science Society of America Journal* vol. 81, pp. 490–502, 2017

[10] J. Wang, E. Engman, T. Mo, T. Schmugge, J. Shiue, “The Effects of Soil Moisture, Surface Roughness, and Vegetation on L-Band Emission and Backscatter,” *IEEE Trans. Geosci. Remote Sens.*, vol. GE-25, pp. 825–833, 1987

[11] T. J. Jackson, T. J. Schmugge, “Passive microwave remote sensing system for soil moisture: Some supporting research,” *IEEE Trans. Geosci. Remote Sens.*, vol. 27, pp. 225–235, 1989

[12] W. Wagner, S. Hahn, R. Kidd, T. Melzer, Z. Bartalis, S. Hasenauer, et al, "The ASCAT soil moisture product: A review of its specifications, validation results, and merging applications," *Meteorologische Zeitschrift*, vol. 22, no.1, pp. 5–33, 2013

[13] L. Karthikeyan, M. Pan, N. Wanders, D. N. Kumar, "Four decades of microwave satellite soil moisture observations: Part 1. A review of retrieval algorithms," *Advances in Water Resources*, vol. 109. pp. 106–120, 2017

[14] M. Drusch, E. F. Wood, T. J. Jackson, "Vegetative and atmospheric corrections for the soil moisture retrieval from passive microwave remote sensing data: Results from the Southern Great Plains Hydrology Experiment 1997," *J. Hydrometeorol.*, vol. 2, pp. 181–192, 2001

[15] M. Owe, R. de Jeu, T. Holmes, "Multisensor historical climatology of satel- lite-derived global land surface moisture," *J. Geophys. Res. Earth Surf*., VOL. 113 F01002, 2008, doi:10.1029/2007JF000769

[16] M. Owe, R. de Jeu, J. Walker, "A methodology for surface soil moisture and vegetation optical depth retrieval using the microwave polarization difference index," *IEEE Trans. Geosci. Remote Sens*., vol. 39, pp. 1643–1654, 2001

[17] T. J. Jackson, D. Chen, M. Cosh, F. Li, M. Anderson, C. Walthall, P. Doriaswamy, E.R. Hunt, "Vegetation water content mapping using Landsat data derived normalized difference water index for corn and soybeans," *Remote Sens. Environ*., vol. 92, pp. 475–482, 2004

[18] X. Li, J-P. Wigneron, L. Fan, et al., "A new SMAP soil moisture and vegetation optical depth product (SMAP-IB): Algorithm, assessment and inter-comparison," *Remote Sens. Environ*., vol. 271, no. 112921, 2022, https://doi.org/10.1016/j.rse.2022.112921

[19] Q. Zhang, Q. Yuan, J. Li, et al., "Generating seamless global daily AMSR2 soil moisture (SGD-SM) long-term products for the years 2013–2019," *Earth Syst. Sci. Data*, vol. 13, pp. 1385–1401, 2021

[20] L. Karthikeyan, A. K. Mishra, "Multi-layer high-resolution soil moisture estimation using machine learning over the United States," *Remote Sens. Environ*., vol. 266, no. 112706, 2021, https://doi.org/10.1016/j.rse.2021.112706

[21] F. Lei, V. Senyurek, M. Kurum, "Quasi-global machine learning-based soil moisture estimates at high spatio-temporal scales using CYGNSS and SMAP observations," *Remote Sens. Environ*., vol. 276, no. 113041, 2022, https://doi.org/10.1016/j.rse.2022.113041

[22] J. R. Eagleman, W. C. Lin, "Remote sensing of soil moisture by a 21-cm passive radiometer," *J. Geophys. Res*., vol. 81, pp. 3660–3666, 1976

[23] T. J. Jackson, D. M. Le Vine, A. Y. Hsu, A. Oldak, P. J. Starks, C. T. Swift, J. D. Isham, M. Haken, "Soil moisture mapping at regional scales using microwave radiometry: The southern great plains hydrology experiment," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, pp. 2136–2151, 1999

[24] E. Pekel, "Estimation of soil moisture using decision tree regression," *Theoretical and Applied Climatology*, vol. 139, pp. 1111–1119, 2020

[25] V. Senyurek, F. Lei, D. Boyd, M. Kurum, A. C. Gurbuz, R. Moorhead, "Machine learning-based CYGNSS soil moisture estimates over ISMN sites in CONUS," *Remote Sens*., vol. 12, no. 7, pp. 1168, 2020

[26] V. Senyurek, F. Lei, D. Boyd, A. C. Gurbuz, M. Kurum, R. Moorhead, R., "Evaluations of a machine learning-based CYGNSS soil moisture estimates against SMAP observations," Remote Sens., vol. 12, no. 21, pp. 3503, 2020

[27] T. Chen, C. Guestrin, C., "Xgboost: A scalable tree boosting system," In: Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016

[28] Y. Liu, X. Xia, L. Yao, W. Jing, C. Zhou, W. Huang, et al., "Downscaling satellite retrieved soil moisture using regression tree-based machine learning algorithms over Southwest France," *Earth and Space Science*, vol. 7, pp. e2020EA001267, 2020 https://doi.org/10.1029/2020EA001267

[29] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015

[30] C. Shen, "A transdisciplinary review of deep learning research and its relevance for water resources scientists," *Water Resources Research*, vol. 54, pp. 8558–8593, 2018

[31] P. Yao, H. Lu, J. Shi, T. Zhao, K. Yang, M. H. Cosh, D. J. Short Gianotti, D. Entekhabi, "A long term global daily soil moisture dataset derived from AMSR-E and AMSR2 (2002–2019)," *Scientific Data*, vol. 8, pp. 143, 2021 https://doi.org/10.1038/s41597-021-00925-8

[32] T. Maeda, Y. Taniguchi, K. Imaoka, "GCOM-W1 AMSR2 level 1R product: Dataset of brightness temperature modified using the antenna pattern matching technique," *IEEE Trans. Geosci. Remote Sens*., vol. 54, pp. 770–782, 2016

[33] J. Yin, X. Zhan, J. Liu, "NOAA Satellite Soil Moisture Operational Product System (SMOPS) Version 3.0 Generates Higher Accuracy Blended Satellite Soil Moisture," *Remote Sens*., vol. 12, pp. 2861, 2020, doi:10.3390/rs12172861

[34] J. Yin, X. Zhan, J. Liu, M. Schull, "An intercomparison of Noah model skills with benefits of assimilating SMOPS blended and individual soil moisture retrievals," *Water Resour. Res*., vol. 55, pp. 2572–2592, 2019

[35] J. Yin, X. Zhan, Y. Zheng, J. Liu, L. Fang, C. R. Hain, "Enhancing Model Skill by Assimilating SMOPS Blended Soil Moisture Product into Noah Land Surface Model," *J. Hydrometeorol*., vol. 16, pp. 917–931, 2015

[36] Q. Liu, C. Cao, C. Grassotti, Y.-K. Lee, "How Can Microwave Observations at 23.8 GHz Help in Acquiring Water Vapor in the Atmosphere over Land?", *Remote Sens*., vol. 13, pp. 489, 2021, https://doi.org/10.3390/rs13030489

[37] D. Entekhabi, E. G. Njoku, P. E. O'Neill, K. H. Kellogg, W. T. Crow, W. N. Edelstein, et al., "The Soil Moisture Active Passive (SMAP) mission," *Proceedings of the IEEE*, vol. 98, no. 5, pp. 704–716, 2010

[38] S. K. Chan, R. Bindlish, P. E. O'Neill, E. Njoku, T. Jackson, A. Colliander, F. Chen, M. Burgin, S. Dunbar, J. Piepmeier, S. Yueh, "Assessment of the SMAP passive soil moisture product," IEEE Trans. Geosci. Remote Sens., vol. 54, no. 8, pp. 4994–5007, 2016

[39] R. Bindlish, M. H. Cosh, T. J. Jackson, et al., "GCOM-W AMSR2 Soil Moisture Product Validation Using Core Validation Sites," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, pp. 1-11, 2017, DOI: 10.1109/JSTARS.2017.2754293

[40] C. A. Reynolds, T. J. Jackson, W. J. Rawls, "Estimating soil water-holding capacities by linking the Food Agriculture Organization soil map of the world with global pedon databases, continuous pedotransfer functions," *Water Resour. Res*., vol 36, pp. 3653-3662, 2000

[41] R. Zhang, C. Huang, X. Zhan, H. Jin, X Song, "Development of S-NPP VIIRS global surface type classification map using support vector machines," *International Journal of Digital Earth*, vol. 11, no. 2, pp. 212-232, 2018, DOI:10.1080/17538947.2017.1315462.

[42] G. L. Schaefer, M. H. Cosh, T. J. Jackson, "The USDA Natural Resources Conservation Service Soil Climate Analysis Network (SCAN)," *Journal of Atmospheric and Oceanic Technology*, vol. 24, no.12, pp. 2073–2077, 2007

[43] P. E. Utgoff, N.C. Berkman, J. A. Clouse. Decision Tree Induction Based on Efficient Tree Restructuring. Machine Learning, vol. 29, pp. 5–44, 1997

[44] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001

[45] J. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001

[46] K. Hsu, H. V. Gupta, S. Sorooshian, "Artificial Neural Network Modeling of the Rainfall-Runoff Process," *Water Resour. Res*., vol. 31, pp. 2517–2530, 1995

[47] T. R. H. Holmes, R. A. M. De Jeu, M. Owe, and A. J. Dolman, "Land surface temperature from Ka band (37 GHz) passive microwave observations," *J. Geophys. Res*., vol. 114, pp. D04113, 2009, doi:10.1029/2008JD010257

[48] D. J. Cecil, T. Chronis, "Polarization-Corrected Temperatures for 10-, 19-, 37-, and 89-GHz Passive Microwave Frequencies," *J Appl Meteorol Climatol*., vol. 57, no. 10, pp., 2249–2265, 2018

[49] P. Abbaszadeh, H. Moradkhani, X. Zhan, Downscaling SMAP radiometer soil moisture over the CONUS using an ensemble learning method. *Water Resources Research*, vol. 55, pp. 324–344, 2019

[50] D. Entekhabi, R. H. Reichle, R. D. Koster, W. T. Crow, "Performance Metrics for Soil Moisture Retrievals and Application Requirements," *J. Hydrometeorol*. vol. 11, pp. 832–840, 2010