Refinement of NOAA AMSR-2 Soil Moisture Data Product—Part 2: Development With the Optimal Machine Learning Model

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

***Abstract*— Advanced Microwave Scanning Radiometer-2 (AMSR2) is a successor of AMSR for Earth-Observation System (AMSR-E), while the third generation of AMSR (AMSR3) will be launched in the near future. The AMSR2 soil moisture (SM) product is also an important component of the SM operational products system (SMOPS) datasets that are operationally produced by National Oceanic and Atmospheric Administration (NOAA). The refinement of the NOAA AMSR2 SM data product can not only benefit the past AMSR-E and the upcoming AMSR3 but also improve the SMOPS data quality. In this second article of the two-part series, the extreme gradient boosting (XGB) model was trained using the AMSR2 6.925-, 10.65-, 18.7-, and 36.5-GHz brightness temperature (Tb) measurements in dual polarizations, ancillary maps, and the vegetation index datasets and in turn used to predict the daily global AMSR2 SM retrievals from 2012 to 2021. Validation results show that the refined AMSR2 SM retrievals (AMSRr) show an overwhelming advantage in data accuracy over the currently operational AMSR2 (AMSRc) product. Compared to the AMSRc, the developed AMSRr presents a significant improvement on data availability. Results also indicate that the refined AMSR2 datasets are comparable with the latest version SM active passive (SMAP) SM product. Based on this study, higher quality AMSRr SM data product will be operationally produced in the NOAA and will eventually benefit the NOAA SMOPS blended SM product and its users.**

***Index Terms*— Advanced Microwave Scanning Radiometer-2 (AMSR-2) soil moisture (SM), machine learning, SM operational products system (SMOPS).**

I. Introduction

THE Advanced Microwave Scanning Radiometer-2 (AMSR2) onboard of the Global Change Observation Mission 1st—Water (GCOM-W1) satellite was launched on May 17, 2012. As the successor of AMSR for Earth-Observation System (AMSR-E), AMSR2 is the second-generation satellite-borne microwave radiometers of the Japan Aerospace Exploration Agency (JAXA). Both AMSR2 and AMSR-E have multifrequency microwave receivers to provide 6.925-, 10.65-, 18.7-, 23.8-, 36.5-, and 89-GHZ brightness temperature (Tb) observations, allowing to provide a majority of global water cycle Environmental Data Records (EDRs). Soil moisture (SM) is one of the most important components of global and regional water cycle, as it can not only control the exchanges of water, energy, and carbon between land surface and the atmosphere but also impact global and regional weather, climate, flash flood, and river flow forecasts.

The radiometer receives the land surface emission affected by the physical temperature and the emissivity of the Earth. The microwave emission primarily depends on the soil dielectric constant linking SM and soil emissivity [1], [2]. This theory allows to retrieve SM in a relatively direct manner on the basis of microwave satellite Tb observations. There are currently three AMSR2 SM retrieval methods well known in the passive microwave SM community [3], including the land parameter retrieval model (LPRM), the JAXA SM algorithm (JAXA), and the single-channel algorithm (SCA). Specifically, LPRM uses both vertical- (V-pol) and horizontal-polarized (H-pol) Tb observations to estimate surface SM status with the assumptions of: 1) the same temperature for canopy and soil and 2) the equal vegetation transmissivity and surface albedo for H-pol and V-pol observations [4], [5]. JAXA AMSR2 SM retrieval algorithm uses the index of soil wetness and the polar index with assuming that the surface temperature is a constant 293 K [6], [7], [8].

Compared to the LPRM and JAXA methods, the SCA AMSR2 SM retrievals are more successful to present SM status with respect to in situ measurements [3]. Based on the radiative transfer equation, SCA uses a single radiometer channel along with a bunch of ancillary maps [9]. The SCA retrieval algorithm has been implemented to operationally produce AMSR2 data product in the National Oceanic and Atmospheric Administration (NOAA)-National Environmental Satellite, Data, and Information Service (NESDIS). AMSR2 is an important component of SM operational products system (SMOPS) SM data product that was developed by NESDIS to meet the real-time SM data requirements of NOAA-National Centers for Environmental Prediction (NCEP) users [10], [11], [12], [13]. The SCA-based SM quality is primarily dependent on the nadir vegetation opacity and the vegetation single scattering albedo parameters [3], [9].

Manuscript received 1 April 2023; revised 16 May 2023; accepted 22 May 2023. Date of publication 26 May 2023; date of current version 11 July 2023. This work was supported in part by the NOAA Joint Polar Satellite System (JPSS)-Proving Ground and Risk Reduction (PGRR) Program; in part by the NOAA Infrastructure Investment and Jobs Act (IIJA) Program; in part by the NOAA Climate Program Office (CPO)-Modeling, Analysis, Predictions and Projections (MAPP) Program; and in part by NOAA under Grant NA19NES4320002. *(Corresponding author: Jifu Yin.)*

Jifu Yin, Jicheng Liu, and Ralph R. Ferraro are with the Earth System Science Interdisciplinary Center, Cooperative Institute for Climate and Satellites, University of Maryland, College Park, MD 20740 USA (e-mail: jifu.yin@noaa.gov).

Xiwu Zhan and Huan Meng are 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.

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

Digital Object Identifier 10.1109/TGRS.2023.3280176

In the past years, the following strategies were implemented to improve the operational AMSR2 accuracy in the NOAA-NESDIS: 1) implementation of LPRM-based vegetation opacity retrievals to make the algorithm completely independent of short-wave satellite vegetation index observations; 2) refinement of the land cover-based model parameters for SCA to improve satellite SM retrieval quality; and 3) update cumulative distribution function (cdf) database building on longer term SM estimations to improve both retrieval quality and spatial coverage. However, our recent study indicates that the current operational AMSR-2 data product can still not show a successful performance expected [12], [14].

Distinct from the currently operational AMSR2 (AMSRc) SM retrievals on the basis of the SCA model, we propose to refine the NOAA AMSR2 data quality using an optimal machine learning method. The refined AMSR2 takes advantage of two advances in producing SM data product. The first advance is the "model-free" characteristics that can better meet the requirements of our operational users. In the SCA method, the polarized Tb needs to be converted to emissivity using land surface temperature [9]. In order to seamlessly and conveniently provide data to the operational weather forecast users within the 6-h cutoff time, the land surface temperature simulations from numerical weather prediction (NWP) models are generally used in the AMSR2 retrieval procedure. This disadvantage could be avoided by machine learning approach. Another advance is to reduce the uncertainties caused by building lookup tables. In the SCA AMSR2 retrieval model, the vegetation parameters for calculating vegetation opacity and vegetation water content are assigned as constants based on lookup tables derived from small watershed experiments [9]. AMSR2 SM retrievals could thus contain large uncertainties at the continental domain let alone the global scale, which could also be addressed by the machine learning method.

In the first article of the two-part series, the intercomparisons of the commonly used machine learning models were conducted, and the extreme gradient boosting (XGB) method shows a more successful performance than the other approaches. Based on the XGB machine learning model, this article is to focus on the refinement of the AMSR2 SM retrievals. Datasets used in this study will be introduced in Section II. XGB model training and the relevant evaluation strategies will be described in Section III. The results focused on the assessment of the developed machine learning model and intercomparison of SM quality between the currently operational (AMSRc) and the refined (AMSRr) AMSR2 retrievals will be provided in Section IV. The discussion and future work are given in Section V. Brief summaries are finally introduced in Section VI.

II. Datasets

A. AMSR2 Brightness Temperature

The AMSR2 onboard the GCOM-W1 satellite offers multifrequency Tb observations in either vertical (V-pol) or horizontal (H-pol) polarization, including 6.925, 7.3, 10.65, 18.7, 23.8, and 36.5 GHz with the corresponding footprint resolutions at 35 × 62, 34 × 58, 24 × 42, 14 × 22, 15 × 26, and 7 × 12 km, respectively [15]. The 7.3-GHz channel was specifically designed to mitigate radio frequency interference (RFI). Except the 7.3-GHz Tb observations, the AMSR2 frequency set is identical to that of AMSR for EOS (AMSR-E). In this study, the 7.3-GHz data were thus excluded to enable that the trained machine learning model can be implemented to retrieve AMSR-E SM. Considering that the Tb observations at 23.8 GHz are relevant to precipitation retrievals [16], they were also excluded to offer the independent SM information without the signals used from satellite precipitation retrievals. Finally, the ascending and descending AMSR2 6.925-, 10.65-, 18.7-, and 36.5-GHz Tb measurements in dual polarizations from July 3, 2012 to December 31, 2021 are used in this article. The 6.925- and 10.65-GHz observations are typically used to retrieve satellite SM in a direct manner, while the 18.7- and 36.5-GHz Tb measurements are used to characterize the water surface and land surface temperature conditions, respectively. Based on the nearest neighborhood method, the original footprint AMSR2 Tb observations were resampled as 25-km spatial resolution over the global domain.

B. SMAP SM Observations

The SM active passive (SMAP) was designed to sense surface SM status through incorporating the observations from an L-band radar and an L-band radiometer [17]. At L-band, the Tb emission originates the top 5-cm soil layer, while the measurements are sensitive to SM status up to 5-kg/m^2 water content of vegetation areas. After the malfunction of the radar instrument, the National Aeronautics and Space Administration (NASA) SMAP can still provide L-band Tb measurements over the global domain, which allows to operationally generate the high-quality SM data products [18]. Based on the dual channel algorithm, the most recent version (V8.0) of Level-3 SMAP SM data showed a better performance than the previous versions using the traditional SCA method [18]. The SMAP was successfully launched on January 31, 2015 and began to offer the global SM science data on April 1, 2015. In this article, the SMAP V8.0 SM observations from April 1, 2015 to December 31, 2021 were regridded to 25-km spatial resolution using the nearest neighborhood approach and then used to train the machine learning model. SMAP V8.0 data are accessible from National Snow and Ice Data Center (https://nsidc.org/data/spl3smp/versions/8).

C. Ancillary Data

The ancillary data used in this article include the 1-km Food and Agriculture Organization soil texture map, the 500-m normalized difference vegetation index (NDVI) of the moderate resolution imaging spectroradiometer (MODIS), and the 1-km annual Visible Infrared Imaging Radiometer Suite (VIIRS) land cover type maps. Specifically, the 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/) is used to characterize the soil retention capacity for water and the spatial variations of soil type.

Both land cover type and MODIS NDVI data are used to represent the underlying vegetation conditions. Following the International Geosphere-Biosphere Program (IGBP) classification scheme, the annual VIIRS global land cover maps at 1-km spatial resolution were developed by the land surface type team of NOAA-NESIDS. It provides 17 surface type classes with respect to the IGBP classification scheme. The eight-daily MODIS NDVI datasets were developed by compositing 16-daily MODIS NDVI

observations at 500-m resolution from either Aqua or Terra. VIIRS land cover maps are accessible from the NOAA Comprehensive Large Array-data Stewardship System (https://www.avl.class.noaa.gov/saa/products/welcome), while MODIS NDVI data are obtained from the NASA EARTHDATA (https://www.earthdata.nasa.gov/).

D. NASMD SM Observations

The North American SM Database (NASMD) is a high-quality observational SM database [19]. It was developed to provide quality-controlled and harmonized ground SM measurements for scientists and decision makers. In situ observations from 33 networks and two shorter term SM campaigns have been integrated into the NASMD. Considering sensor types vary significantly from many different in situ observational networks, all SM observations representing the top 5-cm soil layer have been unified by resampling to daily resolution and converting to volumetric soil water content [19]. A robust quality control (QC) procedure has been implemented to assess the data quality based on multiple, complementary data flagging techniques. All NASMD data were quality controlled by the corresponding QC flags over the 2012–2021 time period, while sites with fewer than 1000 days of observations were excluded. Finally, there are a total of 260 sites in the CONUS domain chosen to validate the refined AMSR2 SM data in this article.

E. Currently Operational NOAA AMSR2 SM

To meet the requirements of the users from NOAA-NCEP, the SMOPS has been developed by the NOAA-NESDIS to provide the global SM observation in near rear time [10], [11], [12], [13]. The AMSR2 SM data product was thus operationally produced and ingested into SMOPS blended datasets to improve its spatial and temporal coverage. Based on the SCA method, the AMSR2 Tb observations from the 6.925-GHz channel are converted to emissivity and in turn to be corrected with consideration of vegetation and surface roughness effect. The AMSR2 SM retrievals are finally obtained by determining the dielectric constant and a dielectric mixing model. In this article, the currently operational 25-km AMSR2 V1.0 SM observations are compared with the refined global AMSR2 data product over the July 3, 2012–December 31, 2021 time period.

F. ESA_CCI Combined SM Dataset

European Space Agency (ESA)-Climate Change Initiative (CCI) combines current and past individual microwave satellite data products to provide the climate SM data records from 1979 in support of climate research [19], [20]. Individual retrievals including SMOS, AMSR2, ASCAT-A, and ASCAT-B were gridded to 25-km spatial resolution and then ingested into the ESA_CCI version 4.5 combined SM product through deriving their weights based on signal-to-noise estimates [20], [21], [22]. The newer version ESA_CCI combined data products have a better accuracy with benefiting from merging SMAP observations. The SMAP was used as the reference data to establish the XGB model in this study and may eventually benefit the refined AMSR2 when comparing it with the AMSRc SM retrievals. We thus used version 4.5 ESA_CCI data without combining SMAP retrievals to objectively compare the AMSRr and AMSRc over the 2012–2019 time period.

III. Methodology

A. XGB Model

XGB model has gained popularity in the machine learning community due to its high speed and good efficiency. The XGB ensembles a set of regression trees (RRTs) using a decision tree. Based on the input variables with m features and n samples $x_i (i = 1, 2, \ldots, n, x_i \in \mathbb{R}^m)$ and the reference data y_i, the predicted values are [23], [24]

$$y_i' = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \Theta \tag{1}$$

where f_k and Θ are the kth RRT and the space of all possible RRT, while K is the number of RRTs that have independent tree structure. Given the differentiable convex loss function $E(\cdot)$, the following objective function (T_g) can be minimized:

$$T_g = \sum_{i=1}^{n} E\left(y_i, y_i'\right) + \sum_{k=1}^{K} \Psi(f_k). \tag{2}$$

To avoid overfitting, the regularization function $\Psi(f_k)$ is used to control the complexity of the function

$$\Psi(f) = \gamma T + \frac{1}{2}\lambda \sum_{t=1}^{T} w_t^2 \tag{3}$$

where T and w_t represent the number of leaves in a decision tree and the score associated with the tth leaf, respectively. Each tree's complexity is controlled by regularization parameters γ and λ. To optimize the ensemble model, an additive training procedure and an iterative process are implemented in the XGB model

$$T_g^p = \sum_{i=1}^{n} E\left(y_i, y_i'^{p-1} + f^p(x_i)\right) + \Psi(f^p) \tag{4}$$

where i and p indicate the instance and iteration, respectively. In this way, the resultant objective function in XGB model is minimized to obtain the optimal leaf weights and in turn to improve tree structure quality. In this study, the reference data are the SMAP V8.0 SM data, while the input variables include AMSR2 Tb observations and ancillary datasets.

B. Machine Learning Framework

The ancillary datasets used to construct the XGB machine learning model include soil texture and land cover maps. The soil texture highlights wilting points and saturated SM features, while the land cover is associated with vegetation optical depth (VOD) and vegetation water contend (VOC) characteristics. Satellite NDVI observations were used as the third ancillary dataset to represent the dynamical vegetation changes. Fig. 1 shows the overall framework of developing the global AMSRr data products. Specifically, our strategy was straightforwardly to train an ascending model and a descending model separately, using the AMSR2 Tb observations for the corresponding set of passes and referring the matched SMAP SM retrievals over the April 1, 2015–December 31, 2021 period. The H-pol and V-pol AMSR2 6.925-, 10.65-,

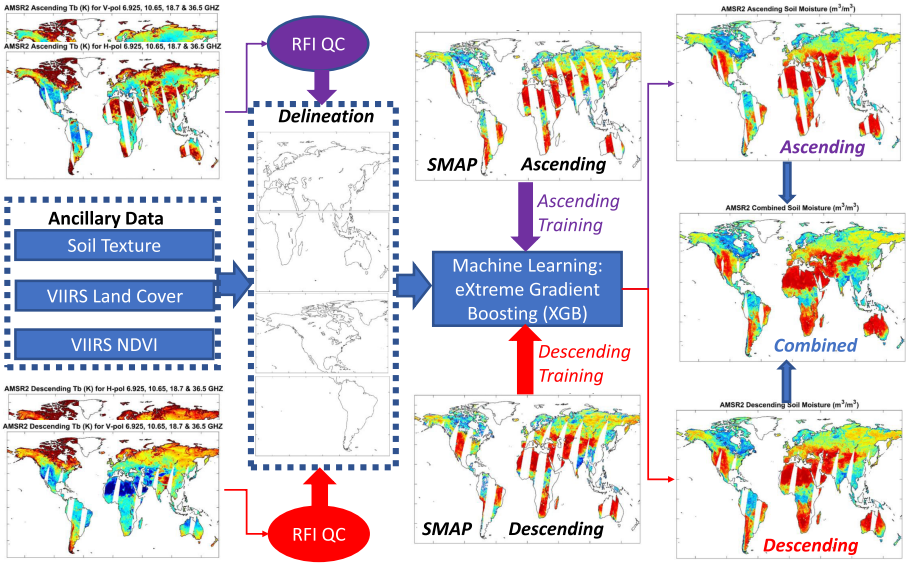


Fig. 1. Schematic framework of comparing the commonly used machine learning models. The abbreviation RFI indicates radio frequency interference, while the abbreviations XGB, ANN, RRT, MLR, RFT, and GBR are extreme gradient boosting, artificial neural network, regression tree, multiple linear regression, random forest and gradient boosting machine learning models, respectively.

18.7-, and 36.5-GHz Tb measurements were quality controlled by the corresponding RFI flags before model training and application.

Instead of training a single model for the global domain, the XGB models were constructed for four subregion domains that basically cover the North America from −180°E, 15°N to −20°E, 90°N, the South America −180°E, −90°N to −20°E, 15°N, the Eurasia from −20°E, 15°N to 180°E, 90°N, and the Africa and Australia from −20°E, −90°N to 180°E, 15°N. This clustering technique takes advantage of three advances in producing the refined AMSR2 SM datasets. The first advance is to characterize the major climate categories with considering the continental geolocations and in turn to reduce model uncertainties. Another advance is the feasibility of operational applications. Additional criterion may benefit the machine learning model, whereas it also increases the computational cost for implementing the trained model. Four major climate zone categories can ensure seamlessly and conveniently provide the XGB-based AMSRr data to the operational weather forecast users within the 6-h cutoff time. The third advance is to stably produce the global AMSRr datasets over the global domain. The performance of the constructed model depends on the data quality and the sample size used in the training procedure. QC of the input and reference data can occasionally reduce the sample size at a small area, which may make the system instable [25].

C. Evaluation Metrics

The refined AMSR2 SM retrievals are evaluated by the quality-controlled ground measurements from the NASMD network within the CONUS domain over the 2012–2021 time period. Considering that the in situ observations are limited coverage and inconsistency at global scale, the supplementary assessments are conducted with the ESA_CCI SM data products over the global domain. The metrics used in this article include Pearson's correlation coefficient (r), root-mean-square error (RMSE)/root-mean-square difference (RMSD), and unbiased RMSE (ubRMSE). Specifically, the correlation coefficient measures the dynamic trend consistency between NASMD and satellite SM observations, while the RMSE and ubRMSE provide the measurements of their differences with and without biases, respectively. Given the ground θ_g and satellite θ_s observations, the evaluation metrics for each pixel (j, i) are given as follows:

$$r(j,i) = \frac{\sum_{k=1}^{N}[\theta_g(j,i)-\overline{\theta_g(j,i)}][\theta_s(j,i)-\overline{\theta_s(j,i)}]}{\sqrt{[\theta_g(j,i)-\overline{\theta_g(j,i)}]^2}\sqrt{[\theta_s(j,i)-\overline{\theta_s(j,i)}]^2}} \tag{5}$$

$$\text{RMSE} = \sqrt{\sum_{k=1}^{N}[\theta_g(j,i)-\theta_s(j,i)]^2/N-1} \tag{6}$$

$$\text{ubRMSE} = \sqrt{\sum_{k=1}^{N}\left\{\left[\theta_g(j,i)-\overline{\theta_g(j,i)}\right]-\left[\theta_s(j,i)-\overline{\theta_s(j,i)}\right]\right\}^2/N-1} \tag{7}$$

where the sample size N indicates the available day numbers of the specific satellite SM data products. In this article, the abovementioned three metrics were computed separately for each NASMD site. Based on the quantitative validations, intercomparisons of the AMSRr and the AMSRc SM retrievals are conducted to highlight the improvements and degradations of the developed AMSRr datasets.

IV. Results

A. Evaluations With Reference Data

Based on the XGB machine learning model, the goal of this study is to develop the high-quality AMSR2 SM data

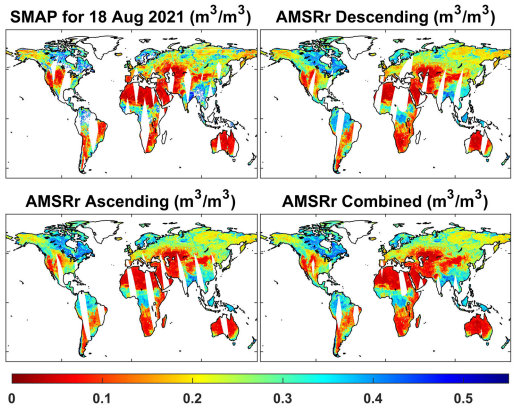


Fig. 2. Daily examples of SMAP combined and the refined AMSR2 (AMSRr) ascending, descending and combined soil moisture retrievals for August 18, 2021.

products. The refined AMSR2 data are first evaluated by comparing with the reference data SMAP. Fig. 2 shows the globally spatial patterns of daily SMAP and the descending, ascending, and the combined AMSRr SM retrievals. Daily examples of the descending and ascending AMSRr present similar spatial variabilities in respect of wetness and aridity levels (Fig. 2). As a result, the combined AMSRr exhibits a reasonable spatial pattern without swath boundaries of the satellite pass-set overlaps (Fig. 2). Similar to the SMAP, the AMSRr SM datasets exhibit wet patterns in Amazon, the Central Africa, South Asia, and the North Forest areas, while smaller AMSRr values can be found in the semi-arid and arid areas, such as Sahara Desert, Australia, Arabian Peninsula, the western United States, and the Central Eurasia. Given AMSR2 has a wider swath (∼1450 km) than SMAP (∼1000 km), the ASMRr presents a better daily spatial coverage in comparison with the SMAP (Fig. 2).

It is of interest to understand whether the XGB model is successfully trained to well respect to the reference data during the training period from 2015 to 2021. The developed AMSRr is estimated by the daily SMAP SM retrievals in Fig. 3. The greater sample density area in warm color is closer to the back 1:1 line, while the lower sample density area shading in the cold color departs from the perfect regression curve. This indicates that the developed AMSRr is consistent with the daily SMAP, including dynamic changes and climatological patterns. As a result, the regression lines for both ascending and descending AMSRr SM datasets are overlapping with the perfect-matching 1:1 curve. The global domain-averaged correlation coefficients between the daily AMSRr and the daily SMAP can reach to 0.883 and 0.884 for descending and ascending pass sets, respectively. The robust agreements suggest that the XGB model has been successfully trained with an expected performance during the training time period.

Given Fig. 3 is only focused on the training time period, however, it is still unknown whether the XGB model here can propagate the SMAP information reasonably beyond the training period. Time-series estimations on SMAP and AMSRr are thus conducted to bridge this gap. In Fig. 4, the time period for the AMSRr is from July 3, 2012 to December 31, 2021, while the SMAP starts from April 1, 2015. The zonal areas from 25°N to 45°N are selected to represent the regions, where

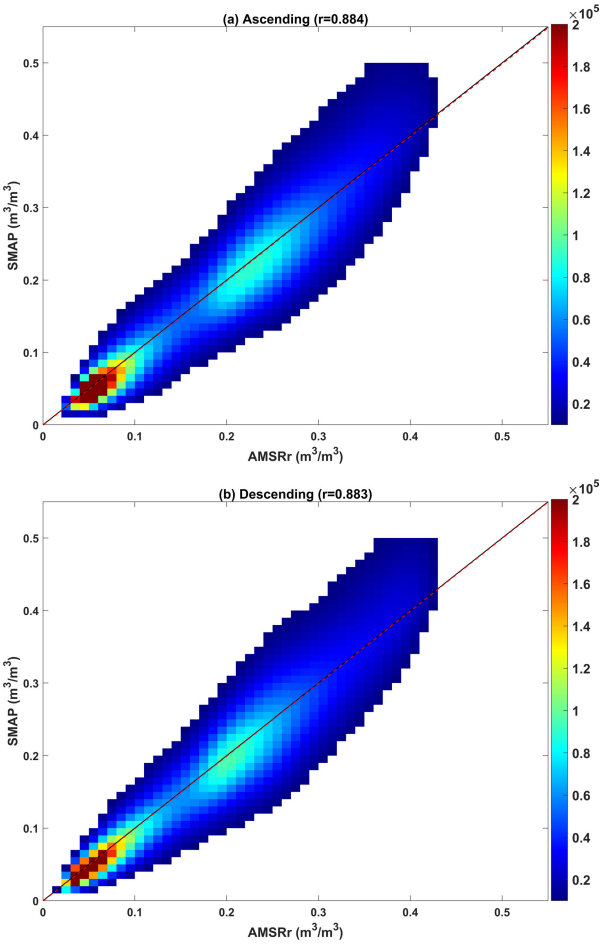


Fig. 3. Daily AMSRr versus the daily SMAP soil moisture data product over the global domain from January 1, 2016 to December 31, 2021. (a) Ascending and (b) descending pass sets. The black 1:1 line indicates that AMSRr perfectly matches with SMAP, while the red dashed curve represents the linear regression line. The color bar indicates sample density.

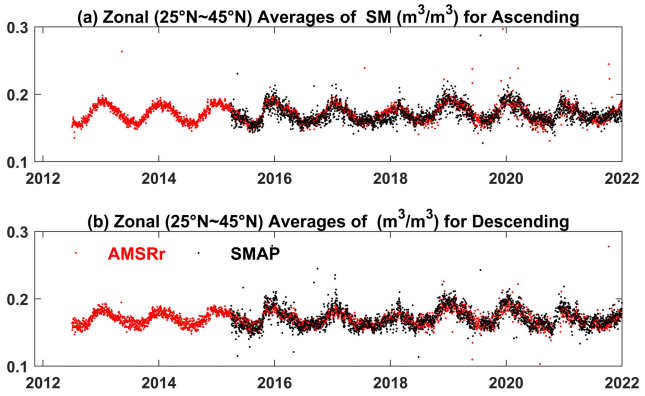


Fig. 4. Daily soil moisture time series of the zonal (25°N–45°N) averages for (a) ascending and (b) descending AMSRr and SMAP datasets. The AMSRr and SMAP data are from July 3, 2012 and April 1, 2015 to December 31, 2021, respectively.

SMAP has a better performance [26]. It can be found that the dynamic trends, seasonal changes, and climatological patterns for the ascending and descending AMSRr SM retrievals are strongly consistent with that for SMAP during the training period, while the AMSRr shows consistent patterns during the

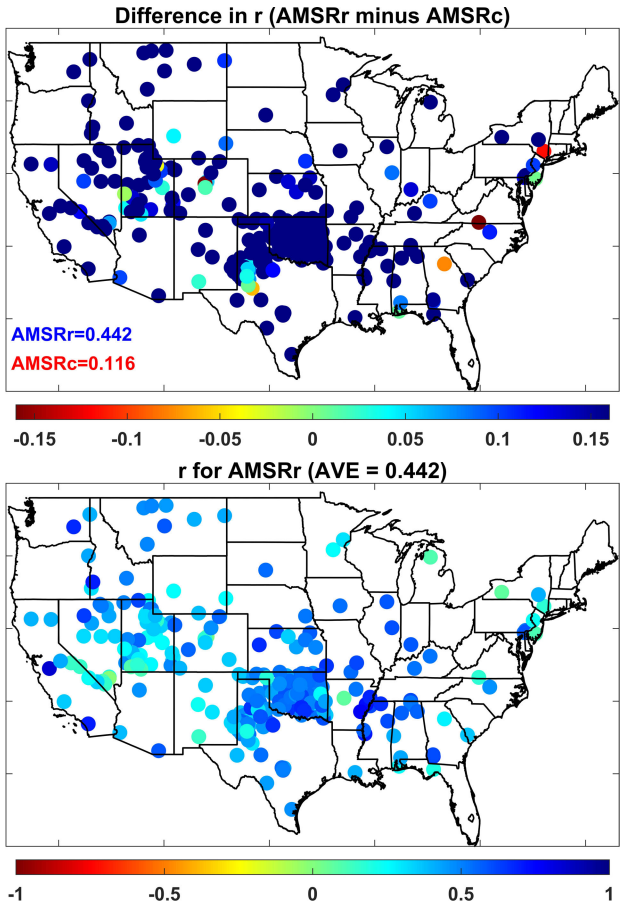


Fig. 5. Validations with the quality-controlled NASMD soil moisture observations over the July 3, 2012–December 31, 2021 time period: (top) r for AMSRr and (bottom) the r differences between AMSRr and AMSRc.

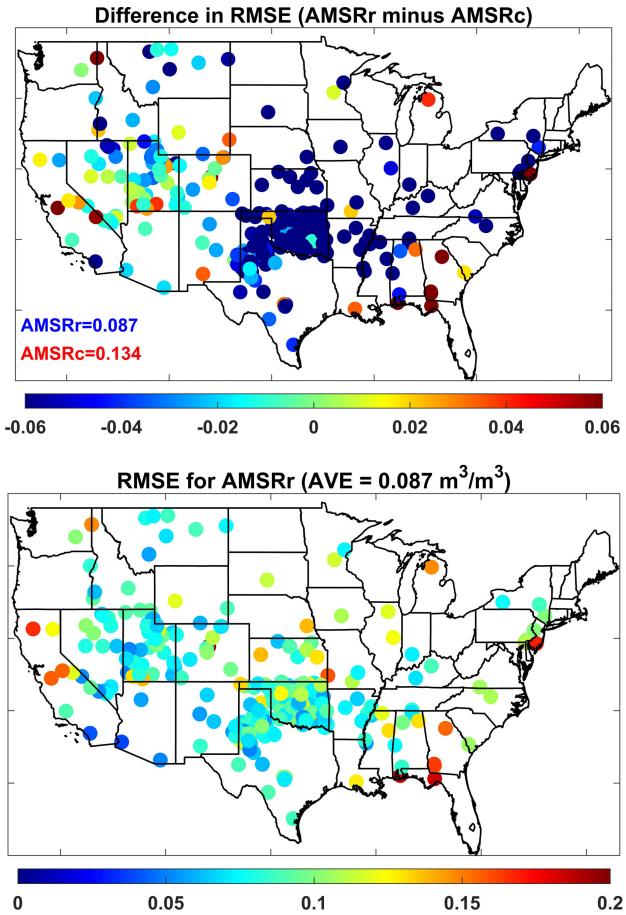


Fig. 6. Validations with the quality-controlled NASMD soil moisture observations over the July 3, 2012–December 31, 2021 time period: (top) RMSE (unit: m^3/m^3) for AMSRr and (bottom) RMSE differences (unit: m^3/m^3) between AMSRr and AMSRc.

time period before April 1, 2015, when SMAP is unavailable. It thus expected that the trained XGB model can successfully retrieve AMSRr SM when the AMSR2 Tb data are accessible. This feature can enable the trained XGB model in this study to predict past and future AMSR2 SM retrievals without a new training procedure.

B. Validations With In Situ Observations

With respect to the quality-controlled NASMD observations, Fig. 5 shows the spatial correlation coefficient (r) distributions of AMSRr SM datasets over the 2012–2021 period. In general, the developed AMSRr presents a good performance on the CONUS domain with r values basically greater than 0.5, whereas few sites with low correlation coefficients (<0.5) primarily scatter in the western mountain areas and the eastern densely vegetated areas. Fig. 5 also exhibits differences in correlation coefficients between the refined and the AMSRc data products. Site in blue color indicates improvement with benefits of the XGB model-based AMSRr, whereas in red color means degradation. Compared to the AMSRc, the AMSRr shows an overwhelming advantage in improving the dynamical trend agreements with the NASMD SM measurements. The CONUS-domain averaged correlation coefficient for AMSRc is 0.110, which can be significantly increased to 0.442 by the refined AMSR2 SM retrievals (Table I).

TABLE I

WITH RESPECT TO THE NASMD SM OBSERVATIONS, CONUS DOMAIN-AVERAGED CORRELATION COEFFICIENTS, RMSE (m^3/m^3) AND UBRMSE (m^3/m^3) FOR AMSRR AND AMSRC SM DATA PRODUCTS OVER THE JULY 2012–DECEMBER 2021 PERIOD

AMSR2 SM	r	RMSE(m^3/m^3)	ubRMSE(m^3/m^3)
AMSRc	0.442	0.134	0.074
AMSRr	0.116	0.087	0.065

Besides of correlation coefficient, we also validated the XGB model-based AMSRr SM retrievals using the RMSE metric over the July 3, 2012–December 31, 2021 time period (Fig. 6). It shows that AMSRr is successful to track surface SM status on the CONUS domain. The AMSRr exhibits a good performance at the most NASMD stations except few of them with larger RMSE values (>0.1 m^3/m^3) mainly distributing in the western and the southeastern areas. Compared to AMSRc, AMSRr significantly improves SM data quality in the western mountain areas, the Great Plains, and the eastern densely vegetated areas with yielding smaller RMSE values. However, the degradations can be found in the western and the eastern CONUS. The CONUS domain-averaged RMSE value for AMSRc is 0.134 m^3/m^3, which can be tremendously reduced

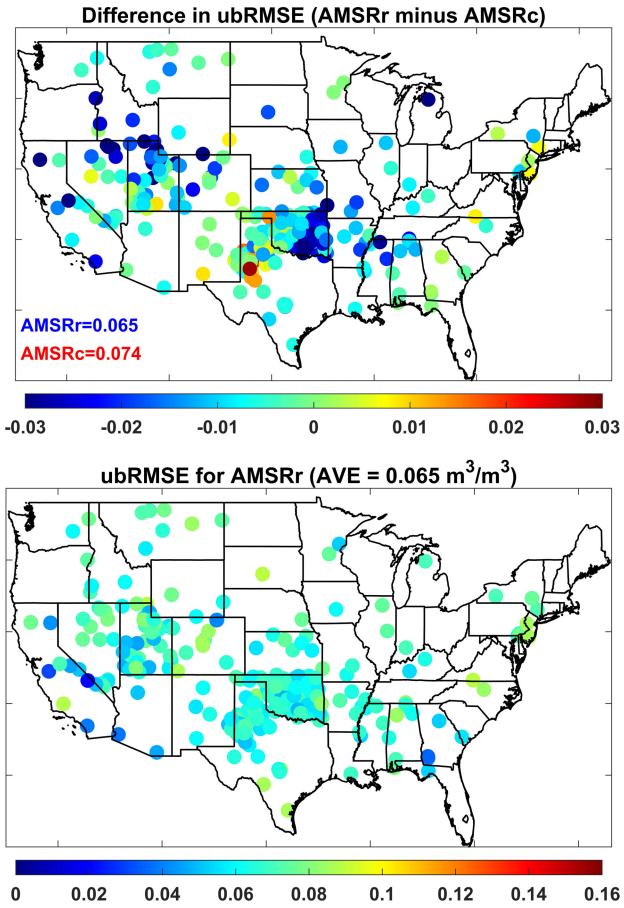


Fig. 7. Validations with the quality-controlled NASMD soil moisture observations over the July 3, 2012–December 31, 2021 time period: (left) ubRMSE (unit: m^3/m^3) for AMSRr and (right) the RMSE differences (unit: m^3/m^3) between AMSRr and AMSRc.

by 0.047 m^3/m^3 (54.02% reduction versus AMSRc) by the developed AMSRr SM retrievals (Table I).

The ubRMSE capturing time-random errors is used as the third metric to evaluate the developed AMSRr, which generally represents the target accuracy level of satellite SM data products [3], [17]. The accuracy requirement of JAXA AMSR2 SM data product is 0.10 m^3/m^3 [3], which can be met by both the currently operational and the refined AMSR2 datasets (Fig. 7). Specifically, the AMSRr SM retrievals on the basis of XGB machine learning model show lower ubRMSE values in the Great Plains and the western sparsely vegetated areas. Compared to AMSRc, AMSRr exhibits improvements not only in the western mountain areas but also in the eastern densely vegetated areas. The CONUS domain-averaged ubRMSE value for the AMSRc is 0.074 m^3/m^3, which can be significantly reduced by 0.009 m^3/m^3 (13.85% reduction) by the refined AMSR2 SM retrievals (Table I). It is worth to note that the developed AMSRr can meet our target accuracy of ubRMSE below 0.065 m^3/m^3.

Apart from the commonly used measurement metrics, the developed AMSRr and the operational AMSRc are also further assessed by comparing against the time series of NASMD measurements. Fig. 8 shows the five representative stations located in Mississippi (MS), Arkansas (AR), Montana (MT), Utah (UT), and South Dakota (SD) states of the US, where the AMSRr SM retrievals have similar ubRMSE values with the AMSRc data. As the time series shown in MT, UT, and SD, the AMSRc presents the flat patterns of seasonal dynamics over the 2012–2021 time period. This unreasonable situation can be improved by AMSRr with successfully respecting to the dynamical changes of NASMD observations. AMSRc has a good performance in MS and AR with exhibiting different values in cold and warm seasons, while the developed AMSRr is more successfully to track SM status with much closer to the time series of in situ SM measurements. The modest performance of time series may result in the low correlation coefficients for AMSRc, while this kind of shortcomings can be addressed by the XGB model-based AMSRr.

C. Complementary Evaluation With ESA_CCI Data

In this article, the XGB model was trained and implemented individually at the four subregion domains, resulting in four independent machine learning models for North America, South America, Eurasia, and the Africa-Australia areas, respectively. In Section IV-B, validations with respect to the quality-controlled in situ observations are conducted in the CONUS domain, and we thus set up a complementary evaluation on AMSRr using the ESA_CCI combined SM data products over the global domain. With respect to ESA_CCI datasets, Fig. 9(a) shows the temporal RMSD for the refined AMSR2 SM retrievals on the global domain from 2012 to 2019. In general, AMSRr can successfully respect to the ESA_CCI data product on the global domain except the high RSMD values (>0.1 m^3/m^3) distributed in the eastern Canada and the northern Eurasia areas. Relatively, AMSRc presents greater differences with ESA_CCI mainly in the low-latitude areas, Europe, the eastern USA, and South America [Fig. 9(b)].

Differences in the ESA_CCI-based RMSDs between the AMSRr and the AMSRc are shown in Fig. 9(c). Area shading in blue color indicates that AMSRr is more consistent with ESA_CCI, whereas region shading in red color means AMSRc is better. Compared to the AMSRc, it is obviously found that the developed AMSRr has an overwhelming advantage with significantly reducing the ESA_CCI-based RMSD values. The RMSD cumulative density functions (CDFs) for AMSRr and AMSRc over the global domain also indicate that the developed AMSRr can tremendously decrease the RMSD values [Fig. 9(d)]. Respecting to the daily ESA_CCI combined SM data product, the global domain-averaged RMSD value for AMSRc is 0.133 m^3/m^3, which can be reduced by 0.053 m^3/m^3 (66.25% reduction versus AMSRc) by the refined AMSR2 SM retrievals.

V. Discussion

The results presented in Section IV indicate that the refined AMSR2 SM retrievals show more robust capability to track surface SM conditions in comparison with the AMSRc data product. The statistical results can be greatly affected by sample size that is relative to the level of confidence in the evaluations. Meanwhile, the results in Section IV-A suggest that the developed AMSRr can well match the reference SMAP data; however, the quantitative comparisons between AMSRr and SMAP still deserve us to pay more attention. The challenges and opportunities are thus discussed further associated with data availability and data accuracy.

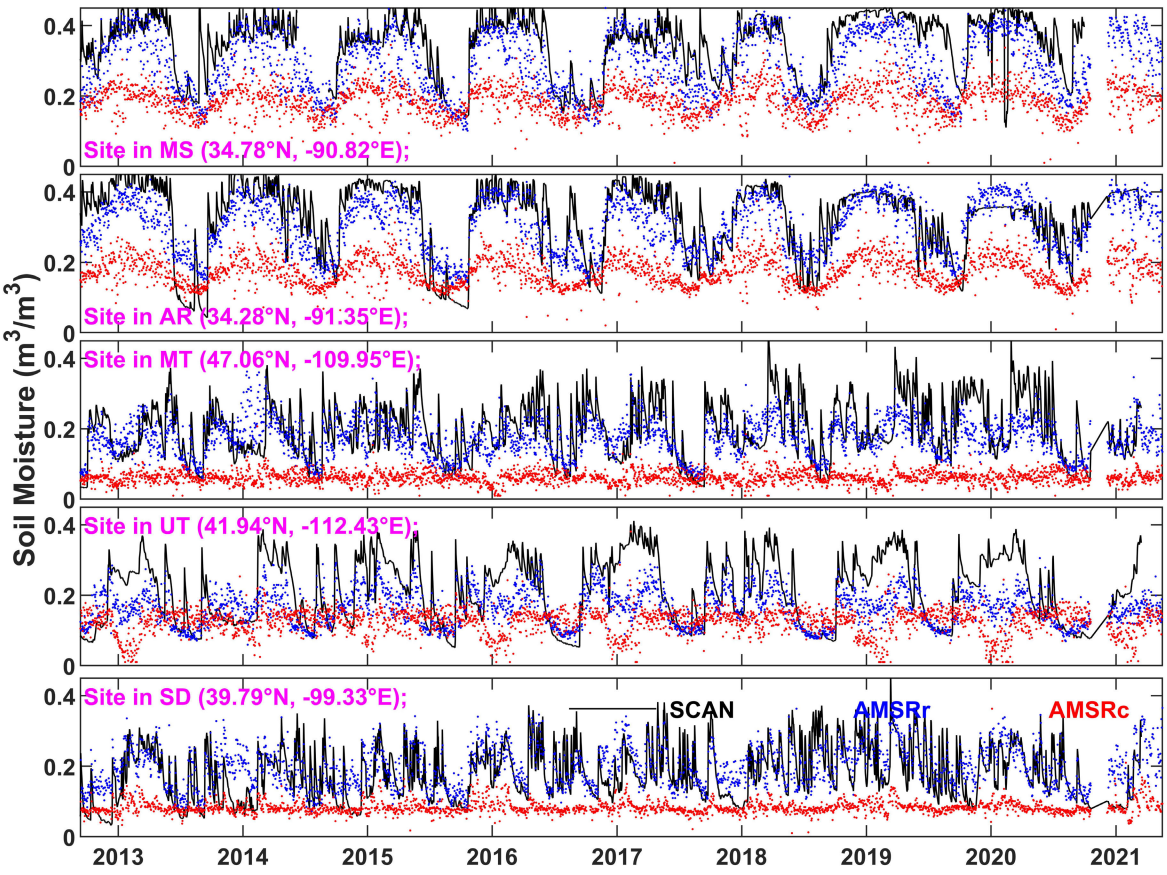


Fig. 8. Soil moisture time series (m^3/m^3) for AMSRr (blue dot), AMSRc (red dot), and NASMD (black line) at five example sites from 2012 to 2021. From top to bottom: sites in Mississippi (MS), Arkansas (AR), Montana (MT), Utah (UT), and South Dakota (SD) state of the US.

A. Spatial Coverage

Sample size indicates the number of observations used in the statistics, which can be reflected by data availability of satellite retrievals. Specifically, the data availability is defined as the fraction of available day number over total day number during the study period for each land grid on the global domain [12]. Fig. 10 exhibits data availability for AMSRc and AMSRr SM data products over the 2012–2021 time period. Suffering from the lookup table retrieval parameters, the AMSRc cannot provide any valuable observations in the areas, including Sahara, central Eurasia, the northern Canada, Russia, and the central and the western Australia, although it has the same spatial coverage with AMSRr in the rest areas [Fig. 10(a)]. Relatively, the AMSRr yields better data availability on a gridded global domain [Fig. 10(b)]. The lower spatial coverage for AMSRr can only be found in the high latitude areas that cannot be covered by the NDVI in the cold season. As a result, the daily global domain-averaged data availability for AMSRr is generally greater than 80% with the maximum value reaching 90% [Fig. 10(c)], which is much higher than that for AMSRr (40% in general). The comprehensive validations are conducted on the CONUS domain in Section IV-B, while the sample sizes for both AMSRr and AMSRc should be the same given the same data availability in Fig. 10(a) and (b). The statistical differences in Figs. 5–7 are thus only dependent on the data quality of AMSRc and AMSRr datasets. Beyond the CONUS areas, AMSRr shows an equivalent and even better spatial coverage than AMSRc, while the AMSRr is more successfully to respect to the ESA_CCI SM observations. This makes it much clearer that the developed AMSRr has a better performance than the AMSRc.

B. Intercomparisons Between AMSRr and SMAP

With respect to the quality-controlled NASMD SM observations, Fig. 11 shows the ubRMSE differences (m^3/m^3) between the developed AMSRr and the reference data SMAP datasets over the April 1, 2015–December 31, 2021 time period. Sites in warm color indicate SMAP is better, while in cold color mean AMSRr is more accurate. Relative to the SMAP, the AMSRr shows smaller ubRMSE values mainly in the central CONUS, whereas the modest performance for AMSRr can be found in the west mountain areas and the eastern densely vegetated areas. The SMAP L-band measurements can sense SM status with penetrating through vegetation of 5-kg/m^2 water content [17], which ensures SMAP to perform better than AMSRr in the eastern CONUS. The AMSR2 Tb observations from C-band, X-band, and Ka-band contend with a smaller penetration depth for SM retrievals. Relatively, the feature of the L-band Tb emission originating from top 5-cm soil layer can benefit the validations on SMAP data product, as the in situ observations from the NASMD networks represent the top 5 cm of soil [19]. This kind of benefit can promote the SMAP performance, but the refined AMSR2 is still comparable with

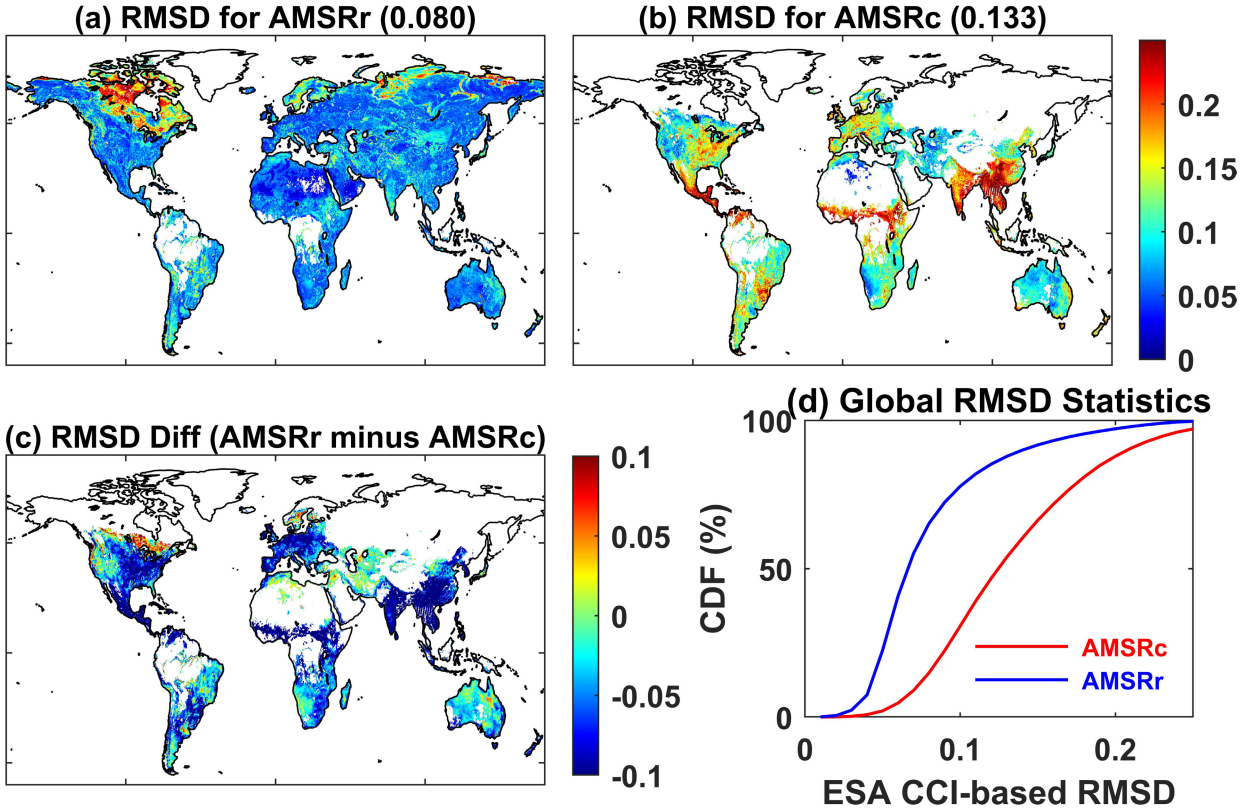


Fig. 9. With respect to the quality-controlled ESA_CCI soil moisture observations, (a) root-mean square difference (RMSD, unit: m^3/m^3) for AMSRr, (b) RMSD for AMSRc, and (c) RMSD differences with areas shading in blue color indicating that the developed AMSRr is more consistent with ESA_CCI, as well as (d) the RMSD CDFs (in unit %) for AMSRr and AMSRc over the global domain during the 2012–2019 time period.

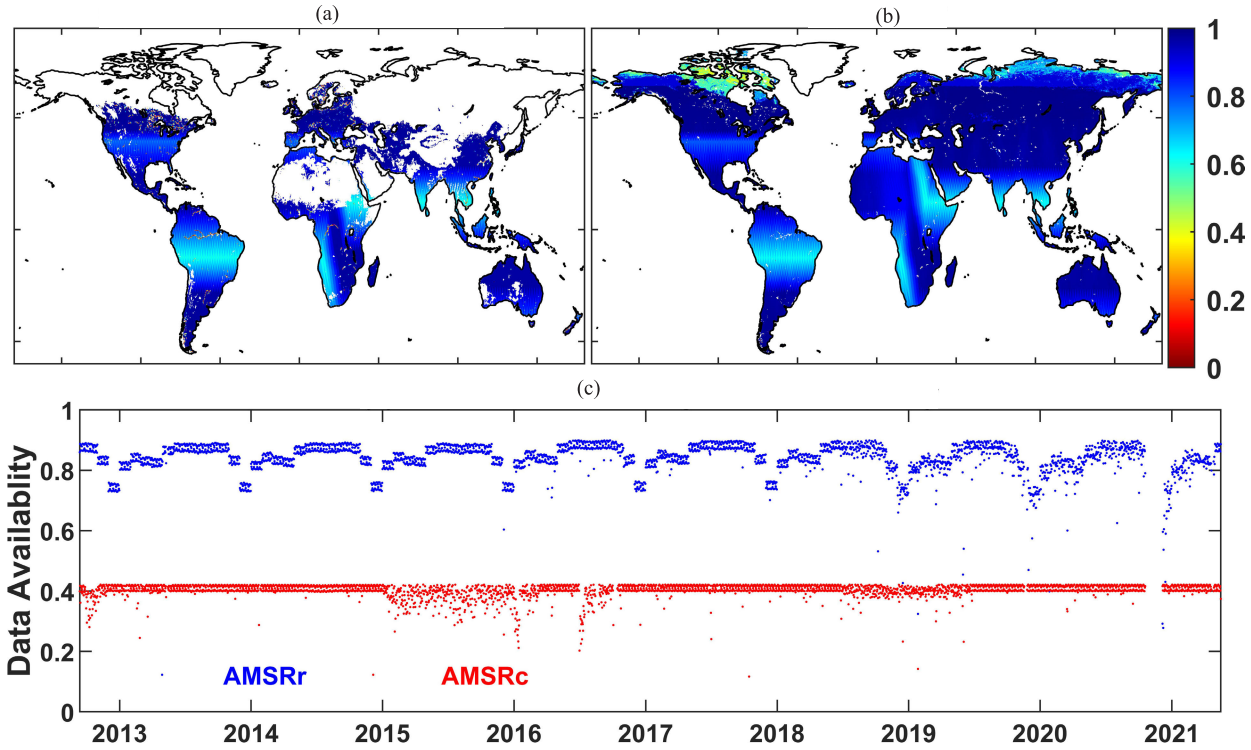


Fig. 10. Data availability (%) for (a) AMSRc and (b) AMSRr, as well as (c) daily global domain-averaged data availability for the both over the 2012–2021 time period.

the reference SMAP product. The CONUS domain-averaged ubRMSE values for AMSRr and SMAP are 0.065 m^3/m^3 and are 0.063 m^3/m^3, respectively. The strong year-to-year consistency results shown in this study suggest that the data accuracy for the refined AMSR2 SM retrievals is qualitatively stable.

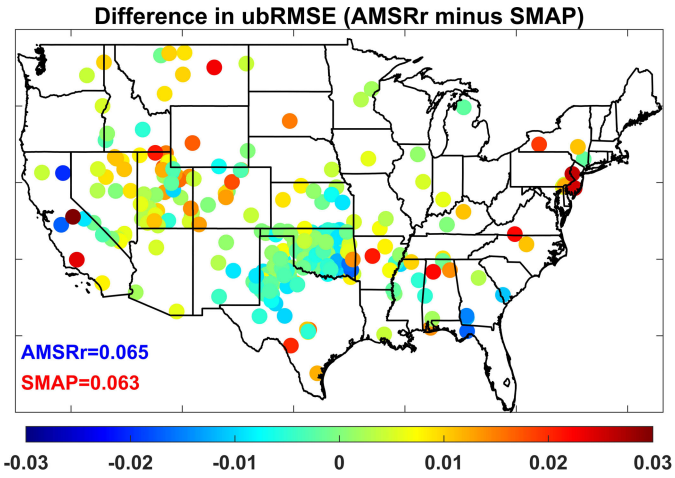


Fig. 11. With respect to the quality-controlled NASMD soil moisture observations, ubRMSE differences (m^3/m^3) between the refined AMSR2 (AMSRr) and the reference data SMAP datasets over the April 1, 2015–December 31, 2021 time period.

C. Future Work

Prior to AMSR2 mission, the AMSR-E has been operated for about ten years starting from June 2002 and stopping operations in October 2011. With experience of the AMSR-E and AMSR2 development and operations, the third generation of AMSR (AMSR3) was targeted to launch within the time window of April 2023–March 2024. The AMSR3 will succeed AMSR series observations, which allows to produce long-term continuous low-frequency Tb data records. Through calibrating the past AMSR-E and the upcoming AMSR3 observations toward the ongoing AMSR2 datasets, it is expected to develop long-term consistent low-frequency Tb measurements in dual-polarization from the three-generation AMSR missions. Benefiting from the XGB machine learning model trained in this article, a longer than two-decade daily AMSR SM dataset over the global domain will be produced in the near future.

Additionally, AMSR2 SM is an important component of the operational SMOPS blended SM data products at the NOAA-NESDIS [10], [11], [12], [13]. To meet the requirements of the National Weather Service users, the latency for 6-h SMOPS is less than 3 h, while the daily SMOPS has latency within 24 h. Based on the XGB model, the refined AMSR2 SM can be produced within 1 h, ensuring it can be integrated into the SMOPS with the 3-h time cutoff limit. The ascending and descending refined AMSR2 SM retrievals can improve the data quality of 6-h SMOPS blended product, while the daily AMSRr can benefit the daily SMOPS. The refined AMSR2 SM will thus be ingested into the SMOPS system to promote its quality in the near future.

VI. Conclusion

The target of this study is to enhance the quality of the current AMSR2 SM data product operationally produced by the NOAA-NESIDIS. The development and comprehensive assessment of the refined AMSR2 retrievals are introduced in this second article of the two-part series. Based on the trained XGB machine learning model, the refined AMSR2 (AMSRr) can well respect to the reference data SMAP spatial and temporal patterns. The developed XGB model is able to reasonably predict AMSRr retrievals when the SMAP is not unavailable. Respecting to the quality controlled in situ observations, the developed AMSRr shows overwhelming advantages over the AMSRc SM data products with significantly increasing correlation coefficient and tremendously reducing RMSE and unRMSE values. Compared to AMSRc, AMSRr is more successfully to respect to the ESA_CCI SM data product over the global domain. Relatively, the AMSRr shows higher spatial coverage than the AMSRc, which allows to operationally offer more available SM observations. Validation results in this article also indicate that the AMSRr is comparable with the latest version SMAP. The better quality of AMSR2 will improve the NOAA SMOPS data product and will eventually benefit our operational users.

Acknowledgment

The authors would like to 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.

References

[1] J. Wang, E. Engman, T. Mo, T. Schmugge, and J. Shiue, "The effects of soil moisture, surface roughness, and vegetation on L-band emission and backscatter," *IEEE Trans. Geosci. Remote Sens.*, vols. GE–25, no. 6, pp. 825–833, Nov. 1987.

[2] T. J. Jackson and T. J. Schmugge, "Passive microwave remote sensing system for soil moisture: Some supporting research," *IEEE Trans. Geosci. Remote Sens.*, vol. 27, no. 2, pp. 225–235, Mar. 1989.

[3] R. Bindlish et al., "GCOM-W AMSR2 soil moisture product validation using core validation sites," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 1, pp. 209–219, Jan. 2018, doi: 10.1109/JSTARS.2017.2754293.

[4] A. G. C. A. Meesters, R. A. M. DeJeu, and M. Owe, "Analytical derivation of the vegetation optical depth from the microwave polarization difference index," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 2, pp. 121–123, Apr. 2005.

[5] M. Owe, R. de Jeu, and 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, no. 8, pp. 1643–1654, 2011.

[6] T. T. Koike et al., "Development of an advanced microwave scanning radiometer (AMSR-E) algorithm of soil moisture and vegetation water content," *Ann. J. Hydraulic Eng., Japanese Soc. Civil Eng.*, vol. 48, pp. 217–222, Feb. 2004.

[7] H. Fujii, T. Koike, and K. Imaoka, "Improvement of the AMSR-E algorithm for soil moisture estimation by introducing a fractional vegetation coverage dataset derived from MODIS data," *J. Remote Sens. Soc. Jpn.*, vol. 29, no. 1, pp. 282–292, 2009.

[8] I. Kaihotsu, J. Asanuma, K. Aida, and D. Oyunbaatar, "Evaluation of the AMSR2 L2 soil moisture product of JAXA on the Mongolian Plateau over seven years (2012–2018)," *Social Netw. Appl. Sci.*, vol. 1, no. 11, p. 1477, Nov. 2019, doi: 10.1007/s42452-019-1488-y.

[9] T. J. Jackson, "III. Measuring surface soil moisture using passive microwave remote sensing," *Hydrological Processes*, vol. 7, no. 2, pp. 139–152, Apr. 1993.

[10] J. Liu et al., "NOaa soil moisture operational product system (SMOPS) and its validations," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Beijing, China, Jul. 2016, pp. 3477–3480.

[11] J. Yin, X. Zhan, and J. Liu, "NOAA satellite soil moisture operational product system (SMOPS) version 3.0 generates higher accuracy blended satellite soil moisture," *Remote Sens.*, vol. 12, no. 17, p. 2861, Sep. 2020, doi: 10.3390/rs12172861.

[12] J. Yin, X. Zhan, J. Liu, and M. Schull, "An intercomparison of Noah model skills with benefits of assimilating SMOPS blended and individual soil moisture retrievals," *Water Resour. Res.*, vol. 55, no. 4, pp. 2572–2592, Apr. 2019.

[13] J. Yin, X. Zhan, Y. Zheng, J. Liu, L. Fang, and C. R. Hain, "Enhancing model skill by assimilating SMOPS blended soil moisture product into Noah land surface model," *J. Hydrometeorology*, vol. 16, no. 2, pp. 917–931, Apr. 2015.

[14] J. Yin, X. Zhan, J. Liu, and R. R. Ferraro, "A new method for generating the SMOPS blended satellite soil moisture data product without relying on a model climatology," *Remote Sens.*, vol. 14, no. 7, p. 1700, 2022, doi: 10.3390/rs14071700.

[15] T. Maeda, Y. Taniguchi, and 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, no. 2, pp. 770–782, Feb. 2016.

[16] Q. Liu, C. Cao, C. Grassotti, and 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, no. 3, p. 489, Jan. 2021, doi: 10.3390/rs13030489.

[17] D. Entekhabi et al., "The soil moisture active passive (SMAP) mission," *Proc. IEEE*, vol. 98, no. 5, pp. 704–716, May 2010.

[18] S. K. Chan et al., "Assessment of the SMAP passive soil moisture product," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4994–5007, Aug. 2016.

[19] S. M. Quiring et al., "The North American soil moisture database: Development and applications," *Bull. Amer. Meteorological Soc.*, vol. 97, no. 8, pp. 1441–1459, Aug. 2016, doi: 10.1175/BAMS-D-13-00263.1.

[20] W. Dorigo et al., "ESA CCI soil moisture for improved Earth system understanding: State-of-the art and future directions," *Remote Sens. Environ.*, vol. 203, pp. 186–215, Dec. 2017.

[21] A. Gruber, T. Scanlon, R. van der Schalie, W. Wagner, and W. Dorigo, "Evolution of the ESA CCI soil moisture climate data records and their underlying merging methodology," *Earth Syst. Sci. Data*, vol. 11, no. 2, pp. 717–739, May 2019, doi: 10.5194/essd-11-717-2019.

[22] W. Preimesberger, T. Scanlon, C. Su, A. Gruber, and W. Dorigo, "Homogenization of structural breaks in the global ESA CCI soil moisture multisatellite climate data record," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 2845–2862, Apr. 2021.

[23] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

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

[25] F. Lei et al., "Quasi-global machine learning-based soil moisture estimates at high spatio-temporal scales using CYGNSS and SMAP observations," *Remote Sens. Environ.*, vol. 276, Jul. 2022, Art. no. 113041, doi: 10.1016/j.rse.2022.113041.

[26] F. Chen et al., "Global-scale evaluation of SMAP, SMOS and ASCAT soil moisture products using triple collocation," *Remote Sens. Environ.*, vol. 214, pp. 1–13, Sep. 2018.

Jifu Yin received the M.S. and Ph.D. degrees in atmospheric physics and atmospheric environment from the Nanjing University of Information Science and Technology (formerly Nanjing Institute of Meteorology), Nanjing, China, in 2011 and 2015, respectively.

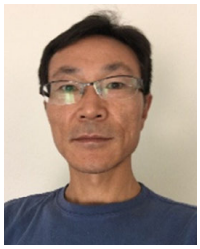
He is currently an Associate Research Scientist with the Earth System Science Interdisciplinary Center (ESSIC) and the Cooperative Institute for Satellite Earth System Studies (CISESS), University of Maryland, College Park, MD, USA. He has been involved with several National Aeronautics and Space Administration (NASA)/National Oceanic and Atmospheric Administration (NOAA) satellite missions. His research interests include satellite soil moisture retrieval and its applications, data assimilation, land surface model, and machine learning model.

Xiwu Zhan received the Ph.D. degree in soil, crops, and atmospheric sciences from Cornell University, Ithaca, NY, USA, in 1995.

He is currently a Physical Scientist with National Oceanic and Atmospheric Administration (NOAA) National Environmental Satellite, Data, and Information Service (NESDIS) Center for Satellite Applications and Research, College Park, MD, USA. His research interests include remote sensing and modeling of land surface soil moisture and vegetation remote sensing for numerical weather, climate and water predictions, and societal applications.

Michael Barlage received the Ph.D. degree in atmospheric, oceanic, and space sciences from the University of Michigan, Ann Arbor, MI, USA, in 2001.

He is currently a Physical Scientist with the National Oceanic and Atmospheric Administration (NOAA)/NWS/National Centers for Environmental Prediction (NCEP) Environmental Modeling Center, College Park, MD, USA. He leads land model and land data assimilation development with the NOAA/NWS/NCEP Environmental Modeling Center. He also conducts applied research on land–atmosphere interactions through parameterization improvement in the Noah-MP and Noah land-surface models. His work incorporates large satellite-based datasets of land properties, such as snow, soil moisture, vegetation index, and albedo.



Jicheng Liu received the Ph.D. degree in geography from Boston University, Boston, MA, USA, in 2005.

He is currently a Visiting Associate Research Scientist with the Earth System Science Interdisciplinary Center (ESSIC), University of Maryland, College Park, MD, USA. His current research interests include generating merged operational land surface soil moisture products using all available data from different satellite sensors and their applications in drought monitoring and assimilations in weather forecasting models.

Huan Meng received the M.S. degree in physical oceanography from Florida State University, Tallahassee, FL, USA, in 1993, and the Ph.D. degree in hydrology from Colorado State University, Fort Collins, CO, USA, in 2004.

She is currently a Physical Scientist with NOAA/NESDIS Center for Satellite Applications and Research. Her research interest include passive microwave remote sensing including algorithm development, calibration and validation, and satellite product climate data record (CDR). Her current research focuses on precipitation and hydrology product development and applications.

Ralph R. Ferraro is currently the Associate Director of the Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA. He has over 40 years of professional experience in the earth science, with an emphasis on remote sensing. During his previous career at National Oceanic and Atmospheric Administration (NOAA)/National Environmental Satellite Data and Information Service (NESDIS), he was the focal point for engagement with National Aeronautics and Space Administration (NASA) on several earth science missions, including EOS, TRMM, and GPM. Precipitation retrieval has been his primary focus with an emphasis on both weather and climate applications. Because of this engagement, NOAA/NESDIS was able to accelerate its use NASA sensor data and products to support NOAA operations.