GIScience & Remote Sensing

ISSN: 1548-1603 (Print) 1943-7226 (Online) Journal homepage: www.tandfonline.com/journals/tgrs20

A stand-alone framework for predicting spatiotemporal errors in satellite-based soil moisture using tree-based models and deep neural networks

Subin Kim, Hyunglok Kim, Yonghwan Kwon & Hoang Hai Nguyen

To cite this article: Subin Kim, Hyunglok Kim, Yonghwan Kwon & Hoang Hai Nguyen (2025) A stand-alone framework for predicting spatiotemporal errors in satellite-based soil moisture using tree-based models and deep neural networks, GIScience & Remote Sensing, 62:1, 2475572, DOI: 10.1080/15481603.2025.2475572

To link to this article: https://doi.org/10.1080/15481603.2025.2475572

GISCIENCE & REMOTE SENSING
2025, VOL. 62, NO. 1, 2475572
https://doi.org/10.1080/15481603.2025.2475572

A stand-alone framework for predicting spatiotemporal errors in satellite-based soil moisture using tree-based models and deep neural networks

Subin Kim [a], Hyunglok Kim [a], Yonghwan Kwon [b] and Hoang Hai Nguyen [a]

[a]Department of Environment and Energy Engineering, Gwangju Institute of Science and Technology (GIST), Gwangju, Republic of Korea; [b]Evaluation Team, Data Assimilation Group, Korea Institute of Atmospheric Prediction Systems (KIAPS), Seoul, Republic of Korea

ABSTRACT

Soil moisture (SM) has been recognized as one of the essential climate variables in Earth science. Incorporating spatiotemporal information from satellite SM observations into land surface models via land data assimilation (LDA) has emerged as a promising method to improve continuous SM modeling and climate extreme monitoring. Since the reliable satellite SM error dynamics are crucial for successful LDA applications but often assumed to remain static over time, this study presents a novel framework that combines triple collocation analysis (TCA) with advanced machine learning techniques, such as Light Gradient Boosting Machine (LGBM) and Deep Neural Networks (DNN), to accurately quantify spatially and temporally continuous satellite-based SM characteristics on a global scale. In this study, the stand-alone TCA-based time-variant SM error prediction models, which rely exclusively on data used for the Soil Moisture Active Passive (SMAP) retrieval, were developed and comprehensively evaluated. These models successfully recover error information over areas where SMAP SM error data were previously unusable due to uncertain error characteristics. In addition, the stand-alone SMAP-based model demonstrates superior performance compared to model that relies on external datasets, such as the Global Land Data Assimilation System (GLDAS). These findings provide valuable insights into the dynamic nature of satellite-based SM error under various environmental conditions and present a novel way to improve LDA. Furthermore, the proposed methodology can be extended to predict error dynamics for other satellite-based geophysical datasets, broadening its potential applications beyond SMAP.

ARTICLE HISTORY
Received 7 September 2024
Accepted 27 February 2025

KEYWORDS
Triple collocation analysis; soil moisture; machine learning in geophysics; microwave satellite system; time-variant error

1. Introduction

The soil moisture (SM) is a key variable for predicting weather and natural disasters, as understanding its accurate dynamics offers valuable insights into land-atmosphere interactions, such as global-scale energy, water resources, and carbon cycles (Seneviratne et al. 2006). Various approaches have been proposed to characterize spatiotemporal SM variation, including observation-based (e.g. in-situ and remote sensing measurements) and simulation-based methods (e.g. land surface models). While in-situ sensors (Kim, Cosh, et al. 2020; Nguyen, Kim, and Choi 2017) can provide accurate SM data and are often used as benchmark for the calibration and validation of satellite- and model-based SM data, their limited point-scale representativeness, labor-intensive construction, and uneven global distribution hinder their effective applications over large-scale areas (Dorigo et al. 2011).

On the contrary, remote sensing technologies have emerged as a useful instrument for periodic, large-scale, and less labor-intensive SM retrieval. Microwave satellite sensors, typically the Soil Moisture and Ocean Salinity (SMOS), Soil Moisture Active Passive (SMAP), Advanced Scatterometer (ASCAT), and the novel bistatic radar-based Cyclone Global Navigation Satellite System (CYGNSS), provide a promising tool for SM monitoring and environmental applications at high temporal resolution due to the high sensitivity of microwave signals to the land surface and subsurface water content (Calvet et al. 2011; Engman and Chauhan 1995; Entekhabi et al. 2010; Kim and Lakshmi 2018; Lakhankar, Krakauer, and Khanbilvardi 2009; Nguyen et al. 2025). However, problems including radio frequency interference, discontinuous observations, poor spatial resolution, and numerous

CONTACT Hyunglok Kim hyunglokkim@gist.ac.kr

Submitted to GIScience & Remote Sensing (Article Collection "Soil moisture for Earth ecosystems: methods, products, and applications").
Supplemental data for this article can be accessed online at https://doi.org/10.1080/15481603.2025.2475572

parameter constraints significantly contribute to the inherent uncertainties of satellite-based techniques more worse (Li et al. 2021; Petropoulos, Ireland, and Barrett 2015). On the other hand, simulation-based methods, such as land surface models (LSMs), which are based on a set of physical and conceptual equations, have become a viable way to provide multiscale and continuous SM data. Despite their potential, LSMs have several limitations, such as imperfect underlying physics assumptions and uncertainties in model calibration and validation. These factors may create discrepancies between simulated and actual SM conditions, hindering their widespread utilization (Fisher and Koven 2020).

While microwave satellite systems are limited in providing continuous SM measurement compared to the LSMs, they offer a unique advantage in capturing extreme events- and human-induced sudden changes, which are often not well represented in LSMs. The complementary relationship between space-borne observations and LSM simulations highlights the need to integrate satellite-derived SM data into LSM through land data assimilation (LDA) techniques to improve SM estimation. Many previous studies have explored LDA to improve LSM and weather prediction model's performance in estimating SM by incorporating SM products from various satellite missions, including the SMOS (Zhao et al. 2014), ASCAT (Aires et al. 2021; Kwon et al. 2024), SMAP (Kwon et al. 2024), and CYGNSS (Kim et al. 2021). These studies demonstrate the improved accuracy of SM estimation through LDA. Nevertheless, these previous studies applied a fixed error variance for satellite systems in LDA, assuming spatially and temporally constant observation errors. This assumption limits the effectiveness of their SM LDA systems, making them suboptimal for capturing the true variability in SM error dynamics. In addition to the importance of understanding accurate error characteristics in satellite-based SM data for the optimal LDA, merging different sources of satellite-based SM datasets also depends on knowing their relative accuracy (Gruber et al. 2016; Kim et al. 2018). However, to the best of our knowledge, no studies have addressed time-variant error characteristics in merging SM data, largely due to the challenge of obtaining continuous time-variant SM error information within each satellite system.

Since the traditional satellite-based products are mainly validated using ground-based measurement networks, their applicability to regional scales is still restricted and requires further reference datasets for improving the evaluation. To address these limitations, this study employed triple collocation analysis (TCA), a widely used method for estimating global-scale error variance in satellite data without reference dataset, to generate time-variant SM error information on a global scale (Kim et al. 2023; Stoffelen 1998; Wu et al. 2021; Yilmaz and Crow 2014). Nevertheless, TCA results exhibit several limitations, particularly when it is challenging to apply assumptions are difficult to implement or when there are insufficient simultaneous SM retrievals (Kim et al. 2023; Yilmaz and Crow 2014). Therefore, when TCA assumptions are violated or simultaneously available triplet data are missing, the computed error variance becomes physically infeasible (Kim, Wigneron, et al. 2020). This lack of error information in satellite-based SM data constrains its comprehensive use in diverse research fields, making it essential to explore alternative strategies and methodologies.

Data-driven machine learning (ML) and deep neural networks (DNN) have recently emerged as potential approaches in hydrology research field (Xu and Liang 2021). Nevertheless, the application of advanced ML and DNN for dynamic SM error prediction remains underexplored, and these approaches often rely on external data products. Therefore, it is necessary to investigate their feasibility and robustness for reconstructing time-variant error information from satellite-based SM observations on a global scale. The major goal of this study is to leverage the combination of the TCA with ML and DNN to improve accurately quantifying spatiotemporal error estimations of global satellite SM data. We selected the SMAP SM dataset – observations from the widely used microwave satellite sensor designed specifically for SM measurement – as the target SM dataset for dynamic error prediction.

The methodologies to reach our primary objectives are organized as follows: First, we calculated the TCA error variance to quantify the uncertainty in SMAP SM data, considering both time-invariant and time-variant error assumptions and demonstrated the significance of understanding time-variant error characteristics. Second, we developed tree-based ML and DNN error prediction models with two different sets of predictors: one for a stand-alone error prediction model relying solely on data from SMAP, and the other incorporating external data sources such as the Global Land Data Assimilation System (GLDAS). We also highlighted the superior performance of the stand-alone error prediction models. Finally, we reconstructed the missing time-variant error information for SMAP, which has not been addressed in previous studies, and demonstrated how our proposed methods effectively overcome the limitations of TCA while providing new insights into the time-varying error characteristics of SMAP SM datasets. Therefore, through accomplishing the major objectives, we investigated the feasibility of developing effective ML- and DNN-based time-variant error prediction models relying exclusively on data used to retrieve SMAP SM data, which does not depend on the support from other products.

2. Datasets

In this study, we aim to investigate the application of tree-based ML and DNN models to predict the time-variant error in SMAP SM data using variables from its own observations (i.e. stand-alone model) or external products (i.e. external-source model). To achieve this, we developed models from different predictors: a stand-alone model that uses only variables from SMAP and an externally sourced model that incorporates external predictors from GLDAS data. Both models were designed to predict the error in the SMAP SM data, as derived from the TCA method.

Note that no quality control (QC) flags were applied to the raw data used in this study for the SMAP and ASCAT SM data. As previously noted by Kim et al. (2023), the QC flags supplied with SMOS, SMAP, and ASCAT SM data might be misleading, perhaps eliminating good-quality data or keeping bad-quality data, hence decreasing their reliability for such applications. Therefore, by using SM data without QC flags, the raw data quality is preserved, enabling us to develop a comprehensive error prediction model without modifying the SM data.

2.1. SMAP soil moisture and additional predictor variables

The SMAP mission, launched in January 2015, is National Aeronautics and Space Administration (NASA)'s first sun-synchronous Earth observation mission designed specifically for global SM measurement purposes, leveraging both L-band passive (radiometer) and active (radar) sensors, with an expected performance of Root Mean Square Error (RMSE) of 0.04 m^3m^{-3} over sparsely to moderately vegetated regions (Entekhabi et al. 2010). However, due to the malfunction of the SMAP radar system, only measurements from the SMAP passive radiometer are currently available for this study. Particularly, the current SMAP radiometer provides measurements at a spatial resolution of 36 km with a revisit time of 2–3 days across two overpasses encompassing the descending (06:00 local time) and ascending (18:00 local time) passes. In this study, the global 36 km SM data from the SMAP Level-3 (SMAP L3) version 9 product for the 2015–2023 period, which adopts the dual channel algorithm and is projected on the Equal-Area Scalable Earth version 2.0 (EASE-2) Grid system, was considered (O'Neill et al. 2023). The data are freely accessible via the NASA National Snow and Ice Data Center Distributed Active Archive Center (NSIDC DAAC).

In the present study, variables in the SMAP L3 product, together with the SMAP ancillary data, were utilized as predictors to develop the SMAP-based model for obtaining the target SMAP time-variant SM error. A list of SMAP L3 and SMAP ancillary variable products used in this study is summarized in Table S1.

2.2. ASCAT soil moisture

The C-band active microwave ASCAT sensor onboard the Meteorological Operational (MetOp) satellite mission, operated by the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), was selected to provide supplementary unbiased satellite SM observations for supporting the TCA error characterization. The MetOp mission consists of three satellites: MetOp-A, launched in October 2006, MetOp-B, launched in 2012, and MetOp-C, launched in 2018. Notably, MetOp-A was recently decommissioned, with its operational duties taken over by MetOp-B and MetOp-C, continuing the global SM monitoring. ASCAT provides measurements at an original spatial resolution of 25–34 km across two overpasses: the descending pass at 09:30 local time and the ascending pass at 21:30 local time. The ASCAT SM data was developed using the Technische Universität Wien (TU Wien) change detection algorithm and are currently monitored by the EUMETSAT Satellite Application Facility on Support to Operational Hydrology and Water Management (H-SAF) (Wagner et al. 2013). From 0% (totally dry soil) to 100% (fully saturated soil), the ASCAT SM value is expressed in the degree of saturation relative to pore volume and represents an estimate of the water content in the top 0–5 cm of the soil layer. Thus, we converted the ASCAT SM's unit into absolute volumetric units (m^3/m^3) by using soil porosity information. This study employed the MetOp ASCAT global SM Climate Data Record (CDR) version 7 (v7) resampled to a 12.5 km grid during the study period of 2015–2023.

2.3. GLDAS soil moisture and additional predictor variables

Along with the remotely sensed observation products, the SM dataset simulated from the model-based GLDAS was also employed in this study to support the error characterization using the TCA. GLDAS is a reanalysis product that was developed by NASA's Goddard Space Flight Center, aiming to provide optimal simulations of land surface states and fluxes by assimilating ground- and satellite-based observational data into offline LSMs such as Noah, Mosaic, Community Land Model, and Variable Infiltration Capacity (Rodell et al. 2004). Here, we employed the global surface soil layer (0–10 cm) dataset, 3-hourly, 0.25° × 0.25° gridded GLDAS Noah Land Surface Model version 2.1 product (GLDAS_NOAH025_3H 2.1) for the period of 2015 to 2023, which is distributed via the NASA Goddard Earth Sciences Data and Information Service Center (GES DISC) (Beaudoing and Rodell 2020).

Moreover, to evaluate the potential of satellite-based dynamic error data prediction from model-based simulations using ML, as a comparison to that from the SMAP observations, several meteorological and hydrological variables provided by the GLDAS were also considered as predictors to construct the GLDAS-based error model. Datasets of the GLDAS variables used in this study were obtained for the same study period and are summarized in Table S2.

2.4. MODIS LULC

Apart from the major SM and other variables used for time-variant SM error data prediction, the Land Use Land Cover (LULC) dataset is required to analyze the proposed method across different land surface conditions. To this end, the commonly used satellite-based LULC product from the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor onboard the Terra and Aqua satellites was employed. Specifically, this study utilized the L3 MODIS Land Cover Climate Modeling Grid (MCD12C1) version 6.1 product, which provides yearly data at a spatial resolution of 0.05° (Friedl and Sulla-Menashe 2015). It is important to note that, as of the time this paper was written, LULC data for 2023 was unavailable. Therefore, only data from the period of 2015–2022 were taken into account; the missing 2023 data was replaced with data from 2022.

3. Methodology

We used Random Forest Regressor (RFR) for the feature selection process and subsequently developed both Light Gradient Boosting Machine (LGBM) and DNN models for the error prediction (Ma et al. 2024). The overall research workflow is depicted in Figure 1, which outlines eight distinct processes. Detailed explanations of each methodology are provided in the following subsections.

3.1. Data preprocessing

3.1.1. Conversion of UTC to LT

Since the triplet components for the TCA-based error calculation should be closely aligned in time, the GLDAS data that is provided in Coordinated Universal Time (UTC) needs to be converted to Local Time (LT). As we consider SMAP, ASCAT, and GLDAS as triplet, we select GLDAS LT as 8 AM and 8 PM to align as closely as possible with the SMAP and ASCAT acquisition times, which are 6 AM/PM and 9:30 AM/PM, respectively. Given that the GLDAS Noah dataset is provided at 3-hour intervals, there are eight datasets available per day. As 1 hour roughly corresponds to 15° of longitude, we extracted data within a 45° range, which indicates selecting data within ± 1 hour of the target LT, from every eight files per day. These

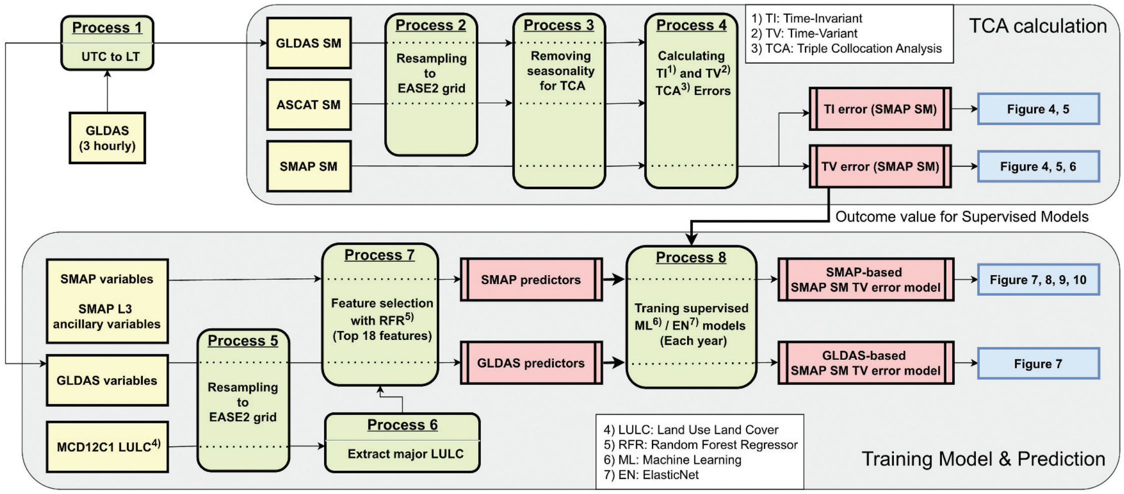

Figure 1. Overall research flowchart. Yellow boxes represent the datasets, green boxes indicate the processes, red boxes highlight the key results, and blue boxes represent the corresponding figures in this study.

patched UTC datasets were then integrated to create a single LT dataset for each day (Process 1 in Figure 1)

3.1.2. Data resampling

To obtain a spatially consistent dataset, the GLDAS and ASCAT datasets were resampled to align with the 36-km EASE-2 grid system. The raw dataset was divided into nine pixels per original pixel and then aggregated by averaging the values within the same EASE-2 grid cell. During this process, interpolation with nearest-neighbor sampling was applied to the original grid system. This indicates that for any given input, the value from the closest available point in the original dataset is used. Meanwhile, the MCD12C1 datasets were resampled to align with the same grid using the same interpolation method, but aggregated by selecting the most frequent value within each EASE-2 grid cell. These processes are illustrated as Process 2 and 5 in Figure 1.

3.1.3. Extracting major LULC type

As the LULC dataset is acquired annually, we treated it as a static variable by extracting the most frequent LULC type for each spatial pixel across the study period (2015–2023). This static LULC map can be used as a predictor in GLDAS-based model, similar to static ancillary datasets for soil properties in SMAP-based model. Furthermore, the extracted static LULC information was used to analyze the distribution of SMAP SM and TCA error with respect to different LULC types. The process of extracting the major LULC types is depicted as Process 6 in Figure 1.

3.2. Triple collocation analysis (TCA)

3.2.1. TCA for error quantification

Due to the inherent uncertainty – stemming from detection limitations – various methods have been developed to quantify the relative error without a true reference. In this study, we employed the TCA method to quantify the error in the observation dataset. This method determines the unknown error variance among three independent measurement systems, without assuming any one system to be a perfectly true reference. The TCA method relies on a linear error model where errors are uncorrelated with each other and the target variable. The error model for TCA typically follows this form:

$$i = \alpha_i + \beta_i \Theta + \varepsilon_i \quad (1)$$

where $i \in [X, Y, Z]$ are three collocated datasets, Θ is the true SM state, α_i is additive bias, β_i is multiplicative bias, and ε_i is zero-mean random noise of dataset i (Stoffelen 1998). Using this error model, the variance and covariance of each dataset can be calculated.

$$\sigma_i^2 = \langle (i - \bar{i})(i - \bar{i}) \rangle = \beta_i^2 \sigma_\Theta^2 + 2\beta_i \sigma_{\Theta\varepsilon_i} + \sigma_{\varepsilon_i}^2 \quad (2)$$

$$\sigma_{ij} = \langle (i-\bar{i})(j-\bar{j}) \rangle = \beta_i \beta_j \sigma_\Theta^2 + \beta_i \sigma_{\Theta \varepsilon_j} + \beta_j \sigma_{\Theta \varepsilon_i} + \sigma_{\varepsilon_i \varepsilon_j} \quad (3)$$

where σ_i^2 is the variance of SM product of i, $\sigma_{\varepsilon_i}^2$ is the variance of error of i, and σ_{ij} is the covariance between i and j SM products.

The TCA requires several underlying assumptions as follows: (i) linearity between the true SM signal and the estimation, (ii) signal and error stationarity, (iii) error orthogonality (i.e. $\sigma_{\Theta\varepsilon_i} = 0$), and (iv) zero error-cross correlation (i.e. $\sigma_{\varepsilon_i\varepsilon_j} = 0$). If these assumptions are satisfied, the variance and covariance can be expressed in a simplified form:

$$\sigma_i^2 = \beta_i^2 \sigma_\Theta^2 + \sigma_{\varepsilon_i}^2 \quad (4)$$

$$\sigma_{ij} = \beta_i \beta_j \sigma_\Theta^2 \quad (5)$$

Since a higher β_i indicates a stronger response of dataset i to SM changes, $\beta_i^2 \sigma_\Theta^2$ is considered as the sensitivity of the dataset to SM variations. This sensitivity can be estimated by combining covariances as follows:

$$\beta_i^2 \sigma_\Theta^2 = \frac{\beta_i \beta_j \sigma_\Theta^2 \beta_i \beta_k \sigma_\Theta^2}{\beta_j \beta_k \sigma_\Theta^2} = \frac{\sigma_{ij}\sigma_{ik}}{\sigma_{jk}} \quad (6)$$

Therefore, the TCA-based error variance for each individual SM product can be calculated using the variance of the individual dataset and the covariance between different datasets acquired from SM observations.

$$\sigma_{\varepsilon_i}^2 = \sigma_i^2 - \frac{\sigma_{ij}\sigma_{ik}}{\sigma_{jk}} \quad (7)$$

Moreover, this TCA-based error variance can be utilized to calculate the widely used metric that quantifies the relationship between observed signal and error: the signal-to-noise ratio (SNR). The SNR can be calculated as follows:

$$\mathrm{SNR}_i = \frac{\beta_i^2 \sigma_\Theta^2}{\sigma_{\varepsilon_i}^2} \quad (8)$$

Using the definition of SNR, the fractional Mean Squared Error (*f*MSE) can be calculated, offering a normalized and intuitive representation of SNR (Gruber et al. 2016):

$$f\mathrm{MSE}_i = \frac{\sigma_{\varepsilon_i}^2}{\sigma_i^2} = \frac{\sigma_{\varepsilon_i}^2}{\beta_i^2 \sigma_\Theta^2 + \sigma_{\varepsilon_i}^2} = \frac{1}{1+\mathrm{SNR}_i} \quad (9)$$

When *f*MSE is close to 0, the SNR is extremely high, indicating that data noise is significantly low. Conversely, when *f*MSE is close to 1, the SNR approaches to 0, meaning data noise is significant. When *f*MSE is 0.5, the SNR is 1, indicating that the observed SM signal is equal to the noise variance. The calculated *f*MSE results in this study will be discussed in detail in Section 4.

3.2.2. Calculation of standard deviation for time-invariant and time-variant errors

When calculating the covariance across three datasets, our focus is on covariance values that reflect noise levels, not background seasonalities. If background seasonalities are present, they can obscure the noise level in the covariance. Hence, deseasonalizing datasets before calculating TCA error is necessary. In this study, we applied deseasonalization by subtracting a moving average with a window size of 201, which corresponds to 100 days of both ascending and descending data, roughly representing the length of one season (Process 3 in Figure 1).

After removing seasonality from the datasets, we characterized TCA error standard deviation (SD) with two types: time-invariant error SD (hereinafter $\sigma_\varepsilon^{\mathrm{TI}}$) and time-variant error SD (hereinafter $\sigma_\varepsilon^{\mathrm{TV}}$). For the calculation of $\sigma_\varepsilon^{\mathrm{TI}}$, we set the threshold for the minimum number of valid points, to be 100. The $\sigma_\varepsilon^{\mathrm{TI}}$ assumes that errors remain constant (i.e. unchanged over time) throughout the entire period, which may oversimplify the error characteristics as it overlooks potential error sources that fluctuate over time. Beyond that, $\sigma_\varepsilon^{\mathrm{TV}}$, which is calculated with a smaller threshold of available SM dataset (i.e. 25) and a window size (i.e. 201), provides a more nuanced and accurate understanding of the data compared to the $\sigma_\varepsilon^{\mathrm{TI}}$, as it is more effective approach in dynamic and realistic estimations of σ_ε. The selected window size is same in the part on removing seasonality: 201 points in the time series, which includes 100 points (i.e. 50 days) before and after the point of interest, plus the point itself. If this window size is too short and the valid data threshold is high, the $\sigma_\varepsilon^{\mathrm{TV}}$ becomes sparse. Therefore, considering a threshold of 25 valid data points, we determined the window size based on one season length of approximately 3 months. The calculation of σ_ε^{TI} and σ_ε^{TV} is depicted as Process 4 in Figure 1, with a detailed diagram provided in Figure 2.

3.2.3. Four masking flags in TCA calculation

Due to the inherent limitations of TCA, gaps can occur in both σ_ε^{TI} and σ_ε^{TV} results. These gaps arise because data points were masked using four specific flags: (i) the correlation flag, which indicates whether the correlation between two covariance values is below a specified threshold (set to 0 in this study), (ii) the valid number flag, which checks if the number of valid data points below a certain node threshold (set to 100 for σ_ε^{TI} and 25 for σ_ε^{TV} in this study), (iii) the *f*MSE flag, which identifies cases where the *f*MSE value is either less than 0 or greater than 1, and (iv) the negative error variance flag, which marks points where the variance of the error is less than 0 (Kim et al. 2023). If the TCA calculation result triggers any of these flags, the corresponding data points are masked, leading to substantial spatial and temporal gaps. To address this issue, we proposed the tree-based ML and DNN framework to predict σ_ε^{TV} of SMAP. This model fills in missing σ_ε^{TV} where the SMAP SM dataset is available by using other simultaneously existing parameters as predictors within an ML model. We evaluated this approach using two thematic models: (i) a GLDAS-based ML model (i.e. external sourced model), trained on GLDAS variables, and (ii) a SMAP-based ML/DNN model (i.e. stand-alone model), trained on SMAP variables.

3.3. Uncertainty prediction

3.3.1. Feature selection with random forest regressor

For effective model performance, the application of feature selection algorithms is widely recommended in the ML field (Belgiu and Drăguţ 2016). Among various algorithms, Random Forest (RF) is a suitable strategy for this task, as it can capture intricate interactions among features by constructing multiple decision trees, each based on a random subset of features (Jovic, Brkic, and Bogunovic 2015). After training the RF model, features with high importance scores are selected as the most relevant. This RF-based feature selection algorithm identifies the key variables as predictors, facilitating the construction of an effective ML model. In this study, we used a Random Forest Regressor (RFR) for the feature selection process. Note that before inputting the supervised predictor and outcome into the RFR or predictive models, we applied standard scaling to both the predictor and outcome values separately to standardize the datasets. This process prevents imbalanced scaling within the predictor values and makes the outcome value's range more manageable, allowing the loss function to be effectively computed without unintentionally weighting variables with larger values. After the ML model predicted the outcome variable, we converted it back to its original scale.

Given the large size of the datasets, which consist of 406 (latitude) $\times$ 964 (longitude) $\times$ N_{doy} (days of year from 2015 to 2023) $\times$ 2 (ascending and descending) $\times$ $N_{\text{variables}}$ (number of variables), we applied the feature selection process to each year's dataset individually. We carefully designed the feature

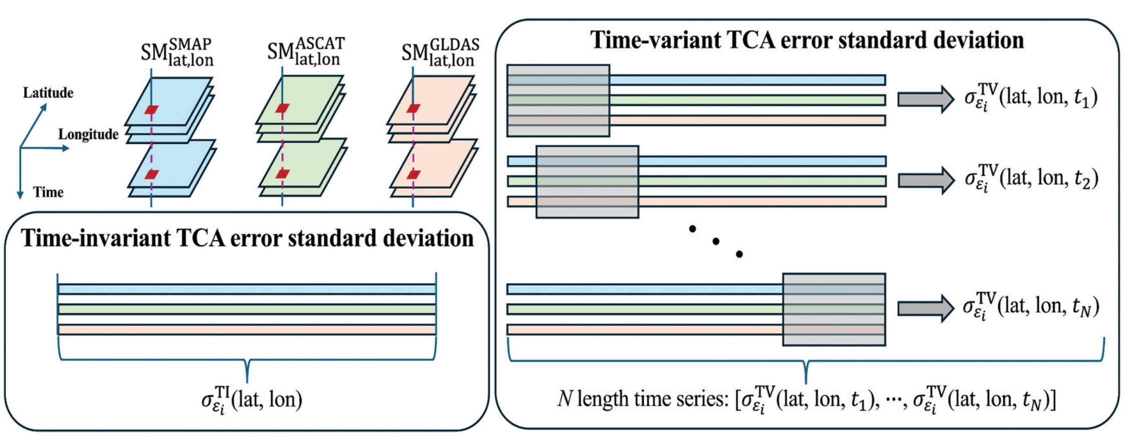

Figure 2. Schematic diagram illustrating the calculation of time-invariant TCA error standard deviation (σ_ε^{TI}) and time-variant TCA standard deviation (σ_ε^{TV}).

selection process using a two-step approach: initially identifying important features using the RFR, followed by a detailed correlation analysis. The RFR enabled us to rank predictors based on their importance for the target variable, while the correlation analysis using Spearman's rho allowed us to identify and remove the redundant variable with high inter-correlation, as shown in Figure S1. For example, we excluded *tb_v_uncorrected* due to its high correlation with *tb_v_corrected*, which rendered it redundant. Nevertheless, we retained predictors with meaningful physical relevance to SMAP $\sigma_{\varepsilon}^{TV}$ predictions, even if their correlation values were high. For instance, we retained variables associated with digital elevation model (DEM), even though they exhibited high correlation with each other, because they capture meaningful terrain information that supports SM prediction. Furthermore, the selection of predictors based on the correlation matrix also considered the physical properties of the data. For example, we removed *boresight_incidence* because, even though it is a critical variable in retrieving SM data from SMAP, it showed minimal variation around SMAP's constant incidence angle of 40 degrees and had very low correlation with the response variable.

These rigorous processes result in the exclusion of unnecessary variables, reducing our top 20 predictors (Figure S1) to 18. This comprehensive consideration ensured that the selected predictors were relevant, minimally redundant, and physically meaningful, resulting in a more robust and efficient input set for the error prediction model. The selected features are highlighted in light blue in Table S1 for SMAP-based model and in Table S2 for GLDAS-based model. It is important to note that some variables were considered static, assuming they do not change over time: SMAP ancillary variables for the SMAP-based model (i.e. *CoastalMask*, *DEM*, *DEMSLP*, *DEMSLPSTD*, *DEMSTD*, *dist2water*, *roughness*, and *sand*) and LULC variable in the GLDAS-based model (i.e. *Majority_Land_Cover_Type_1*). To construct the predictor dataset, these static variables were duplicated across time and combined with other dynamic variables by matching each pixel location and time. This approach allowed us to develop a single ML/DL model that is better generalized for various land property conditions. Figure S2 displays the bar charts of feature importance scores for these top 18 predictors across nine years (2015–2023) for both models (Process 7 in Figure 1). The mutually influential relationship among several variables may affect the feature importance scores. For instance, Figure S2(a) shows that the scores for *roughness* and *vegetation_opacity* in 2016 and 2020 differ from those in other years. In these two years, the feature importance scores of *roughness* are low, while the scores for *vegetation_opacity* are high, indicating a trade-off connection.

3.3.2. Light gradient boosting machine

We utilized the LGBM algorithm which is based on the gradient boosting decision tree framework and is widely used because LGBM can achieve similar or even better performance with less computational cost than many other gradient boosting frameworks (Ke et al. 2017). Additionally, the hyperparameter tuning of the LGBM model was performed by Bayesian optimization. Bayesian optimization constructs a probabilistic model of the objective function and iteratively selects hyperparameters by maximizing an acquisition function, which predicts where the most promising results are likely to be found. This approach is especially beneficial for intricate, high-dimensional spaces where traditional grid search or random search methods might be inefficient (Shahriari et al. 2016). The negative Mean Squared Error (MSE) was used as the optimization score. Since Bayesian optimization's objective is to maximize the acquisition function, using a positive MSE is inappropriate because a lower MSE indicates better model performance, which is opposite to the direction of maximization. Therefore, by using negative MSE as the score, maximizing the score effectively means finding the best hyperparameters. As noted in the previous section, due to the large dataset size, we trained the error prediction models separately for each year using the same selected predictors. The hyperparameters of each year are listed sequentially in Table S3.

3.3.3. Deep neural network

To expand our research into the deep learning (DL) field, we used the DNN model, which is well-known for its foundational architecture that mimics the human brain's recognition process, such as the function of neurons.

DNNs are well-suited for processing high-dimensional data because of their deep hidden architectures, which allow them to grasp complex, non-linear relationship between inputs and outputs. Using a simple DNN structure, we examined this potential to see whether DL can achieve promising performance. The network architecture consists of an input layer with 18 neurons (reflecting the top 18 predictors), four hidden layers, and an output layer with one neuron (representing the predicted $\sigma_{\varepsilon}^{TV}$ value). We designed the four fully connected hidden layers with neuron counts [128-64-32-16], which indicate a decreasing number of neurons. In order to capture the non-linear relationship in the data, we also used the Rectified Linear Unit (ReLU) activation function after each hidden layer. To effectively reduce the MSE loss function, we utilized the Adaptive Moment estimation (Adam) optimizer with a learning rate of 0.001 (Kingma and Ba 2017). During model training, we set the batch size to 128, and the validation set was randomly 20% selected from the training set for each epoch. To avoid overfitting, we used an early stopping callback. This callback monitors the validation loss and saves the best model, terminating training when no improvement is shown after a predetermined patience time (30 epochs in this study). The DNN model architecture is illustrated in Figure S3.

3.3.4. Elastic Net

To demonstrate the significance of ML's non-linear approach, we compared the results from LGBM and DNN models with those obtained using Elastic Net (EN), a regularized multivariate linear regression model that combines L1 and L2 priors as regularizers (Zou and Hastie 2005). The hyperparameters for EN, specifically alpha and L1 ratio, were also optimized using Bayesian optimization. The optimal hyperparameters were determined to be an alpha of 0.1 and an L1 ratio of 0.1 across all years. However, because the data do not exhibit a linear relationship, these hyperparameters poorly represent the dataset, as evidenced by the prediction results. The features and advantages of each prediction model that we considered in this study are summarized in Table S4.

3.3.5. Prediction and reconstruction

As mentioned above, due to the large dataset size, three model structures (i.e. LGBM, DNN, and EN) were considered separately for each year, and the results were concatenated to create the total predicted $\sigma_{\varepsilon}^{TV}$. Additionally, using the trained models with the optimized hyperparameters, we reconstructed the missing $\sigma_{\varepsilon}^{TV}$ values where SM data were available but $\sigma_{\varepsilon}^{TV}$ values were missing. Likewise, these reconstructed results were obtained for each year and the total central tendency and variability of $\sigma_{\varepsilon}^{TV}$ were obtained by calculating the temporal average and SD of concatenated $\sigma_{\varepsilon}^{TV}$ (Process 8 in Figure 1).

4. Results

4.1. Comparison of time-invariant vs. time-variant TC-based error

Figure 3(a) highlights dry and wet regions, employing a range of colors to signify SM levels, with light blue denoting drier areas and dark blue representing wetter areas. The regions with low SM values include western North America, eastern and southern South America, northern and southern Africa, Southwest Asia, Central Asia, northern East Asia, and Australia. On the contrary, the regions with high SM values are found in eastern and northern North America, western and northern South America, Central Africa, northern Europe, southern East Asia, and Southeast Asia. This spatial variability of SM underscores the importance of understanding the error characteristics for each pixel in the SMAP SM product. Furthermore, Figure 3(b) shows the static global map of LULC, which includes 17 categories, and Figure 3(c) presents a violin plot illustrating time-averaged SMAP SM values (top) and data availability (bottom) across different LULC types, with SM values arranged from highest to lowest. Evergreen Broadleaf Forest (EBF) has the highest SM values, while Barren or Sparsely Vegetated (BSV) has the lowest. However, due to the challenges of acquiring SM in densely vegetated regions, EBF shows relatively lower data availability, whereas BSV has higher data availability. This could be attributed to the fact that dense canopy in the forested regions tends to keep more water in the soil, resulting in the highest SM values observed in EBF areas. On the other hand, the data availability increases with decreasing vegetation density, with relatively higher data counts in BSV and lower data counts in EBF (Figure 3(c)), highlighting the

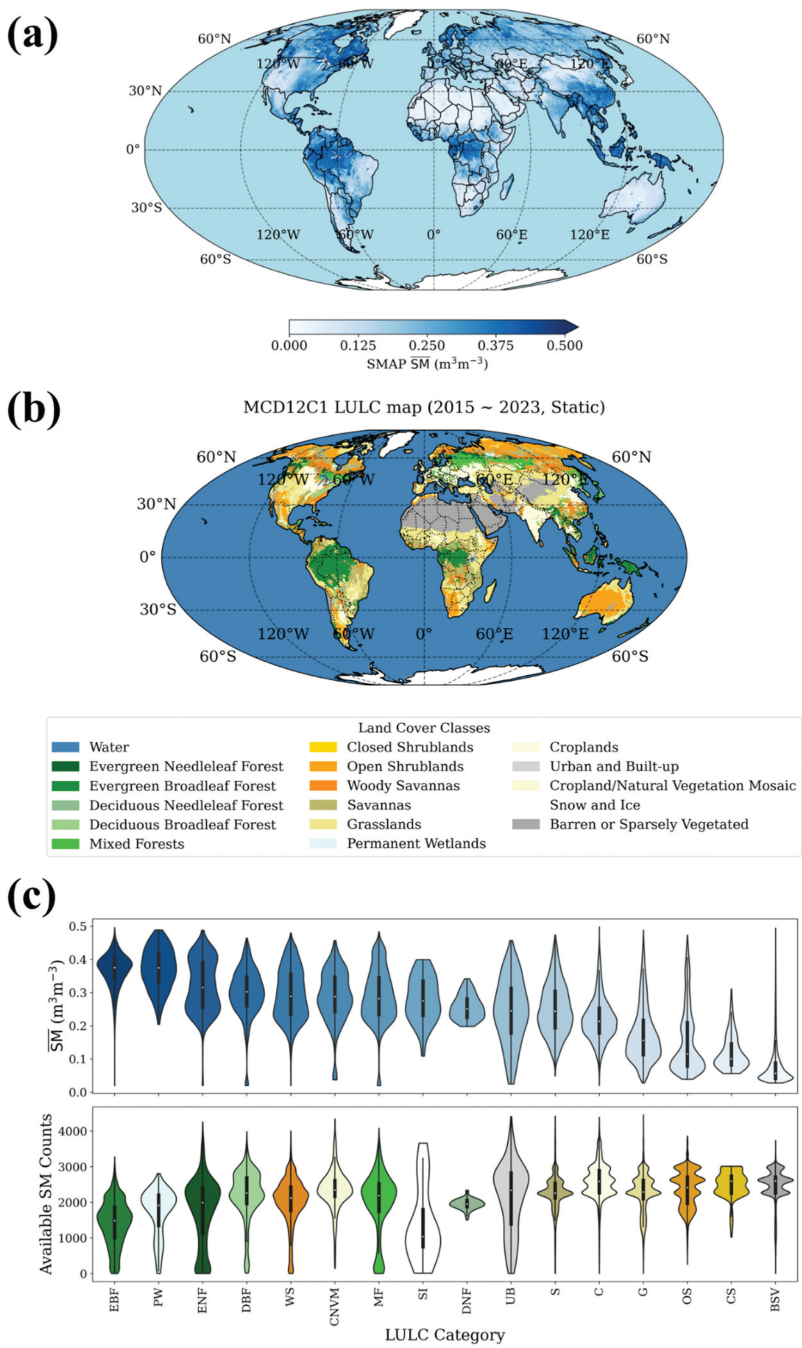

Figure 3. Analysis of SM and associated standard deviation of errors considering LULC. (a) Global map of SMAP SM values, where light blue indicates lower values and dark blues indicates higher values. (b) Global map of LULC types from MCD12C1, categorized into 17 types. (c) Violin plot of data from SM and available SM counts with respect to the LULC types in (b)

limitations associated with vegetation effect on SM retrieval from SMAP.

These available SM distributions across LULC types also result in variations in TCA calculation. For example, the TCA calculation is restricted in areas like EBF, where insufficient observations make it impossible to calculate. Meanwhile, the TCA-based error estimates are likely more reliable in regions such as BSV, which possess more SM observations. As a result, locations with sparse data have been less explored regarding the uncertainty of satellite-based SM data than those with more frequent observations.

Additionally, we investigated the number of available SMAP SM data points for each EASE2 grid pixel (from

2015 to 2023). As expected, and as shown in the map at the center of Figure 4, the global distribution of SMAP SM data is uneven. For example, the Amazon region has fewer data points, indicated by light purple, whereas Europe has a plentiful data points, represented by dark purple.

We further examined the temporal evolution of SMAP SM and its associated $\sigma_{\varepsilon}^{TI}$ and $\sigma_{\varepsilon}^{TV}$. Their corresponding time series of randomly selected pixels, with one pixel chosen from each of the six continents (i.e. Africa, Asia, Australia, Europe, North America, and South America), are extracted and shown in Figure 4, together with the SMAP SM data availability map. For each time-series plot, the top section displays whether any of the four flags (described in Section 3.2.3) were raised during the calculation of TCA. For example, if all four flags are raised, the entire row of the top subplot would be filled, while if only flag 1 is raised, only the first row will be marked in black. This table-like flag plot is useful for investigating why, at certain times, the time-variant error variance, $\sigma_{\varepsilon}^{TV}$, of SMAP SM data could not be calculated using TCA. The middle subplot shows SM values, with blue dots representing solely existence without the $\sigma_{\varepsilon}^{TV}$ time series and pink dots indicating concurrent existence with the $\sigma_{\varepsilon}^{TV}$ time series. The bottom subplot illustrates the associated $\sigma_{\varepsilon}^{TI}$ and $\sigma_{\varepsilon}^{TV}$ over

Figure 4. Global map of available SMAP SM counts and time series of six randomly selected points from six continents, marked with red stars: North America (37.08°N, 93.92°W), South America (4.95°S, 67.78°W), Europe (48.58°N, 18.11°E), Africa (20.93°N, 5.04°E), Asia (27.10°N, 111.47°E), and Australia (22.44°S, 126.41°E). In the central global map, light purple indicates fewer available SMAP SM counts, while dark purple indicates more available SMAP SM counts. In the time series plot, there are three subplots. In the upper subplot, the black lines represent the points that satisfy flags (flag 1: correlation flag, flag 2: valid number flag, flag 3: *f*MSE flag, and flag 4: negative error variance flag). In the middle subplot, blue indicates SM values that do not coincide with $\sigma_{\varepsilon}^{TV}$, while pink indicates SM values that do. In the lower subplot, the green dot represents $\sigma_{\varepsilon}^{TI}$ and the red dot represents $\sigma_{\varepsilon}^{TV}$.

time, with the green dots representing σ_ε^{TI} and the red dots depicting σ_ε^{TV}. For example, in the bottom subplot (beneath the center map), which represents a point in Africa, the average SM value is around 0.05 (indicated by the blue dots in the middle subplot), and both σ_ε^{TI} and σ_ε^{TV} could not be calculated, as explained by the flag information in the top subplot. In contrast, the lower left subplot, representing a point in the Amazon region of South America, highlights fewer SM data points in vegetated regions due to masking. These results highlight that even when SMAP SM data is available, the error for many SM values could not be calculated due to various factors, which represents a significant data loss for the SMAP mission. In summary, both σ_ε^{TI} and σ_ε^{TV} are not always concurrently available with the SMAP SM data, due to the issues highlighted in the top flag plots.

4.2. Analysis of spatial SMAP SM error patterns and their seasonality

Figure 5 presents the spatial maps of both σ_ε^{TI} and time-averaged σ_ε^{TV} on a global scale. In Figure 5(a), the σ_ε^{TI} shows higher error in wetter regions with dense vegetation, such as the Amazon, and lower error in the drier regions, such as Australia. In Figure 5(b), the time-averaged σ_ε^{TV} values for each pixel were calculated to demonstrate the central tendency of SMAP SM σ_ε^{TV}. The time-averaged σ_ε^{TV} map resembles the overall error patterns of the σ_ε^{TI} map and fills in many of the gaps present in the σ_ε^{TI} map. Additionally, as seen in Figure 5(c), the standard deviation of σ_ε^{TV} was plotted in order to investigate its variability. The central tendency and variability of σ_ε^{TV} show similar patterns in certain locations, but not in others. For example, northern Africa has both low averaged σ_ε^{TV} and low SD of σ_ε^{TV}, whereas the Amazon shows high averaged σ_ε^{TV} and high SD of σ_ε^{TV}. On the other hand, Australia displays low averaged σ_ε^{TV} but high SD of σ_ε^{TV}. Since these areas, characterized by a higher SD in σ_ε^{TV}, are regions where reliance on σ_ε^{TI} is insufficient, calculating σ_ε^{TV} is particularly necessary in these regions for practical SMAP SM applications. Special attention should be given to these regions, particularly for LDA and data merging. In other words, areas with high SD of σ_ε^{TV} can benefit more from applying σ_ε^{TV} instead of σ_ε^{TI}. Moreover, the temporally averaged global map for *f*MSE, as explained in Equation (9) in Section 3.2.1, is provided in Figure 5(d). A lower *f*MSE indicates that the noise influence is significantly lower than the SM signal, whereas a higher *f*MSE represents that noise has a stronger influence compared to the SM signal. For example, although northern Africa has a low averaged σ_ε^{TV} value, the high averaged *f*MSE indicates that the SM

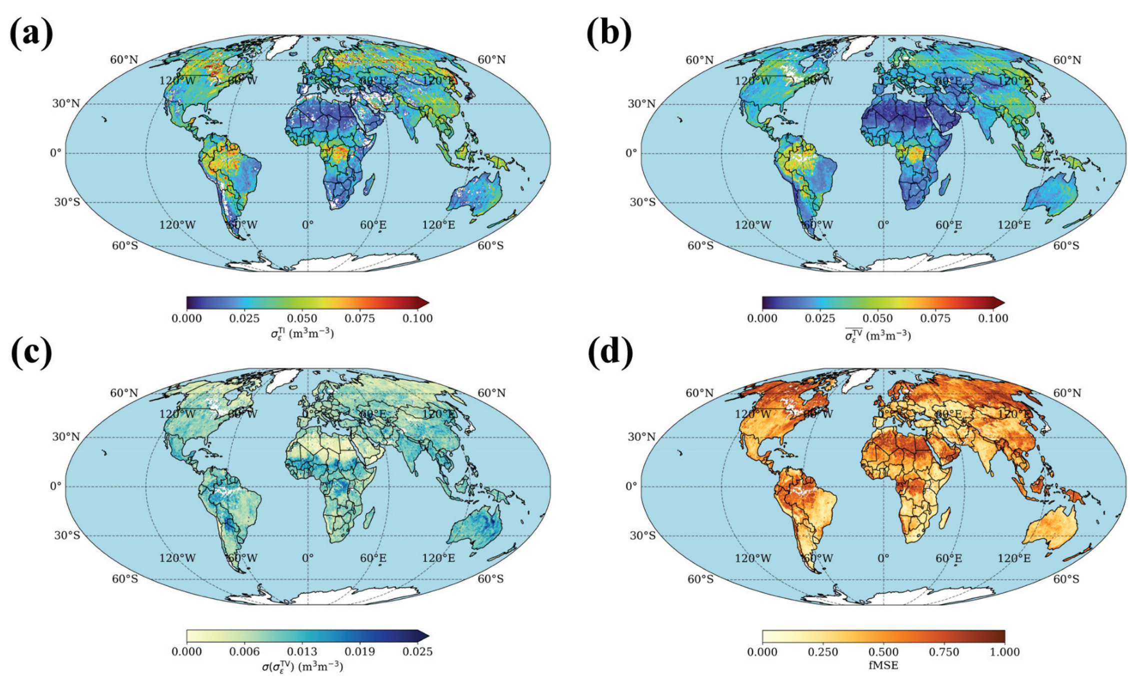

Figure 5. Global maps of (a) σ_ε^{TI}, (b) temporal mean of σ_ε^{TV}, and (c) temporal standard deviation of σ_ε^{TV}, and (d) temporal mean of *f*MSE.

signal has relatively lower impact than the noise signal. Therefore, considering both σ_ε^{TV} and *f*MSE together provides a more comprehensive assessment of SM signal quality. For readers who are interested in exploring the spatial patterns of *f*MSE over all seasons, Figure S4 offers global maps of seasonal values for *f*MSE.

Furthermore, we investigated the spatiotemporal variation of σ_ε^{TV}, by making Hovmöller diagrams, a graphical tool commonly used in meteorology and climatology to display the temporal evolution of a spatial variable (Hovmöller 1949; Peng et al. 2023; Wu et al. 2021). We employed this Hovmöller diagram to examine the spatial variability along a line of constant latitude. Figure 6 shows distinct seasonal patterns, particularly in regions closer to the poles where σ_ε^{TV} tends to increase during specific periods. This finding is consistent with the fact that satellite-based SM retrievals are more prone to inaccuracy in areas that experience seasonal variations, such as snow cover, freeze-thaw cycles, and changes in vegetation density. For example, during the winter months at higher latitudes (e.g. around 60°N), the presence of snow and frozen ground can introduce significant uncertainty in SMAP's SM retrieval, as the satellite's ability to accurately detect SM is reduced under such conditions. This leads to an increase in σ_ε^{TV}, which is clearly depicted in the peaks observed at higher latitudes in the diagram.

The mid-latitudes (e.g. around 30°N) also exhibit some degree of seasonal fluctuation and this can be attributed to agricultural activities, precipitation variability, and vegetative cycles that influence the accuracy of SM retrieval. In contrast, lower latitudes (e.g. around 0°), where tropical climates dominate, generally exhibit less seasonal fluctuation in σ_ε^{TV}, though still present periodic increases likely due to tropical rainfall or other localized events that temporarily affect SM accuracy.

Note that the use of a relatively large window size (201 points) in calculating σ_ε^{TV}, helps smooth out some of the data gaps that would otherwise be more prominent, given the challenges of consistent satellite retrieval in certain regions. However, despite this, some gaps in the time series remain, corresponding to periods when SMAP SM retrievals were unavailable (e.g. 2019 and 2022), further emphasizing the difficulty of consistently measuring SM in certain geographic areas.

Overall, Figures 5, 6, and S4 highlight how the spatial and temporal distribution of σ_ε^{TV} is influenced by both seasonal and latitudinal factors. Comprehending these patterns is important for improving satellite-based SM error correction and ensuring more reliable data integration into LDA systems, especially in areas prone to significant seasonal variation.

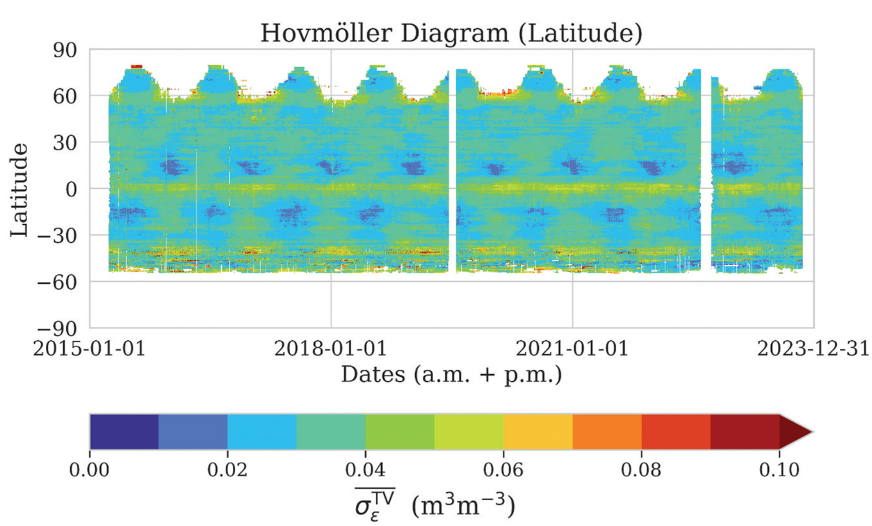

Figure 6. Hovmöller diagrams of latitudinal-mean σ_ε^{TV}. The areas close to the blue color indicate low values, whereas the red color represents high values.

4.3. SMAP SM time-variant error prediction models

We utilized two thematic LGBM model architectures to develop the SMAP SM σ_ε^{TV} prediction models using different predictors: the external-source SMAP SM σ_ε^{TV} prediction model and the stand-alone SMAP SM σ_ε^{TV} prediction model. The external-source model utilized the GLDAS dataset as predictors, while the stand-alone model used the SMAP L3 and ancillary dataset as predictors. Additionally, to demonstrate the importance of the LGBM and DNN models' ability to capture non-linear relationships in the data, we compared the results of the these models with those of the EN model, a form of multivariate linear regression model. Hence, Figure 7 presents the prediction results for the combination of the two thematic predictors (i.e. GLDAS-based vs. SMAP-based model) and the three model architectures (i.e. LGBM, DNN, and EN).

First, by comparing Figures 7(a,b), we observed that the stand-alone LGBM model demonstrates greater accuracy than the external-source LGBM model, as indicated by its closer alignment with the diagonal 1:1 line. We used Mean Absolute Error (MAE) as the evaluation metric because it assigns equal weight to all individual errors, making it less sensitive to outliers compared to MSE. The MAE for the stand-alone LGBM model is 0.006891, whereas the MAE for the external-source LGBM model is 0.008797. This indicates that the stand-alone model is better equipped to capture the dynamic properties of SMAP σ_ε^{TV} owing to the concurrent acquisition of SM and other parameters from the same SMAP satellite mission. In addition, the stand-alone model offers an advantage in developing the error prediction models as it involves fewer data preprocessing steps compared to the external-source model.

Next, Figure 7(c) shows the effectiveness of the stand-alone DNN model, with an MAE of 0.007249. Despite the DNN model exhibiting a slightly higher MAE compared to the LGBM model, the overall prediction outcomes remain closely aligned with the 1:1 line, indicating the potential of DL approach. This suggests that further research to find optimized DL model architecture could be beneficial. Furthermore, a comparison of Figures 7 (b,c) with Figure 7(d) clearly demonstrates that

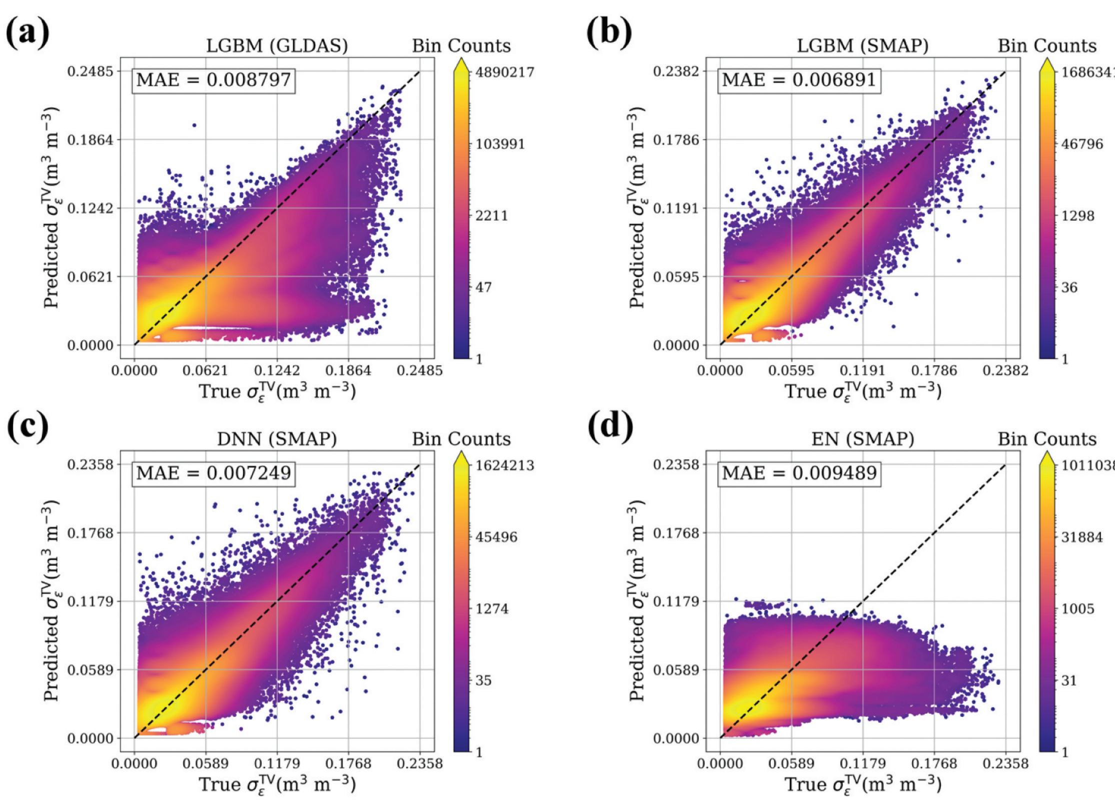

Figure 7. Stand-alone model (LGBM, DNN, and EN) vs. external-source model. (a) external-source LGBM model, (b) stand-alone LGBM model, (c) stand-alone DNN model, (d) stand-alone EN model.

both ML and DNN models outperform the EN model. The stand-alone EN model's MAE is 0.009489, which is even higher than that of the external-source LGBM model. This emphasizes the importance of accounting for non-linear relationships between predictors and σ_ε^{TV} values.

4.4. SMAP SM time-variant error reconstruction from stand-alone error prediction models

In the previous section, we observed that the SMAP data-based (i.e. stand-alone models) outperformed both the external-source ML model and stand-alone EN model. Therefore, in this section, we focus on reconstructing σ_ε^{TV} values where SM and other SMAP predictors are available using two stand-alone LGBM and DNN models. For the LGBM model, we utilized the best hyperparameters identified through Bayesian optimization to construct each annual model. We trained these LGBM annual models using all available datasets for supervised learning and then temporally concatenated the annual results. Using these trained models, we predicted σ_ε^{TV} values for points where SM and other predictor values were available but σ_ε^{TV} values were missing. For the DNN model, we used the same annual models from Figure 7(c) for reconstruct σ_ε^{TV}, instead of retraining the model with all available datasets, as was done with the LGBM model. This decision was made because re-training the DNN model is a time-consuming process. Furthermore, since we had previously split the data into training and testing sets with an 8:2 ratio, we assumed that the training set sufficiently reflects the distribution of all available data.

Figure 8 illustrates the reconstructed SMAP SM σ_ε^{TV} from the stand-alone LGBM and DNN models. Please note that this map was produced by integrating the predicted σ_ε^{TV} values from LGBM and DNN models (hereinafter ensemble ML models), to improve the robustness of σ_ε^{TV}predictions. Figure 8(a) shows the temporal mean global map of the reconstructed σ_ε^{TV} values, while Figure 8(b) presents the temporal SD of the reconstructed σ_ε^{TV} values. The reconstructed σ_ε^{TV} values exhibit regional differences, with some areas showing lower and others higher central tendencies and variability compared to Figures 5(b,c), which depict the temporal mean and SD of σ_ε^{TV} prior to reconstruction.

To clearly compare the temporal mean and SD before (Figures 5(b,c)) and after the reconstruction, Figures 8(c,d) illustrate how the distribution of σ_ε^{TV} changed after reconstruction. This comparison highlights the adjustments made by the reconstruction

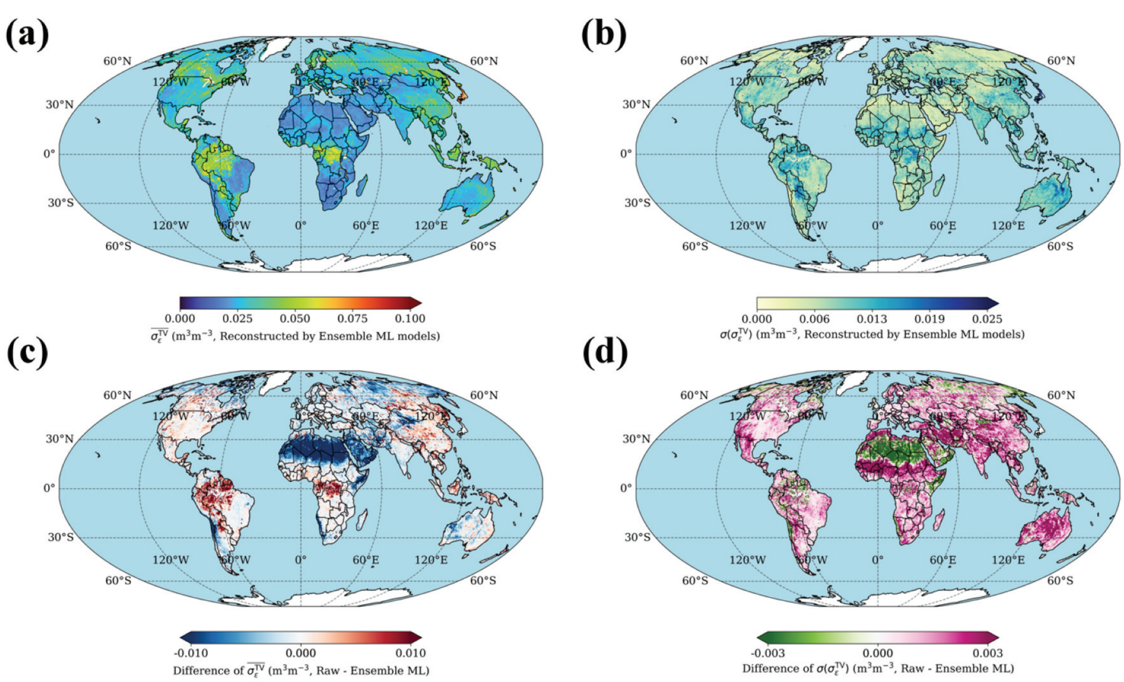

Figure 8. Global maps of (a) temporally averaged σ_ε^{TV}, (b) SD of σ_ε^{TV}, and (c) the difference between temporal averaged σ_ε^{TV} vs. temporal averaged reconstructed σ_ε^{TV}, and (d) the difference of temporal SD between before and after reconstruction. (i.e. raw – ensemble ML).

process, showing where the models captured increased or decreased error variability, improving the representation of spatial and temporal dynamics in the reconstructed σ_ε^{TV} values. The blue regions in Figure 8(c) illustrate areas where the ensemble ML models yielded greater σ_ε^{TV} than raw values, such as northern Africa, while red regions show areas where the models produced lower σ_ε^{TV} than raw values, like the Amazon and central Africa. Similarly, in Figure 8(d), green regions indicate areas where the ensemble ML models produced higher variability in σ_ε^{TV} than raw values, such as northern Africa, whereas purple regions indicate areas where the models produced lower variability of σ_ε^{TV} than raw values, as seen in overall except of northern Africa.

In Figure 9, the central global map presents the ratio of recovery relative to the available SM counts, ranging from 0 to 1. This ratio is defined by dividing the count of reconstructed σ_ε^{TV} by the count of available SMAP SM data. A value of 0 indicates that

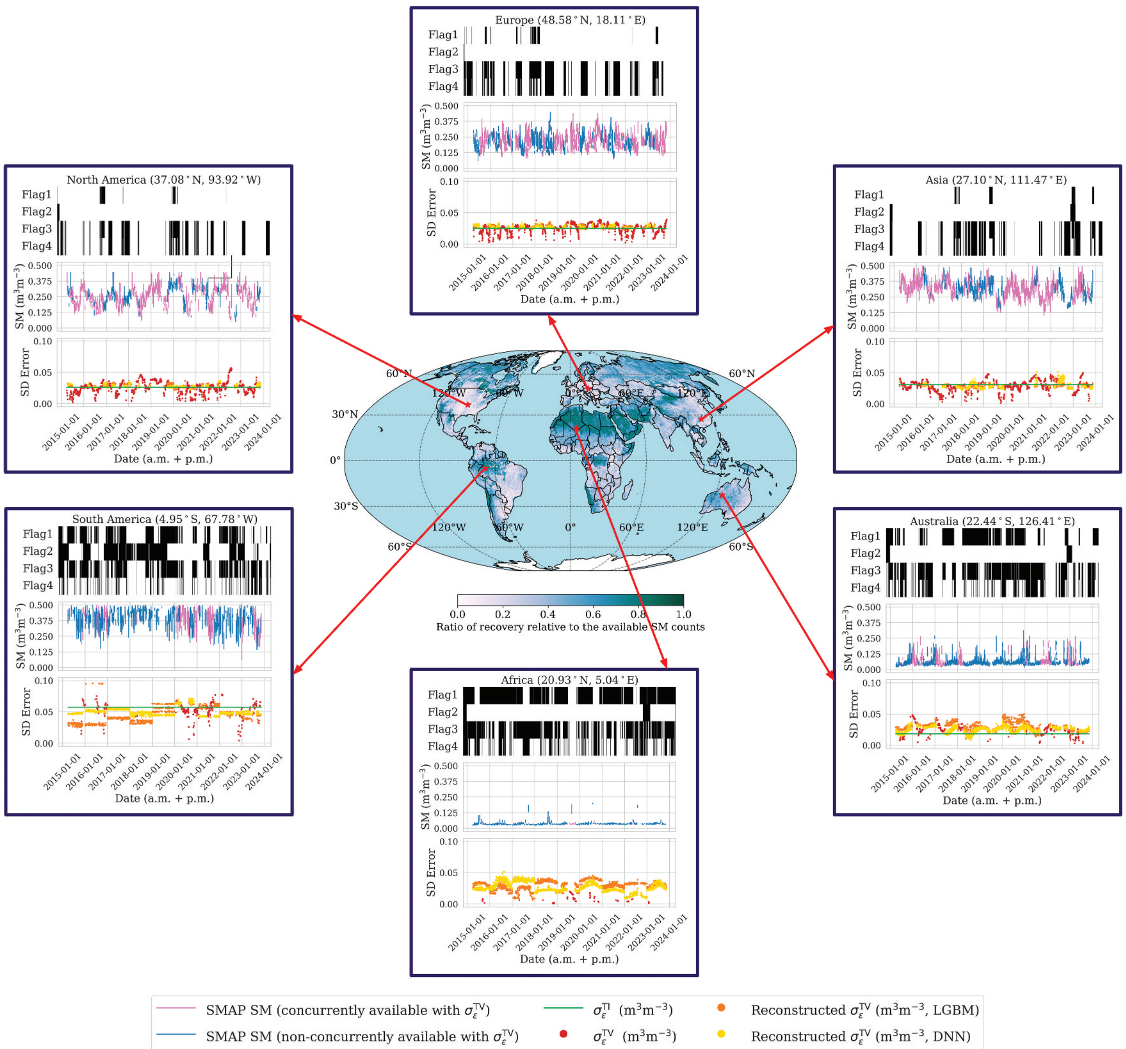

Figure 9. Global map of ratio of recovery relative to the available SMAP SM counts and time series of six randomly selected points from six continents, marked with red stars: North America (37.08°N, 93.92°W), South America (4.95°S, 67.78°W), Europe (48.58°N, 18.11°E), Africa (20.93°N, 5.04°E), Asia (27.10°N, 111.47°E), and Australia (22.44°S, 126.41°E). In the central global map, light purple indicates lower ratio of recovery, while green-blue indicates higher ratio of recovery. In the time series plot, there are three subplots corresponding to Figure 4. In the upper subplot, the black lines represent the points that satisfy flags (flag 1: correlation flag, flag 2: valid number flag, flag 3: *f*MSE flag, and flag 4: negative error variance flag). In the middle subplot, blue indicates SM values that do not coincide with σ_ε^{TV}, while pink indicates SM values that do. In the lower subplot, the green dot represents σ_ε^{TI} and the red dot represents σ_ε^{TV}. The newly added elements compared to Figure 4 are the lgbm-reconstructed σ_ε^{TV}, shown in orange, and the dnn-reconstructed σ_ε^{TV}, shown in yellow.

σ_ε^{TV} values were available whenever SMAP SM data was present, while a value of 1 means that all σ_ε^{TV} values were missing despite the availability of SMAP SM data. Regions with high recovery ratios are indicated in green-blue, while regions with low recovery ratios are shown in light purple.

To further illustrate the reconstruction results over time, Figure 9 also presents time series of both the LGBM and DNN reconstructed versions across the same randomly selected points as in Figure 4. The σ_ε^{TV} reconstructed by the LGBM model is shown by the orange scatter plots, whereas the σ_ε^{TV} reconstructed by the DNN model is indicated by the yellow scatter plots. Although the overall trends of both σ_ε^{TV} predictions are similar, slight differences in the spread are observed, particularly in locations with fewer raw σ_ε^{TV} data points, such as in South America, Africa, and Australia. These locations may exhibit reduced robustness in the proposed framework due to insufficient training data. Nonetheless, the predicted points reflect dynamic properties in response to the status of the predictors, as compared to σ_ε^{TI}.

Finally, for a more detailed sensitivity analysis, we utilized violin plots in Figure 10, categorized by LULC. These plots show the distribution of the error recovery ratio, shown in Figure 9, across different LULC types. BSV regions exhibit the highest median error recovery ratio, while Cropland/Natural Vegetation Mosaic (CNVM) has the lowest. Additionally, EBF regions, which has the highest SM values, but relatively low SM counts in Figure 3(c), has wide distribution of recovery ratio. This result aligns with previous challenges in calculating σ_ε^{TV} in these areas, underscoring the limitations of SM retrieval in regions with complex land cover.

5. Discussion

The proposed error prediction model can provide the increase in the temporal resolution of SM error estimation, approximately 119.47% (Figures 9 and 10). This implies that LDA and data merging can benefit significantly from the current method, as it maximizes the possible error recovery rate, allowing for improved data integration and application. However, although the proposed error prediction model's feasibility and robustness were shown, we identified several drawbacks. Thus, it is important to discuss the limitations and applications of the current study and compare our findings with those of previous research.

First, error recovery is not possible when predictor values are not concurrently available with SMAP SM. For instance, the "sand" variable in SMAP ancillary data does not always coincide with the availability of SM in certain areas, resulting in still missing time-variant error results even when SMAP SM data is available. To overcome this issue, one potential approach can be to either filling the gaps

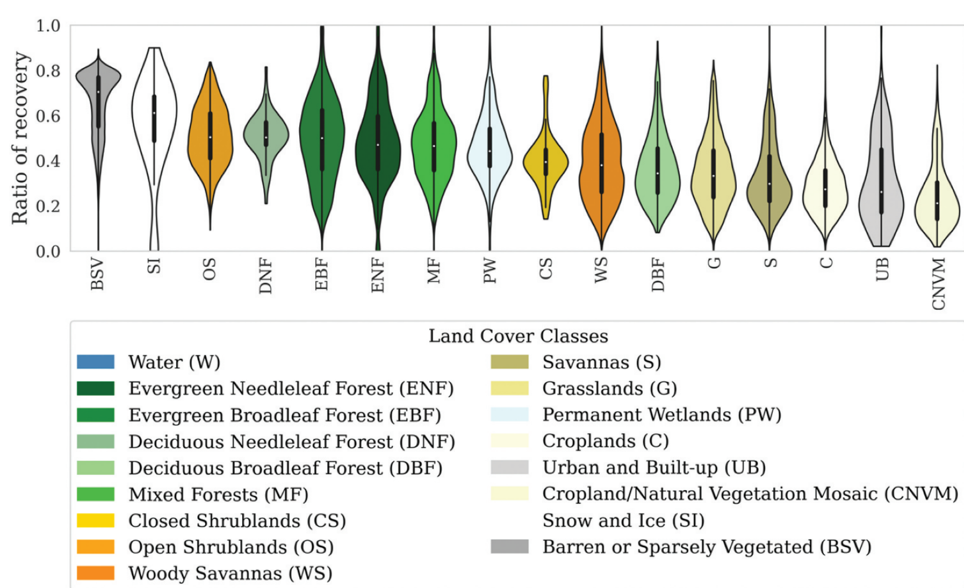

Figure 10. Violin plots of the ratio of recovery relative to the available SM counts with respect to the LULC types.

in the predictor variables before designing the error prediction model or replacing the predictor with an alternative data source. As shown in Figure 7(a), our results indicate that the GLDAS-based model performs reasonably well. Therefore, we expect that the proposed framework can incorporate some (though not all) variables from different sources when predictors from a single source are insufficient to construct a stand-alone error prediction model.

Second, setting the threshold in TCA is arbitrary without the optimal selection process. Wu et al. (2023) set the window size of 101-day – roughly equivalent to our study's 201 window (accounting for both AM and PM) – and set the threshold of minimum sample for $\sigma_{\varepsilon}^{TV}$ to be 90. However, to obtain the sufficient number of time-variant errors, Wu et al. (2023) applied the temporal interpolation on the SMOS and SMAP SM datasets before applying TCA calculation. They mentioned that temporal interpolation method can induce additional errors into SM time series, but it has relatively small impacts. In contrast, to create training dataset for supervised learning, our study set a low threshold (i.e. 25). This lower threshold allowed us to generate more $\sigma_{\varepsilon}^{TV}$ values across different regions and time periods, however, it resulted in reduced robustness compared to using a higher threshold.

Third, our study only considers the one combination of SM triplet acquired from SMAP, ASCAT, and GLDAS. Meanwhile, Gruber et al. (2017) employed SM datasets obtained from ASCAT, Advanced Microwave Scanning Radiometer for EOS (AMSR-E), and GLDAS as SM triplet, and Kim et al. (2023) explored various triplet combinations. Therefore, for future study, a more robust framework can be constructed by considering multiple triplet combinations as suggested by these studies.

Finally, the methodology proposed in this study possesses considerable potential for broader applications beyond SMAP SM data. This framework can be adapted to other satellite systems for hydrometeorological variables such as SM data from the SMOS mission or precipitation data from the Global Precipitation Measurement (GPM) mission, enabling precise error prediction and offering a reliable solution for estimating time-variant errors in a variety of remote sensing products. According to Wu et al. (2023), other SM products, such as ASCAT and SMOS, had superior efficacy in estimating SM dynamics under specific conditions based on land cover analysis. Therefore, further research may explore incorporating other SM products to estimate error characteristics and developing region-specific ML/DL models.

6. Conclusion

In this study, stand-alone triple collocation analysis (TCA)-based time-variant soil moisture (SM) error prediction models were developed using tree-based machine learning and deep learning methods: Light Gradient Boosting Machine (LGBM) and Deep Neural Networks (DNN). The SM dataset from Soil Moisture Active Passive (SMAP), Advanced Scatterometer (ASCAT), and Global Land Data Assimilation System (GLDAS) are utilized as the triplet to calculate the time-variant TCA error in SMAP SM estimates globally over the period from 2015 to 2023.

The robustness of the proposed stand-alone SMAP SM error prediction model was evaluated with that of a model developed using external datasets, such as GLDAS. The stand-alone LGBM model achieved an MAE of 0.0069, outperforming the external source model's MAE of 0.0088 by approximately 21.7%. Furthermore, our reconstructed time-variant error standard deviation ($\sigma_{\varepsilon}^{TV}$) by the proposed SM error prediction models results in an approximately 119.47% increase in the temporal resolution of $\sigma_{\varepsilon}^{TV}$ on global scale. These findings demonstrated that the proposed error prediction models effectively capture global spatiotemporal changes in SMAP SM error under varying environmental conditions and can effectively retrieve error information whenever SMAP SM data is available, addressing a significant gap witnessed in previous SMAP applications.

As global environmental monitoring efforts continue to expand, the ability to dynamically predict error characteristics will be critical for improving the accuracy and reliability of satellite-derived datasets. Knowing the accuracy of satellite-based SM data will eventually enhance land data assimilation (LDA) practices and advance our understanding of Earth's climate processes. For future research, we anticipate that the spatiotemporal dynamic error information of satellite-based SM data will not only support the provision of data masking flags but also facilitate the merging of different SM products through diverse LDA methods.

Acknowledgments

The authors thank the teams from NASA and EUMETSAT for making their data sets publicly available. We utilized generative artificial intelligence tools (ChatGPT-4o) to enhance the clarity and grammar of the sentences, improving readability and comprehension.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the AI-based GIST Research Scientist Project grant funded by the GIST in 2024. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A5A1028138). This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education [RS-2024-00392418].

ORCID

Subin Kim http://orcid.org/0009-0002-1575-1619
Hyunglok Kim http://orcid.org/0000-0002-5790-7012
Yonghwan Kwon http://orcid.org/0000-0001-9901-8670
Hoang Hai Nguyen http://orcid.org/0000-0002-2000-9294

Data availability statement

SMAP L3 dataset is freely accessible at https://www.earthdata.nasa.gov/eosdis/daacs/nsidc. ASCAT dataset is available at https://user.eumetsat.int/catalogue/EO:EUM:DAT:0656. GLDAS Noah land surface model L4 3 hourly 0.25 degree dataset is available at https://disc.gsfc.nasa.gov/datasets/GLDAS_NOAH025_3H_2.1/summary.
MCD12C1 dataset is available at https://lpdaac.usgs.gov/products/mcd12c1v061. The reconstructed TCA-based error results of Figure 9 are available at Zenodo (https://doi.org/10.5281/zenodo.14850520). Please be aware that if the data is updated, the Zenodo link will change. In that case, please visit hydroai.net or contact the authors directly to download the dataset.

References

Aires, F., P. Weston, P. De Rosnay, and D. Fairbairn. 2021. "Statistical Approaches to Assimilate ASCAT Soil Moisture Information—I. Methodologies and First Assessment." *Quarterly Journal of the Royal Meteorological Society* 147 (736): 1823–1852. https://doi.org/10.1002/qj.3997.

Beaudoing, H., M. Rodell, and NASA/GSFC/HSL. 2020. "GLDAS Noah Land Surface Model L4 3 hourly 0.25 X 0.25 Degree V2.1." Greenbelt, Maryland, USA, Goddard: Earth Sciences Data and Information Services Center (GES DISC). Accessed February 12, 2025. https://doi.org/10.5067/E7TYRXPJKWOQ .

Belgiu, M., and L. Drăguţ. 2016. "Random Forest in Remote Sensing: A Review of Applications and Future Directions." *Isprs Journal of Photogrammetry & Remote Sensing* 114:24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011.

Calvet, J.-C., J.-P. Wigneron, J. Walker, F. Karbou, A. Chanzy, and C. Albergel. 2011. "Sensitivity of Passive Microwave Observations to Soil Moisture and Vegetation Water Content: L-Band to W-Band." *IEEE Transactions on Geoscience & Remote Sensing* 49 (4): 1190–1199. https://doi.org/10.1109/TGRS.2010.2050488.

Dorigo, W. A., W. Wagner, R. Hohensinn, S. Hahn, C. Paulik, A. Xaver, A. Gruber, et al. 2011. "The International Soil Moisture Network: A Data Hosting Facility for Global in situ Soil Moisture Measurements." *Hydrology and Earth System Sciences* 15 (5): 1675–1698. https://doi.org/10.5194/hess-15-1675-2011.

Engman, E. T., and N. Chauhan. 1995. "Status of Microwave Soil Moisture Measurements with Remote Sensing." *Remote Sensing of Environment* 51 (1): 189–198. https://doi.org/10.1016/0034-4257(94)00074-W.

Entekhabi, D., E. G. Njoku, P. E. O'Neill, K. H. Kellogg, W. T. Crow, W. N. Edelstein, J. K. Entin, et al. 2010. "The Soil Moisture Active Passive (SMAP) Mission." *Proceedings of the IEEE* 98 (5): 704–716. https://doi.org/10.1109/JPROC.2010.2043918.

Fisher, R. A., and C. D. Koven. 2020. "Perspectives on the Future of Land Surface Models and the Challenges of Representing Complex Terrestrial Systems." *Journal of Advances in Modeling Earth Systems* 12 (4): e2018MS001453. https://doi.org/10.1029/2018MS001453.

Friedl, M., and D. Sulla-Menashe. 2015. "MCD12C1 MODIS/Terra +aqua Land Cover Type Yearly L3 Global 0.05Deg CMG V006 [Data Set]." NASA EOSDIS Land Processes Distributed Active Archive Center. Accessed February 12, 2025. https://doi.org/10.5067/MODIS/MCD12C1.006.

Gruber, A., W. A. Dorigo, W. Crow, and W. Wagner. 2017. "Triple Collocation-Based Merging of Satellite Soil Moisture Retrievals." *IEEE Transactions on Geoscience & Remote Sensing* 55 (12): 6780–6792. https://doi.org/10.1109/TGRS.2017.2734070.

Gruber, A., C.-H. Su, S. Zwieback, W. Crow, W. Dorigo, and W. Wagner. 2016. "Recent Advances in (Soil Moisture) Triple Collocation Analysis." *International Journal of Applied Earth Observation and Geoinformation* 45:200–211. https://doi.org/10.1016/j.jag.2015.09.002.

Hovmöller, E. 1949. "The Trough-And-Ridge Diagram." *Tellus* 1 (2): 62–66. https://doi.org/10.3402/tellusa.v1i2.8498.

Jovic, A., K. Brkic, and N. Bogunovic. 2015. "A Review of Feature Selection Methods with Applications." *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia: IEEE, 1200–1205. https://doi.org/10.1109/MIPRO.2015.7160458.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. 2017. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree."

Kim, H., M. H. Cosh, R. Bindlish, and V. Lakshmi. 2020. "Field Evaluation of Portable Soil Water Content Sensors in a Sandy Loam." *Vadose Zone Journal* 19 (1): e20033. https://doi.org/10.1002/vzj2.20033.

Kim, H., W. Crow, X. Li, W. Wagner, S. Hahn, and V. Lakshmi. 2023. "True Global Error Maps for SMAP, SMOS, and ASCAT Soil Moisture Data Based on Machine Learning and Triple Collocation Analysis." *Remote Sensing of Environment* 298:113776. https://doi.org/10.1016/j.rse.2023.113776.

Kim, H., and V. Lakshmi. 2018. "Use of Cyclone Global Navigation Satellite System (CyGNSS) Observations for Estimation of Soil Moisture." *Geophysical Research Letter* 45 (16): 8272–8282. https://doi.org/10.1029/2018GL078923.

Kim, H., V. Lakshmi, Y. Kwon, and S. V. Kumar. 2021. "First Attempt of Global-Scale Assimilation of Subdaily Scale Soil Moisture Estimates from CYGNSS and SMAP into a Land Surface Model." *Environmental Research Letters* 16 (7): 074041. https://doi.org/10.1088/1748-9326/ac0ddf.

Kim, H., R. Parinussa, A. G. Konings, W. Wagner, M. H. Cosh, V. Lakshmi, M. Zohaib, and M. Choi. 2018. "Global-Scale Assessment and Combination of SMAP with ASCAT (Active) and AMSR2 (Passive) Soil Moisture Products." *Remote Sensing of Environment* 204:260–275. https://doi.org/10.1016/j.rse.2017.10.026.

Kim, H., J.-P. Wigneron, S. Kumar, J. Dong, W. Wagner, M. H. Cosh, D. D. Bosch, et al. 2020. "Global Scale Error Assessments of Soil Moisture Estimates from Microwave-Based Active and Passive Satellites and Land Surface Models Over Forest and Mixed Irrigated/Dryland Agriculture Regions." *Remote Sensing of Environment* 251:112052. https://doi.org/10.1016/j.rse.2020.112052.

Kingma, D. P., and J. Ba. 2017. Adam: A Method for Stochastic Optimization (No. arXiv:1412.6980). arXiv. http://arxiv.org/abs/1412.6980.

Kwon, Y., S. Jun, E. Kim, K. Seol, S. Hong, I. Kwon, J. Kang, and H. Kim. 2024. "Improving Weather Forecast Skill of the Korean Integrated Model by Assimilating Soil Moisture Active Passive Soil Moisture Anomalies." *Quarterly Journal of the Royal Meteorological Society* 150 (765): 5305–5336. https://doi.org/10.1002/qj.4871.

Lakhankar, T., N. Krakauer, and R. Khanbilvardi. 2009. "Applications of Microwave Remote Sensing of Soil Moisture for Agricultural Applications."

Li, Z.-L., P. Leng, C. Zhou, K.-S. Chen, F.-C. Zhou, and G.-F. Shang. 2021. "Soil Moisture Retrieval from Remote Sensing Measurements: Current Knowledge and Directions for the Future." *Earth-Science Reviews* 218:103673. https://doi.org/10.1016/j.earscirev.2021.103673.

Ma, H., J. Zeng, X. Zhang, J. Peng, X. Li, P. Fu, M. H. Cosh, et al. 2024. "Surface Soil Moisture from Combined Active and Passive Microwave Observations: Integrating ASCAT and SMAP Observations Based on Machine Learning Approaches." *Remote Sensing of Environment* 308:114197. https://doi.org/10.1016/j.rse.2024.114197.

Nguyen, H. H., H. Kim, and M. Choi. 2017. "Evaluation of the Soil Water Content Using Cosmic-Ray Neutron Probe in a Heterogeneous Monsoon Climate-Dominated Region." *Advances in Water Resources* 108:125–138. https://doi.org/10.1016/j.advwatres.2017.07.020.

Nguyen, H. H., H. Kim, W. Crow, S. Yueh, W. Wagner, F. Lei, J.-P. Wigneron, A. Colliander, and F. Frappart. 2025. "From Theory to Hydrological Practice: Leveraging CYGNSS Data Over Seven Years for Advanced Soil Moisture Monitoring." *Remote Sensing of Environment* 316:114509. https://doi.org/10.1016/j.rse.2024.114509.

O'Neill, P. E., S. Chan, E. G. Njoku, T. Jackson, R. Bindlish, and J. Chaubell. 2023. "SMAP L3 Radiometer Global Daily 36 Km EASE-Grid Soil Moisture, Version 9 [Data Set]." Boulder, Colorado USA: NASA National Snow and Ice Data Center Distributed Active Archive Center. Accessed February 12, 2025. https://doi.org/10.5067/4XXOGX0OOW1S.

Peng, C., J. Zeng, K.-S. Chen, Z. Li, H. Ma, X. Zhang, P. Shi, T. Wang, L. Yi, and H. Bi. 2023. "Global Spatiotemporal Trend of Satellite-Based Soil Moisture and Its Influencing Factors in the Early 21st Century." *Remote Sensing of Environment* 291:113569. https://doi.org/10.1016/j.rse.2023.113569.

Petropoulos, G. P., G. Ireland, and B. Barrett. 2015. "Surface Soil Moisture Retrievals from Remote Sensing: Current Status, Products & Future Trends." *Physics and Chemistry of the Earth, Parts A/B/C* 83–84:36–56. https://doi.org/10.1016/j.pce.2015.02.009.

Rodell, M., P. R. Houser, U. Jambor, J. Gottschalck, K. Mitchell, C.-J. Meng, K. Arsenault, et al. 2004. "The Global Land Data Assimilation System." *Bulletin of the American Meteorological Society* 85 (3): 381–394. https://doi.org/10.1175/BAMS-85-3-381.

Seneviratne, S. I., D. Lüthi, M. Litschi, and C. Schär. 2006. "Land–Atmosphere Coupling and Climate Change in Europe." *Nature* 443 (7108): 205–209. https://doi.org/10.1038/nature05095.

Shahriari, B., K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas. 2016. "Taking the Human Out of the Loop: A Review of Bayesian Optimization." *Proceedings of the IEEE* 104 (1): 148–175. https://doi.org/10.1109/JPROC.2015.2494218.

Stoffelen, A. 1998. "Toward the True Near-Surface Wind Speed: Error Modeling and Calibration Using Triple Collocation." *Journal of Geophysical Research Oceans* 103 (C4): 7755–7766. https://doi.org/10.1029/97JC03180.

Wagner, W., S. Hahn, R. Kidd, T. Melzer, Z. Bartalis, S. Hasenauer, J. Figa-Saldaña, et al. 2013. "The ASCAT Soil Moisture Product: A Review of Its Specifications, Validation Results, and Emerging Applications." *Meteorologische Zeitschrift* 22 (1): 5–33. https://doi.org/10.1127/0941-2948/2013/0399.

Wu, K., D. Ryu, L. Nie, and H. Shu. 2021. "Time-Variant Error Characterization of SMAP and ASCAT Soil Moisture Using Triple Collocation Analysis." *Remote Sensing of Environment* 256:112324. https://doi.org/10.1016/j.rse.2021.112324.

Wu, K., D. Ryu, W. Wagner, and Z. Hu. 2023. "A Global-Scale Intercomparison of Triple Collocation Analysis- and Ground-Based Soil Moisture Time-Variant Errors Derived from Different Rescaling Techniques." *Remote Sensing of Environment* 285:113387. https://doi.org/10.1016/j.rse.2022.113387.

Xu, T., and F. Liang. 2021. "Machine Learning for Hydrologic Sciences: An Introductory Overview." *WIREs Water* 8 (5): e1533. https://doi.org/10.1002/wat2.1533.

Yilmaz, M. T., and W. T. Crow. 2014. "Evaluation of Assumptions in Soil Moisture Triple Collocation Analysis." *Journal of Hydrometeorology* 15 (3): 1293–1302. https://doi.org/10.1175/JHM-D-13-0158.1.

Zhao, L., K. Yang, J. Qin, Y. Chen, W. Tang, H. Lu, and Z.-L. Yang. 2014. "The Scale-Dependence of SMOS Soil Moisture Accuracy and Its Improvement Through Land Data Assimilation in the Central Tibetan Plateau." *Remote Sensing of Environment* 152:345–355. https://doi.org/10.1016/j.rse.2014.07.005.

Zou, H., and T. Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society: Series B, Statistical Methodology* 67 (2): 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.