AGU ADVANCING EARTH AND SPACE SCIENCES

Water Resources Research

RESEARCH ARTICLE
10.1029/2023WR034457

Temporal Gap-Filling of 12-Hourly SMAP Soil Moisture Over the CONUS Using Water Balance Budgeting

Runze Zhang[1], Seokhyeon Kim[2], Hyunglok Kim[3], Bin Fang[1], Ashish Sharma[4], and Venkataraman Lakshmi[1]

[1]School of Engineering Systems and Environment, University of Virginia, Charlottesville, VA, USA, [2]Department of Civil Engineering, College of Engineering, Kyung Hee University, Yongin, Republic of Korea, [3]School of Earth Sciences and Environmental Engineering, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea, [4]School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW, Australia

Key Points:

- Twelve-hourly satellite soil moisture (SM) data were gap-filled using a water balance based on SM and precipitation observations
- Gap-filled data had good accuracy and temporal consistency with in situ data and captured SM peaks to heavy rainfall
- Exclusive fill-in SM values exhibited comparable performance to the Soil Moisture Active Passive observations

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

S. Kim,
shynkim@khu.ac.kr

Citation:

Zhang, R., Kim, S., Kim, H., Fang, B., Sharma, A., & Lakshmi, V. (2023). Temporal gap-filling of 12-hourly SMAP soil moisture over the CONUS using water balance budgeting. *Water Resources Research*, *59*, e2023WR034457. https://doi.org/10.1029/2023WR034457

Received 8 JAN 2023
Accepted 15 NOV 2023

Author Contributions:

Conceptualization: Runze Zhang, Seokhyeon Kim, Ashish Sharma
Data curation: Runze Zhang, Seokhyeon Kim, Hyunglok Kim
Formal analysis: Runze Zhang
Funding acquisition: Seokhyeon Kim
Investigation: Runze Zhang
Methodology: Runze Zhang, Seokhyeon Kim, Bin Fang, Ashish Sharma, Venkataraman Lakshmi
Resources: Runze Zhang, Seokhyeon Kim
Software: Runze Zhang, Seokhyeon Kim
Supervision: Seokhyeon Kim, Venkataraman Lakshmi
Validation: Runze Zhang

Abstract

Temporal gaps in satellite-based soil moisture (SM) products are a persistent issue. This study presents an entirely observation-based method to derive volumetric SM content for filling gaps in Soil Moisture Active Passive (SMAP) retrievals. Using a water balance equation, 12-hr topsoil water amount variations are determined based on observed precipitation from the Global Precipitation Measurement Mission (inflow) and a hydrologic loss function (outflow) built on SMAP dry-downs. A temporally seamless SM product, composed of SMAP dry-downs and precipitation-driven moisture approximations, was generated as a secondary outcome in determining optimal water balance parameters. This data set maintains the original SMAP SM dynamics with a median Pearson correlation (R) of 0.69 and an unbiased root-mean-square error (ubRMSE) of 0.05 m^3/m^3. Using these parameters and available SMAP observations, a 12-hourly SM product was produced over the conterminous United States. Validated against in situ measurements, this 12-hourly SM product exhibits good performance with a median R of 0.63 and captures most SM peaks induced by heavy rainfall. A time series examination revealed the produced 12-hourly SM product closely corresponds to in situ SM variations and outperforms two other SMAP-based 12-hourly SM products gap-filled using temporal linear interpolation and a three-dimensional smoothing approach, especially during sparse SMAP data periods. The proposed scheme's validity is further verified by the comparable performance of the exclusive filled-on SM estimates. Utilizing the 12-hourly SM data set and its paired hydrologic losses could enhance the quantification connections among the hydrologic components and benefit the understanding of land-surface hydrology.

Plain Language Summary

The use of satellite-based soil moisture (SM) data can be affected by missed observations, which can make it difficult to accurately interpret processes and applications in Earth system science. To fix this, we need a continuous data set that fills in the gaps with reasonable estimates. In this study, we developed a method using observed data to fill gaps in Soil Moisture Active Passive (SMAP) data. We used a simplified water balance equation to estimate the amount of water in the top layer of soil over a 12-hr period based on observed precipitation and information about how water is lost through the soil. We created a continuous SM product called the SMAP-based 12-hourly SM product for the conterminous United States, which performed well when compared to in situ measurements and captured most SM peaks caused by heavy rainfall. Having continuous SM data and information about how water is lost through the soil can help us better understand land-surface hydrology.

1. Introduction

Soil moisture (SM) is a variable of great importance in its capability of influencing land-atmosphere interactions and its critical role in the hydrologic cycle (Koster et al., 2004, 2010; Petropoulos, 2013; Seneviratne et al., 2010). Satellite-based microwave remote sensing has emerged as a reliable tool for monitoring temporal variations of surface SM at a global scale, striking a balance between measurement accuracy, geographical coverage, and cost-effectiveness. The utilization of SM observations from various space-borne microwave sensors has greatly improved our understanding of Earth's systems, including climate variability, drought detection, water resource management, and agricultural monitoring (Fang et al., 2021; Findell et al., 2011; Ge et al., 2011; Koster et al., 2010; Miralles et al., 2014; Taylor et al., 2012). This progress has been facilitated by the availability of high-resolution spatial SM information from active microwave sensors like Advanced Scatterometer (ASCAT) and Sentinel-1 under the Copernicus Service, as well as fine-temporal-resolution SM variations from passive

Visualization: Runze Zhang, Hyunglok Kim
Writing – original draft: Runze Zhang
Writing – review & editing: Runze Zhang, Seokhyeon Kim, Hyunglok Kim, Bin Fang, Ashish Sharma, Venkataraman Lakshmi

microwave missions such as Soil Moisture Active Passive (SMAP) and soil moisture ocean salinity (SMOS) (Balenzano et al., 2021; Entekhabi et al., 2014; Mecklenburg et al., 2016; Wagner et al., 2013). Among them, the exceptional performance of SMAP SM retrievals has been extensively validated by rigorous studies and thus SMAP data have been applied in various hydrologic applications (Chan et al., 2018; Colliander et al., 2017; Ma et al., 2019; R. Zhang et al., 2021).

SMAP SM product has significantly advanced our understanding of SM dynamics at a global scale and offers a temporal resolution that is well-suited for a variety of applications, including climate modeling and agricultural monitoring. However, for certain applications, there is a growing demand for SM data at a higher temporal resolution (sub-daily) and finer spatial scale (kilometer scale). For example, the need for sub-daily SM data sets has been emphasized in studies such as Peng et al. (2021), which highlight their importance in hydrological modeling and numerical weather prediction. Additionally, surface SM exhibits a marked diurnal variation, and understanding the dynamics of diurnal SM is essential for enhancing our fundamental interpretation of the evaporation process (R. D. Jackson, 1973). Research focused on the diurnal SM cycle typically requires hourly SM information (T. J. Jackson et al., 1997). Moreover, there is a particular need for SM data closely neighboring extreme rainfall events to improve our understanding of the relationship between precipitation and flooding, thus enhancing our capability to predict and mitigate flooding events (Sharma et al., 2018; Wasko & Sharma, 2017).

In general, satellite-based SM retrievals are presented as a time series with a regular time step (such as 12-hourly or daily), and any gaps resulting from the revisit return time are represented as "NaN" values. For instance, the SMAP level 3 descending and ascending products can be aggregated into a 12-hourly data set but with a number of temporal gaps. To fill the temporal voids in those data sets and ensure a more complete representation of SM over time, various methods have been developed, primarily categorized into two groups. The first category fills the gaps in SM data sets by analyzing the temporal and/or spatial patterns observed in the available measurements. Simple techniques like linear interpolation (LIP) are commonly used to estimate SM values between two consecutive observations. Advanced statistical modeling techniques, such as the three-dimensional optimization underlying the discrete cosine transforms, have been employed to achieve temporally seamless and even spatially complete SM records (ElSaadani et al., 2021; D. Kim et al., 2016; Pham et al., 2019; G. Wang et al., 2012; Q. Zhang et al., 2021). In addition to analyzing SM patterns alone, many geostatistics and machine learning techniques have been utilized to establish relationships between SM and other relevant geophysical variables, such as brightness temperature, precipitation, soil texture, and temperature (Almendra-Martín et al., 2021; Y. Cui et al., 2019; Y. Liu et al., 2022; Llamas et al., 2020; Tong et al., 2021; Xiao et al., 2016). These techniques leverage the available data of non-SM variables, which often have a higher temporal resolution than SM retrievals, to approximate SM values in the gaps. However, it is important to note that gap predictions based on spatial or temporal modes from existing retrievals may inherit the errors of the input products and often require computationally intensive processing. The second category merges information from multiple satellite platforms to enhance the actual temporal resolution of the SM data set, such as the European Space Agency Climate Change Initiative (ESA CCI) SM data set and the SMOSSMAP-IB product (Dorigo et al., 2017; S. Kim et al., 2021; Li et al., 2022). Another satellite mission called Cyclone Global Navigation Satellite System (CYGNSS), which utilizes L-band GNSS signals of opportunity, has been employed to retrieve SM with a higher temporal resolution (i.e., sub-daily). Although CYGNSS SM retrievals show comparable performance with SMAP data, they are limited in terms of irregular acquisition times and spatial coverage (38°N to 38°S) (H. Kim & Lakshmi, 2018; Ruf et al., 2018).

This research is dedicated to the creation of a process-based water balance scheme, designed to fill the temporal gaps in the SMAP data record, which are resulted from the satellite's revisit frequency. This water balance scheme, inspired by Koster et al. (2017), consists of two main steps. Initially, the integrated hydrologic loss encompassing evaporation and drainage, is estimated using the antecedent SM data. Subsequently, the unknown SM value in the next temporal gap is derived by subtracting the estimated loss from precipitation within a specific temporal interval. Modifications to this loss function (developed in the first step) were previously made by Akbar et al. (2018), who applied the adapted model to estimate the hydrological lengths (ΔZ) across the conterminous United States (CONUS). Specifically, they constructed the loss function using a locally weighed linear smoother (LOWESS) technique with a fixed span of 65% to regress the observed loss rates to SMAP dry downs. However, the use of a constant span could overlook the broad range of observed losses at similar SM magnitudes and may not adequately reflect the pixel-specific characteristics. To address these issues, this study further refines the loss

Table 1
Summary of Data Sets Used in This Study

Variable	Product name	Spatial resolution	Unit	Reference
Volumetric Soil Moisture	SMAP L3 enhanced soil moisture product (version 5)	9 km	m^3/m^3	O'Neill, Chan, et al. (2021)
	ISMN	Point	m^3/m^3	Dorigo et al. (2021)
	TxSON	Point	m^3/m^3	Caldwell et al. (2019)
	ERA5-Land 0–7 cm soil moisture	0.1°	m^3/m^3	Muñoz-Sabater (2019)
Precipitation	GPM IMERG half-hourly final-run Level-3 precipitation product (version 06B)	0.1°	mm/hr	Huffman et al. (2019)
	NLDAS Primary Forcing L4 Hourly Precipitation (version 2.0)	0.125°	$kg/m^2/hr$	Xia et al. (2012)

function algorithm by integrating quantile regression with a pixel-specific parameter. This enhancement aims to provide more accurate sub-daily filled-in SM estimations for the 12-hourly SMAP product.

The process-based water balance model developed in this study falls under the first type gap-filling method mentioned earlier. However, it distinguishes itself from other approaches in this category by its simplicity and independence from purely numerical relationships with other SM-related variables, which can be challenging to interpret. Additionally, the proposed water balance model can approximate SM data at different temporal scales within a single simulation round, offering significant flexibility. In this study, the Global Precipitation Measurement Mission (GPM) product with a temporal resolution of 30 min was utilized as the precipitation input. This high-resolution temporal data allows for the generation of SM simulations at various time scales, ranging from sub-hourly to yearly intervals, thereby accommodating a wide range of research needs. However, the focus of this study is on a fixed 12-hourly interval. This choice aligns well with the time difference between the SMAP descending (6 a.m.) and ascending (6 p.m.) overpasses and takes advantage of the 12-hr GPM product's ability to capture the extreme rainfall events more effectively than its finer-temporal-scale ones (Mazzoglio et al., 2019). It is important to note that the scope of this study does not include spatial gap-filling.

Our analyses focus on CONUS due to its diverse climatic regimes, vegetation conditions, and the availability of numerous ground-based SM measurement stations. Therefore, it provided a suitable setting for evaluating the performance of the continuous SM estimates derived from our process-based water balance scheme by comparing them against in situ benchmarks. To the best of our knowledge, this study represents the first attempt to conduct a direct comparison between the precipitation-driven SM data derived from the proposed scheme and ground-based measurements. In addition to employing conventional validation metrics, this study assessed the capability of the SM product obtained by the proposed algorithm to accurately capture rainfall-induced SM peaks. Again, these immediate positive responses in SM time series play a crucial role in disaster prediction and hydrologic applications (Peng et al., 2021; Sharma et al., 2018).

This paper is organized as follows. In Section 2, the data sets and the corresponding preprocessing methodologies are presented. Following that, Section 3 describes the specific procedures used to forward simulations of SM. It also outlines the assessment strategies implemented to evaluate the accuracy of the gap-filled SM data sets. The results and discussion of the findings are presented in Section 4. Finally, conclusions followed by a summary are provided in Section 5.

2. Data

As summarized in Table 1, various data sets covering the period of 7 years (1 April 2015–31 March 2021) have been adopted in this study. These include (a) the National Aeronautics and Space Administration (NASA) SMAP SM product, (b) the reanalysis SM data set of the land component of the fifth generation of European Re-Analysis (ERA5-Land) developed by the European Center for Medium-Range Weather Forecast (ECMWF) (Muñoz-Sabater, 2019), (c) in situ SM measurements from 1,084 stations of the International Soil Moisture Network (ISMN) (Bell et al., 2013; Caldwell et al., 2019; Cook, 2016; Dorigo et al., 2013, 2021; Larson et al., 2008; Leavesley et al., 2008; Moghaddam et al., 2010; Ojo et al., 2015; Osenga et al., 2019; Schaefer et al., 2007) (see Table S1 in Supporting Information S1 (hereafter Figures and Tables in Supporting Information S1 use the prefix "S")), (d) in situ SM measurements from 40 stations of the Texas Soil Observation Network (TxSON) (Caldwell et al., 2019), (e) half-hourly precipitation estimates from the final-run Integrated Multi-SatellitE Retrievals for GPM IMERG (Hou et al., 2014), and (f) hourly precipitation data from the North American Land Data Assimilation System (NLDAS) (Xia et al., 2012).

Given the objective of building a 12-hourly continuous SM data set at a Coordinated Universal Time (UTC) timescale, the local-time-based SMAP retrievals were interpolated into the closest time slots of UTC 00:00 and 12:00 using the nearest neighboring method. This adjustment process was constructed on a crucial hypothesis of invariable SM within a (±) 3-hr interval. Accordingly, the hourly available ERA5-Land and the in situ SM data synchronized with the UTC 00:00 and 12:00 were extracted for validation. After temporal processing, the gridded GPM IMERG and ERA5-Land products were then resampled into the Equal-Area Scalable Earth (EASE) 9-km scale to be compatible with the SMAP spatial resolution.

2.1. SMAP Soil Moisture

The SMAP mission was launched on 31 January 2015, by NASA for quantifying the representative water content at the top 5 cm of the full soil column and detecting freeze/thaw states at a quasi-global scale (Entekhabi et al., 2014). The SMAP sensor crosses the equator constantly at around 6 a.m. and 6 p.m. (local solar time) and monitors SM variations with a revisit frequency of 2–3 days (O'Neill, Bindlish, et al., 2021). In order to satisfy the research requirements of hydrometeorology and hydroclimatology, SMAP originally intended to incorporate the attributes of active and passive microwave sensors to provide high-resolution SM retrievals. However, the malfunction of the SMAP radar in July 2015 hampered the initial goals. Alternatively, the Backus-Gilbert optimal interpolation technique is adopted on the oversampled measurements of the SMAP radiometer to derive an enhanced SM product posted at the 9-km EASE grids (O'Neill, Bindlish, et al., 2021).

In this study, the SMAP Enhanced Level-3 Radiometer Global Daily 9-km EASE-Grid SM (version 5) product (hereinafter referred to as SMAP) has been selected. An integrated consideration of both geographical coverage and the quality of the filled-in SM estimations, a series of filtering procedures have been adopted. Specifically, the regions of vegetation water content (VWC) below 7 kg/m^2 are retained to include more areas of eastern CONUS (Akbar et al., 2018). It should be noted that this VWC of 7 kg/m^2 is less restricted relative to the recommended threshold of 5 kg/m^2 (O'Neill, Bindlish, et al., 2021). In contrast, a more rigid water fraction threshold of lower than 1% of water bodies within each 9-km pixel has been applied. Moreover, the pixel-wise soil porosity value (Φ) was computed using the bulk density (BD) under the SMAP ancillary data set, following the equation $\Phi = 1 - \mathrm{BD}/2.65$ (Das & O'Neill, 2020). These derived Φ values are essential for constructing the loss functions discussed in the later part of the paper (Section 3.2).

2.2. In Situ Soil Moisture Measurements

In situ measurements are often acknowledged as the most reliable source and widely performed as benchmarks to assess remotely sensed SM retrievals and model estimations. The ISMN (Dorigo et al., 2013, 2021), as a centralized data platform, regularly compiles and reconciles ground SM observations from different networks all over the world. For the purpose of ensuring a rigorous assessment, in situ observations that were not symbolized as "good" quality or exhibited a measuring depth beyond 10 cm were filtered out. Meanwhile, stations presenting fewer than 60 effective SM samples were omitted to uphold the statistical significance. Although the evaluation results based on the sparse networks tend to be slightly inferior to those obtained from the core validation sites, they are still of high value due to the wide geographical coverage (Chan et al., 2018; R. Zhang et al., 2019).

If the average assessment metrics were calculated using all the metrics obtained by comparing SM retrievals against all the stations, there is a risk of placing excessive weight on pixels with multiple stations. To mitigate this issue, a strategy was implemented to select only one station as the representative for each pixel. The selection process followed the steps outlined in Dorigo et al. (2015). Here, for each station (STN_x) within a pixel, the Pearson Correlation (R) values between the SMAP/ERA5-Land and STN_x were separately computed. The averaged R values for R [SMAP, STN_x] and R [ERA5-Land, STN_x] were compared, and the station with the highest mean R value was chosen as the pixel representative.

2.3. GPM IMERG Half-Hourly Final-Run Precipitation

The GPM IMERG half-hourly final-run precipitation Level-3 product (Version 6B) has been applied to provide the water inputs to the upper soil system. It combines observations from multiple satellite missions, including the GPM Core Observatory, along with other passive microwave and infrared sensors, to provide accurate and detailed estimates of precipitation rates and accumulation (Huffman et al., 2019). With a temporal resolution of half-hourly intervals and a spatial resolution of 0.1°, the GPM IMERG final-run product enables study of fine-scale precipitation patterns and variability on a quasi-global scale. To ensure accuracy, the final-run product incorporates measurements from ground-based gauges, albeit with a delayed release of approximately 3.5 months after the observation month.

As our study focuses on a 12-hr time step, the 30-min Level 3 GPM IMERG final-run (hereinafter referred to as GPM) precipitation volumes were aggregated for the time intervals: UTC 00:00 to 12:00 and UTC 12:00 to 00:00. These time slots correspond to local time 18:00 to 06:00 and local time 06:00 to 18:00 in the central CONUS, between the SMAP ascending and descending overpass moments. Subsequently, the GPM precipitation was re-gridded to the 9 km EASE 2.0 grid projection before extracting the SMAP dry-down SM values. By aligning the study period with the overlapping data range between the GPM and SMAP products and a complete annual cycle, we established the study period from 1 April 2015, to 31 March 2021.

2.4. NLDAS Primary Forcing L4 Hourly Precipitation

In this study, the NLDAS hourly precipitation data were utilized to evaluate the model's ability to capture rainfall-induced SM peaks (Xia et al., 2012). The NLDAS precipitation data, included as an independent source, could differ from the GPM precipitation used in identifying rainfall occurrences. The precipitation field in "File A" corresponds to a temporal disaggregation of a gauge-only Climate Prediction Center (CPC) analysis of daily precipitation, adjusted for orographic effects using the widely applied Parameter-elevation Relationships on Independent Slopes Model (PRISM) climatology. The temporal disaggregation process involves deriving hourly weights from WSR-88D Doppler radar-based precipitation estimates, 8-km CPC MORPHing technique (CMORPH) hourly precipitation analyses, or the Northern American Regional Reanalyzes (NARR)-simulated precipitation, in order of availability. For pixels with in situ SM stations, the NLDAS hourly precipitation was extracted and aggregated into a 12-hr precipitation data set. The specific utilization of this 12-hourly NLDAS precipitation is further discussed in Section 3.3 of this paper.

3. Methods

3.1. Estimation of the Integrated Hydrologic Loss

The hydrologic process taking place in the topsoil layer can be simplified into three main components: precipitation (input), hydrologic loss (output), and changes in SM (change in storage) (Equation 1). Given that the water inputs are known in advance, it is crucial to accurately quantify the loss rate (Q) to ensure the reliability of the derived SM.

$$\Delta Z \cdot \frac{\mathrm{SM}_{t+\Delta t} - \mathrm{SM}_t}{\Delta t} = P(t \sim t + \Delta t) - Q(t \sim t + \Delta t) \tag{1}$$

where SM_t and $\mathrm{SM}_{t+\Delta t}$ (m^3/m^3) represent volumetric SM at time points t and $t + \Delta t$, respectively. Δt (day) is the time interval. ΔZ (mm) denotes the hydrologic depth within which SM data have similar dynamics. P and Q are the precipitation (mm/day) and loss rates (mm/day) during the time interval.

Akbar et al. (2018) showed that Q can be expressed as the product of ΔZ and the volumetric loss rate (L) (Equation 2). Additionally, L can be approximately estimated using the SMAP dry-down SM.

$$Q = \Delta Z \cdot L \tag{2}$$

where L is the loss rate in the volumetric unit (m^3/m^3/day).

The SMAP observed dry downs are identified as consecutive periods of decreasing SM. Each dry-down period is defined by at least three consecutive SMAP observations with no recorded precipitation in between (Akbar et al., 2018). SMAP dry-down SM data were identified and collected in a separate set (Figures 1a and 1b). Then, the daily rate of L for each dry-down period is estimated using Equation 3. It is important to note that the loss rates derived from Equation 3 are solely based on observed SM retrievals, operating under the assumption that the variations in SM during non-raining conditions are indicative of loss quantifications. Despite the term loss

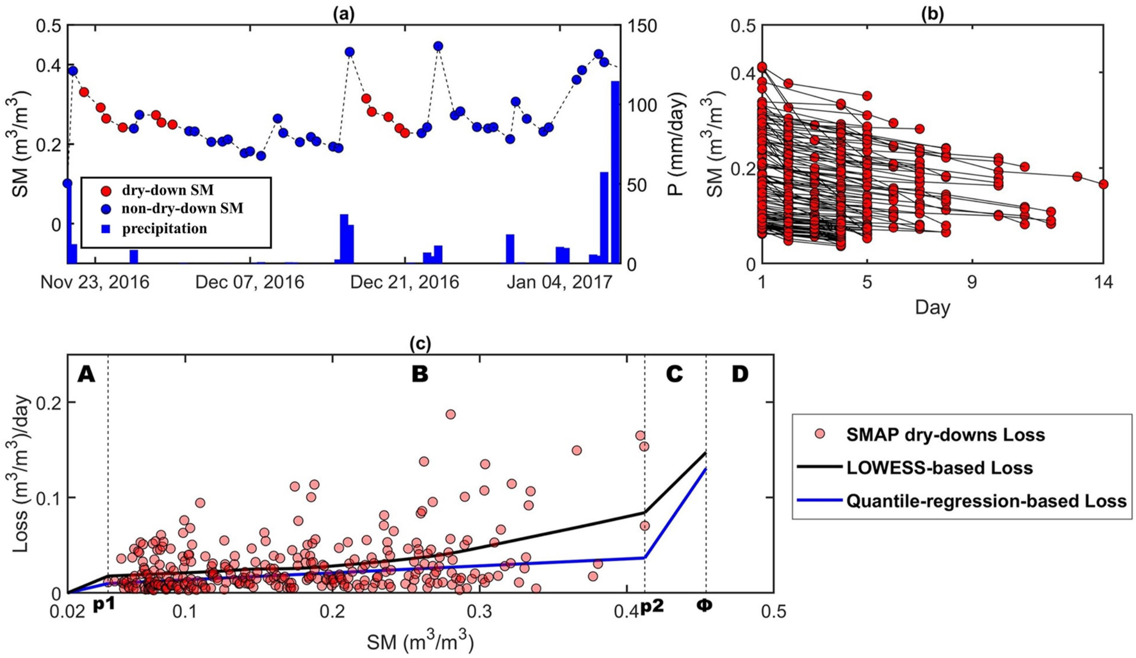

Figure 1. (a) An example of the Soil Moisture Active Passive (SMAP) soil moisture (SM) and the GPM precipitation time series for a pixel centered on (36.25°N, 120.58°W) from 1 November 2016, to 1 January 2017. (b) All the SMAP SM dry-downs are stacked together; (c) The quantitative relationship between the dry-down SM and its paired loss. Regression analyses, including the LOWESS with a constant span of 65% (black line) and the quantile regression with β (blue line), have been performed within the SMAP dry-downs range (Segment B). Lines in Segments A and C are linearly extended along the line of Segment (b). Again, p_1 and p_2 denote the minimum and maximum SMAP dry-down SM while Φ is the soil porosity.

encapsulating various hydrological processes including drainage, runoff, and evapotranspiration, stage II evapotranspiration is typically the principal influence of SM fluctuations during most dry-down phases (McColl, Wang, et al., 2017).

$$L_{\mathrm{obs}} = -\frac{\mathrm{SMAP}_{dd}(t + \Delta t_{\mathrm{obs}}) - \mathrm{SMAP}_{dd}(t)}{\Delta t_{\mathrm{obs}}} \tag{3}$$

where the subscripts obs and *dd* indicate the observation-based derivations, and the SMAP dry-down SM measurements, respectively.

When the dry-down SM data are not available, it becomes impractical to estimate the integrated loss using Equation 3. Additionally, relying on an intermittent product of L is insufficient to maintain a continuous cycle of the water balance equation described in Equation 1. To address this issue, a simple solution is to assume that the value of L over a given time interval is a function of the initial and/or final SM values within that period (Akbar et al., 2018; Brocca et al., 2019; Koster et al., 2018). In this study, we presumed that the estimated loss rate (L_{est}) was predominately governed by the initial SM (SM_t) of a given time slot (Equation 4), and this L_{est} remains consistent over a 12-hr period.

$$L_{\mathrm{est}}(t \sim t + 0.5) = f(\mathrm{SM}_t) \tag{4}$$

where $L_{\mathrm{est}}(t \sim t + 0.5)$ represent the estimated volumetric water loss between t and 0.5 days after t. $f()$ displays the regressed relationship between the loss and initial SM.

The quantitative relationship between L and SM_t can be determined by regressing the available L_{obs} (calculated using Equation 3) to SM_t of the corresponding dry-down limbs. Here, a dry-down limb refers to a subsection of a dry-down SM time series, consisting of two consecutive decreasing SM retrievals. Specifically, two regression methods, namely LOWESS with a span of 65% (Akbar et al., 2018) and the quantile regression approach (Magan et al., 2020; Wasko & Sharma, 2014) with a performance-driven percentile (β) have been adopted and compared.

Since Equation 4 was constructed based on the SMAP dry-downs, it is assumed to be more suitable when SM falls within the minimum (p_1) and maximum (p_2) of SMAP dry-downs. Given the broad range of SM, four distinct formulas have been separately employed for four SM_t domains: [0.02, p_1] (Segment A), [p_1, p_2] (Segment B), [p_2, porosity (Φ)] (Segment C), and values above Φ (Segment D). When the SM_t lies between [0.02, p_1], the loss function defined for Segment A linearly interpolates between the points [0.02, $L_{est}(0.02)$] and [p_1, $L_{est}(p_1)$], with $L_{est}(0.02)$ presumed to be 0 m^3/m^3/day. Similarly, the loss function defined for Segment C linearly interpolates between [p_2, $L_{est}(p_2)$] and [Φ, $L_{est}(\Phi)$]. Given the $L_{est}(\Phi)$ is unknown, a slope parameter α has been introduced, the value of which can be determined through an optimization search in the forward simulation displayed in Section 3.2. Once the initial SM values reach saturation (Segment D), the excessive water majorly owing to runoff and drainage is integrated into the L_{est} for the next 12-hr interval.

$$L_{\text{est}}(t \sim t+0.5) = \begin{cases} (\text{SM}_t - 0.02) \cdot \frac{L_{\text{est}}(p_1) - L_{\text{est}}(0.02)}{(p_1 - 0.02)} & 0.02 < \text{SM}_t < p_1 \quad \text{Segment A} \\ f(\text{SM}_t) & p_1 \le \text{SM}_t \le p_2 \quad \text{Segment B} \\ L_{\text{est}}(p_2) + \alpha \cdot (\text{SM}_t - p_2) & p_2 < \text{SM}_t \le \Phi \quad \text{Segment C} \\ \frac{P(t \sim t+0.5)}{\Delta Z} + \frac{(\text{SM}_t - \Phi)}{\Delta t} & \text{SM}_t > \Phi \quad \text{Segment D} \end{cases} \tag{5}$$

where p_1 and p_2 are the maximum and minimum SMAP dry-down SM, and Φ is the porosity.

The full form of the loss function has been illustrated in Figure 1c where the loss amounts monotonically elevate with the increase of SM. Within an acceptable range of SM values, spanning from 0.02 m^3/m^3 to Φ, the sensitivity of loss values is more pronounced at the dry and wet extremes than at median SM levels (Koster et al., 2018; Salvucci, 2001).

3.2. Forward Simulation of Rainfall-Driven Soil Moisture

Given that L has been described as a function of SM_t, $SM_{t+0.5}$ can be predicted as long as P and SM_t are available (Equation 6). However, there are still two parameters, ΔZ and α, that remain to be determined. A third parameter, β, will also be required if the quantile regression is used to build the function in Segment B. The optimal values of these parameters were obtained by minimizing the root-mean-square error (RMSE) between the precipitation-driven SM estimations and SMAP retrievals.

$$\text{SM}_{t+0.5} = \text{SM}_t + 0.5 \cdot \left[\frac{P(t \sim t+0.5)}{\Delta Z} - L_{\text{est}}(t \sim t+0.5) \right] \tag{6}$$

Using Equation 6, SM can be continuously simulated by inserting one initiated value for SM at the beginning of the study period. Simulated SM products in this way are expressed as PLO (P: Precipitation-driven simulation + L: "LOWESS" regression + O: One initiated SM), and PQO (P: Precipitation-driven simulation + Q: Quantile regression + O: One initiated SM).

However, the precipitation-reconstructed SM estimates are prone to suffer from the errors, possibly resulting in a large deviation between the predicted and observed SM in the late or particular stage of simulation. These errors could be sourced from the occasional mismatching between rainfall and SM data sets as well as from the inappropriate derivations of hydrologic losses. Given this, SMAP dry downs were incorporated to displace their concurrent SM_t in Equation 6 to proceed the precipitation-driven process. This reset segments the simulation process into discrete intervals bounded by SMAP consecutive dry downs, aiming to mitigate the cumulative simulation errors encountered by the PLO over a 6-year span. A preliminary analysis (not shown here) revealed that the PLO SM estimates exhibited excessively rapid drying rates. The incorporation of SMAP dry downs into the simulation procedure has shown promise in ameliorating this issue. Therefore, the SMAP dry-downs could also be labeled as correction points or as lifting-up measures. The number of SMAP dry-down SM is around 12% of the total SMAP observations (4% of 12-hr continuous SM simulations over the 6-year period). Thus, the precipitation-driven SM products including the original SMAP dry downs are entitled as PLD (PL + D: Dry-down SM) and PQD (PQ + D: Dry-down SM), respectively.

Finally, PQF (PQ + F: gap-Filled) was generated to supplement the missed SM at a 12-hourly scale of the combined descending and ascending SMAP L3 product. Similar to PQD, the simulated SM was immediately displaced by the simultaneous SMAP observation. As the optimization objective is to minimize the RMSE between the simulated SM data and their temporally paired SMAP observations, the benchmark data set and the optimization process will become ineffective when all the SMAP data are involved in the simulation. Hence, PQF was derived using the optimal parameters obtained from the PQD, and a rerun of Equation 6. PQF is the ultimate product conforming to the major purpose here that solely fills the SMAP temporal gaps. A summary of different simulated SM products is described in Table S2 in Supporting Information S1.

The selection of the optimal methodology flow from those described above (i.e., PLO, PLD, and PQD) was based on the simulation results from the representative 9-km pixels. In order to cover a wide variety of land surface conditions, the CONUS matrix (285 × 644) composed of 9-km grids was first divided into 15 × 28 coarse-scale blocks. Each block encompassed 19 × 23 9-km pixels, and then a pixel was randomly selected from each block. Before the random procedure, the pixels with dry-down limbs of fewer than 50 and the blocks of fewer than 10 effective pixels were excluded. As a result, 260 pixels out of 420 blocks were determined and extracted as the representative points. By conducting the simulations within the representative pixels instead of the entire CONUS, the computational time was significantly reduced. Although there is some risk associated with determining the optimal methodology flow based on 260 grids, the results were found to be consistently stable when conducting multiple tests using different random pixels.

3.3. Detection of Rainfall-Induced Soil Moisture Peaks

Understanding rainfall-induced SM peaks and their underlying causes is crucial for characterizing SM dynamics. Figure S1 in Supporting Information S1 illustrates the methodology used to examine alignment between the rainfall-induced SM peaks detected by in situ SM measurements and SM estimations derived from precipitation data (e.g., PQD). The process consists of four main steps. First, consecutive NLDAS data with precipitation rates above 0.5 mm/day were grouped separately to represent distinct rainfall events. To illustrate, consider a sequence of six NLDAS data spanning from 12:00:00 on Day 1 to 00:00:00 on Day 4, each registering rates surpassing the 0.5 mm/day threshold. The ensuing observation at 12:00:00 on Day 4 recorded a rate of 0 mm/day. In this context, NLDAS timestamps, whether at 12:00:00 or 00:00:00 on Day X, represent averaged precipitation rates for the preceding 12-hr interval. Given this scenario, the contiguous set of the first six observations would be grouped as a single, independent rainfall event. The timestamps of each independent rainfall event were subsequently logged. These documented times were collected as the precipitation-based anticipated times for SM peaks. As those precipitation-driven SM data sets had been derived based on the GPM observation, an independent NLDAS product was thus incorporated here to avoid overestimating their peak-capturing capacity. Second, a peak detection algorithm (MathWorks, 2020) was employed to identify the time indexes of SM peaks in in situ and simulated SM time series. A peak was determined if the corresponding value was higher than the points before and after it. The time indexes of peaks identified by the aforementioned peak detection algorithm were recognized as the SM-derived expected timestamps for SM peaks. Subsequently, the overlap between the precipitation-based anticipated times for SM peaks and the SM-derived expected timestamps for SM peaks was examined, and the peaks at these overlapping times were defined as rainfall-induced SM peaks. Following the time order of independent rainfall events, the binary array was separately generated for in situ SM product and the precipitation-driven SM data sets, indicating the presence or absence of rainfall-induced SM peaks. Herein, independent precipitation events with rainfall-induced SM peaks were denoted as 1, while events without rainfall-induced SM peaks were labeled as 0. Finally, the binary array derived from the in situ SM measurements served as the reference and was juxtaposed against binary arrays from precipitation-driven SM data sets. This comparison is done by calculating categorical metrics, as described in Section 3.4, to evaluate the agreement between the observed and simulated rainfall-induced SM peaks.

3.4. Statistical Metrics

Statistical metrics are required to reflect the quality and accuracy of different data sets. Here, the quantitative metrics, that is, unbiased RMSE (ubRMSE) and R, were adopted to describe the discrepancies in magnitudes and temporal correlations between the precipitation-based SM and the benchmark data sets. For purposes of capturing rainfall-induced SM peaks, three categorical scores, that is, the probability of detection (POD), false alarm ratio (FAR), and critical success index (CSI) have been computed. The formulas of all the statistical metrics are presented below.

$$\text{ubRMSE} = \sqrt{E\left[(\theta_{\text{sim}} - \theta_{\text{ref}})^2\right] - E[(\theta_{\text{sim}} - \theta_{\text{ref}})]^2} \tag{7}$$

$$R = \frac{E[(\theta_{\text{sim}} - E[\theta_{\text{sim}}])(\theta_{\text{ref}} - E[\theta_{\text{ref}}])]}{\sigma_{\text{sim}}\sigma_{\text{ref}}} \tag{8}$$

$$\text{POD} = \frac{H}{H + M} \quad \text{range } [0, 1], \text{optimal: } 1 \tag{9}$$

$$\text{FAR} = \frac{F}{H + F} \quad \text{range } [0, 1], \text{optimal: } 0 \tag{10}$$

$$\text{CSI} = \frac{H}{H + F + M} \quad \text{range } [0, 1], \text{optimal: } 1 \tag{11}$$

where $E[]$ represents the expected value; θ_{sim} and θ_{ref} denote the SM from simulation product and reference data set; σ_{sim} and σ_{ref} refer to the standard deviations of simulated and referenced SM data; H is the number of rainfall-induced SM peaks simultaneously detected by a simulated data set and in situ measurements; F is the number of rainfall-induced SM peaks detected by a simulated data set but not observed by in situ measurements; and M is the number of rainfall-induced SM peaks observed by in situ measurements but not detected by the simulated product.

3.5. Penalized Least Square Regression Based on Three-Dimensional Discrete Cosine Transform

An additional gap-filling method, a penalized least square regression based on three-dimensional discrete cosine transform (DCT), has also been adopted to complement the SMAP gaps at a 12-hourly temporal step. Then, the magnitudes and dynamics of its filled-on SM were analyzed and compared with those SM estimates derived from the above process-based model by validation against the in situ observations. The DCT was originally proposed by Garcia (2010) to smooth multidimensional incomplete data, and G. Wang et al. (2012) extended this technique for the application of complementing spatiotemporal gaps for large-scale geophysical data sets. Specifically, the penalized least square regression (PLS) is targeted to find a model that fits the data well by minimizing the squared difference between the modeled and actual data while adding a penalty term to avoid overfitting. Thus, the objective function of the PLS purses the least summation of the squared difference and the penalty term (i.e., Equation 12).

$$F(\hat{D}) = \left\| W^{\frac{1}{2}} \circ (\hat{D} - D) \right\|^2 + s\left\| \nabla^2 \hat{D} \right\|^2 \tag{12}$$

where D and $\hat{D}$ represent the three-dimensional SM arrays from the SMAP product and the SMAP-derived data set without missing values, respectively; W is a binary array of the same size of D to identify whether SMAP data are available or not; $\|\ldots\|$ is the Euclidean norm, while $\circ$ and ∇^2 denote the Schur product and the Laplace operator; s is a positive scalar reflecting the smooth degree.

The discrete cosine transformation can convert the multidimensional data into a group of cosine functions with different frequencies and phases, and the PLS regression can be conducted on the data set transformed using the discrete cosine transformation. Hence, the $\hat{D}$ can be obtained via:

$$\hat{D} = \text{IDCT}\left(\Gamma \circ \text{DCT}\left(W \circ (D - \hat{D}) + \hat{D}\right)\right) \tag{13}$$

where IDCT and DCT mean the inverse discrete cosine transformation and discrete cosine transformation conversions; Γ represents a three-dimensional filtering tensor here.

$$\Gamma_{i_1,i_2,i_3} = \left(1 + s\left(\sum_{j=1}^{3}\left(2 - \cos\frac{(i_j - 1)\pi}{n_j}\right)\right)^2\right)^{-1} \tag{14}$$

where i_j is the ith element in the jth dimension (j = 1, 2, or 3 here); n_j represents the size of D along the jth dimension.

The smoothing factor, s, thus becomes the only unknown parameter, and a previous investigation has concluded that the 10^{-6} is an optimal value for s considering the scale of the reconstruction error (G. Wang et al., 2012). The

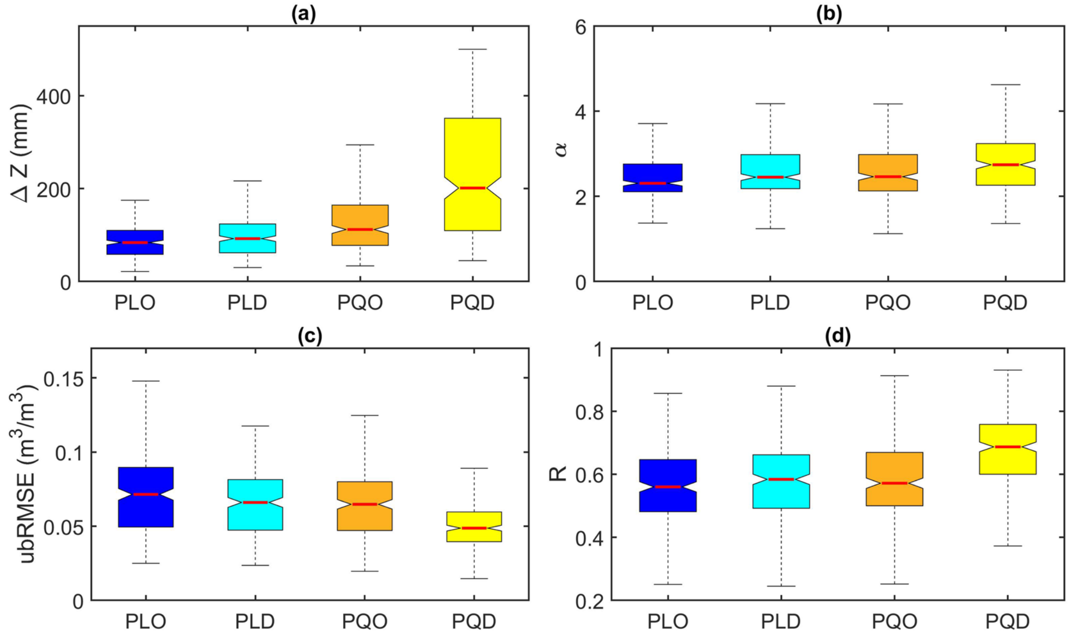

Figure 2. Boxplots of the optimal parameters (a) ΔZ (mm) and (b) α, and performance metrics of (c) unbiased root-mean-square error (m^3/m^3) and (d) R calculated by separately comparing the PLO, PLD, PQO, and PQD against the Soil Moisture Active Passive retrievals.

detailed mathematical derivations and the selection of the smoothing parameters can be referred to Garcia (2010) and Mironov et al. (2012); G. Wang et al. (2012).

4. Results and Discussion

4.1. Determination of Optimal Simulation Flow

The performance of the precipitation-driven SM data set relies on the accuracy of loss estimations and the quality of precipitation input. Comparison of various precipitation products is beyond the scope of this study. In the context of the water balance model proposed herein, this section primarily aims to address two questions: (a) which loss function, LOWESS-based or quantile regression-based, does yield SM estimations of better performance? (b) what is the magnitude of enhancement in the derived SM time series upon integrating dry-down SM into the simulation process, compared to employing the initial SM exclusively? Given these, four methodology flows (i.e., PLO, PLD, PQO, and PQD) were evaluated by comparing against the SMAP observations over the representative pixels (Figure S2 in Supporting Information S1; Section 3.2). The objective behind this incorporation is to discern the optimal simulation workflow conducive to producing high-quality SM.

The common parameters (ΔZ and α) (Figures 2a and 2b) and performance metrics (ubRMSE and R) (Figures 2c and 2d) of precipitation-driven simulations from four different methodology flows (i.e., PLO, PLD, PQO, and PQD) are illustrated below. The PQD shows a pronounced superiority over the PLO, PLD, and PQO (Figures 2c and 2d) whereas the PQO exhibits a comparable performance with the PLO and PLD. The latter result indicates that while the introduced β is pixel dependent, employing a quantile-regression-based loss function with only one initiated SM does not significantly improve the SM simulations conducted via the proposed water balance model.

Since more SMAP data have been incorporated into the PLD and PQD, their validation results tend to outperform the PLO and PQO. Despite this, the improvement of the PLD relative to the PLO is marginal, potentially due to the low occurrence of dry-down events within the SMAP records. This observation implies that the substantial enhancement witnessed in PQD, as opposed to PQO, cannot be solely attributed to the additional dry-down SM observations in the PQD time series. Instead, it is the synergistic effect of combining dry-down SM with the additional parameter β in the simulation process, resulting in the superior performance of PQD. The preliminary investigations (not shown here) reveal that the drying rates of the PLO and PQO are excessively fast, resulting in numerous extremely low SM values. This, in turn, results in the prediction of low loss amounts, subsequently causing abrupt SM spikes upon the occurrence of rainfall events. The integration of dry-down SM into the simulation process can mitigate this issue.

Based on Figure 2b, the parameter α is insensitive to different methodology flows considered here. Conversely, the values of ΔZ exhibit significant variance across the various simulation pathways. In Figure 2a, there is a pronounced stretch and a high average for ΔZ values of the PQD. It should be noted that ΔZ does not serve as an indicator of the microwave radiometer's sensing depth. Rather, ΔZ represents the depth within which SM variations exhibit temporal analogies and its value is predominately controlled by precipitation characteristics (Akbar et al., 2018). Furthermore, the variation in ΔZ is consistent with the findings of H. Kim and Crow (2024), indicating that ΔZ is likely subject to spurious biases stemming from limitations in the temporal resolution and accuracy of satellite-based SM estimates, as well as the omission of other hydrological factors. Despite the distinctive discrepancies of ΔZ values between the PLO and PQD, the comparable spatial distribution of ΔZ across the CONUS and similar median values (139 mm for PQD and 135 mm in Akbar et al. (2018)) have been observed in Figure S3 in Supporting Information S1. The elevated ΔZ values within the PQD framework could signify a tendency toward lower loss amounts. This observation can be interpreted as a strategic adjustment by the PQD methodology, aimed at mitigating the issue of unduly rapid drying rates that characterized the PLO and PQO. Incorporating the additional parameter β into the loss function has endowed ΔZ with greater adaptability. The simultaneous application of ΔZ and β facilitates a balance interchange, ensuring a steady range of SM variations throughout the optimization process. These cannot be achieved by the application of LOWESS with a fixed span of 65%. Moreover, the parameter β could reflect pixel-scale climatic and geographical characteristics. The harmonization of addressing the rapid drying rate issue, the utilization of diverse parameter set combinations, and the application of pixel-specific β values within the optimization procedure collectively contribute to the superior performance of the PQD methodology over other approaches. Although the application of the parameter β still cannot fully capture the scattered losses at higher SM values, this scheme represents an optimal option considering accuracy. Based on all the above findings, the PQD methodology was decided as the operational approach for the precipitation-reconstructed SM simulations and parameter derivations in the following analysis.

4.2. Overall Performance of Precipitation-Driven Soil Moisture

The ubRMSE quantifies the discrepancies in absolute magnitude between the PQD and the SMAP retrievals. The ubRMSE map (Figures 3a, 3c, and 3d) exhibits an east-west gradient in its distribution. The relatively higher deviations of the PQD in the eastern CONUS could be partly attributed to the generally larger SM values caused by more precipitation volumes. Additionally, the SMAP SM data in the eastern sides were mostly retrieved under the VWC of more than 5 kg/m^2, which is a commonly used threshold to screen out low-quality SM estimates from densely vegetated areas (O'Neill, Bindlish, et al., 2021). Therefore, the loss estimations derived from the dry downs of inferior SMAP SM data may not function as quantitatively well as in regions with VWC below 5 kg/m^2.

The spatial distributions of R between the PQD and the SMAP retrievals are illustrated in Figures 4a, 4c, and 4d. The median R of 0.69 indicates a good agreement between the temporal variations described by the PQD and the SMAP observations throughout the entire study period. Previous studies (Akbar et al., 2018; Koster et al., 2017) focused on performing this process-based water balance scheme over summer seasons with sufficiently available SMAP retrievals and at a scale of 36 km. Moving forward, in this analysis, a separate comparison was made between the 9 km PQD and SMAP SM data for both cold seasons (November to April) and warm seasons (May to October). In terms of R, there is a noticeable degradation in the performance of SM simulations over cold seasons relative to warm seasons (Figure 4b). In the high-latitude areas, the R values quickly drop from 0.7 to 0.5 (Figures 4c and 4d), which could be attributed to the low availability of the SMAP benchmarks during winter due to long frozen periods and/or frequent snowfall events. However, it is interesting to note that there is an improvement in R values in the western CONUS during cold seasons (Figure 4d). This enhanced performance during cold seasons can likely be attributed to the distinctive seasonal rainfall patterns in the western CONUS, where precipitation is both more voluminous and frequent in the winter months compared to summer. The observed favorable results over the western CONUS in winter appear to be in contrast with the observations made by Akbar et al. (2018), which indicated potential limitations of the loss-function-based water balance model in this region.

The reduced performance of the PQD product observed throughout the majority of CONUS during winter months can be principally attributed to the evapotranspiration discrepancy between summer and winter. Although the loss

10.1029/2023WR034457

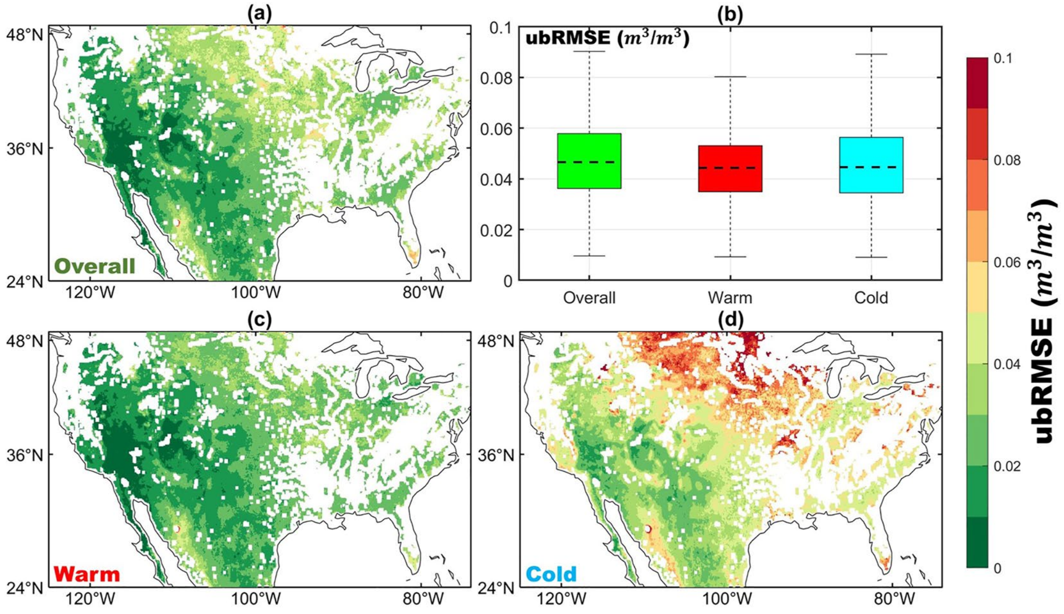

Figure 3. Spatial distribution of unbiased root-mean-square error (m^3/m^3) between the PQD simulations and the Soil Moisture Active Passive observations over the conterminous United States, where (a), (c), and (d) represent the metrics obtained using soil moisture data of the entire study period, warm seasons (May to October), and cold season (November to April).

term in the water balance model encompasses components such as drainage, runoff, and evapotranspiration, evapotranspiration plays the dominant role in most scenarios (Laio et al., 2001; McColl, Wang, et al., 2017). The disparity between summer and winter evapotranspiration is pronounced, primarily due to the substantial differences in temperature, solar radiation, vegetation physiology, and soil water availability across these seasons. Consequently, the application of a uniform loss function throughout the year fails to accurately capture the seasonal variations in SM, as evidenced by Figures 3 and 4, where the performance of PQD significantly

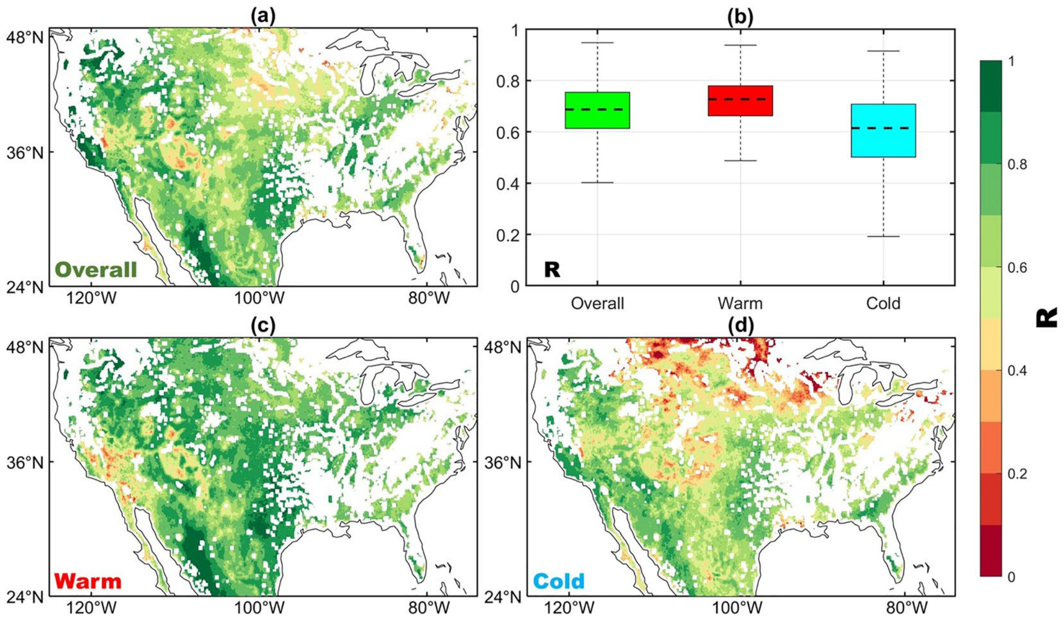

Figure 4. Spatial distribution of R between the PQD simulations and the Soil Moisture Active Passive observations over the conterminous United States, where (a), (c), and (d) represent the metrics obtained using soil moisture data of the entire study period, warm seasons (May to October), and cold season (November to April).

diminishes in winter compared to summer. It should be noted that the loss functions employed in this study are more representative of the SM characteristics in the warmer seasons, given the predominance of dry-down SM data extracted from those periods. The choice of seasonally invariant loss functions and parameters was driven by computational considerations and maintaining methodology simplicity. However, for a more precise representation of SM dry-down behavior and to enhance the quality of the precipitation-driven SM estimations, the implementation of seasonally adaptive loss functions is indispensable.

Additionally, it should also be noted that the contrast between daytime and nighttime evapotranspiration is stark. Some studies posit that nighttime evapotranspiration can be considered negligible, justified by the absence of solar radiation and the closure of plant stomata during nighttime (K. Wang & Dickinson, 2012). Given the 12-hr temporal resolution employed in this study, this discrepancy becomes a pivotal factor when predicting daytime SM from the preceding nighttime values during periods without precipitation. Nonetheless, the utilization of SMAP SM data in itself poses challenges for detecting subtle diurnal changes in SM, primarily due to the improper assumption of thermal equilibrium for ascending retrieval (O'Neill, Bindlish, et al., 2021). Hence, the estimated losses are likely to be inaccurate even if an abundance of consecutive nighttime and daytime dry-down data were available. In this context, the concurrent use of descending and ascending SM retrievals primarily serves to double data count for acquiring more dry-down segments and optimization benchmarks. Consequently, aligning with the method used by Akbar et al. (2018), we made this simplifying assumption that daytime and nighttime evapotranspiration rates are equivalent. However, a more refined water balance model could be achieved through the integration of a priori knowledge regarding the diurnal cycle of SM, along with the segregation of daytime and nighttime evapotranspiration rates, which could be categorized based on solar radiation intensity and diurnal temperature range (K. Wang & Dickinson, 2012).

In addition to the performance of the reconstructed loss function, the precipitation input plays a critical role in this SM simulation scheme. The precipitation data set not only influences the selection of dry-down SM data but also directly affects the estimation of SM evolution over time. Therefore, it is important to account for the errors and limitations present in the GPM data set. To date, the accuracy and utility of the GPM product have undergone extensive validation from different perspectives and across various regions, and several studies have highlighted specific issues with the GPM product (Beck et al., 2019; W. Cui et al., 2020; Mazzoglio et al., 2019; Pradhan et al., 2022; Tran et al., 2023). For example, W. Cui et al. (2020) found that the GPM data set tends to overestimate the rainfall hours and precipitating areas in the central and eastern United States. Mazzoglio et al. (2019) indicated that the GPM product may struggle to accurately capture extreme rainfall events at aggregation intervals shorter than 12 hr. These errors in the GPM product can propagate into the SM simulations. Moreover, the consistency between the precipitation and SM data sets is also crucial. Divergence between the trends of precipitation and SM can lead to a reduced availability of dry-down SM data used for constructing the loss function and directly degrade the filled-in SM estimates through simulation. Moreover, the skills of the PQD and PQF are also limited by the quality of the SMAP SM retrievals as these retrievals were also employed in building the function for Segment B as well as resetting the SM values during the integrations (Equation 6).

The integrated assessment of the ubRMSE (0.05 m^3/m^3) and R (0.69) demonstrates the comparable performance of the PQD relative to the SMAP retrievals, affirming the validity of the derived parameters (i.e., ΔZ, α, and β). These parameters are necessary for quantifying the integrated losses while the conventional gap-filling methods are usually unable to afford such information. The availability of loss estimates holds significant potential for other research endeavors, such as the SM to Rain algorithm (SM2RAIN) (Brocca et al., 2014, 2019).

4.3. Validation of Gap-Filled Soil Moisture Products

The performances of the PLO, PQD, and PQF were further evaluated by comparing them against in situ measured SM over 526 CONUS pixels (Figure S4 in Supporting Information S1). Here, the PLO acts a representative of the product generated through the framework established by Akbar et al. (2018), with its resultant SM data set still pending validation through in situ measurements. The comparative analysis of the PLO, PQD, and PQF elucidates the advantages conferred by the integration of supplementary procedures like quantile regression and SMAP SM observations into our simulation processes. Given the share of same parameters in SM simulation, comparing PQD with PQF enables a thorough examination of the effects of integrating the full suite of SMAP observations into the precipitation-driven simulation, as opposed to exclusively considering dry-down SM. Additionally, assessment metrics of another two SMAP-based continuous SM data sets yielded using the

Table 2
Statistic Metrics of Different Gap-Filled Products by Comparing Against In Situ Measurements

Validation type[a]/product	Metric	PLO	PQD	PQF	DCT	LIP
Overall validation	ubRMSE (m^3/m^3)	0.08	0.07	0.06	0.06	0.06
	R	0.46	0.59	0.63	0.65	0.65
SMAP-non-synchronous validation	ubRMSE (m^3/m^3)	0.09	0.07	0.06	0.06	0.06
	R	0.46	0.59	0.63	0.67	0.66
SMAP-synchronous validation	ubRMSE (m^3/m^3)	0.08	0.07	0.06	0.06	0.06
	R	0.45	0.59	0.64	0.64	0.64

[a]Overall, SMAP-non-synchronous, and SMAP-synchronous validation metrics are computed using all the paired data, all the paired simulations non-synchronous with the SMAP observations, and all the paired simulation synchronous with the SMAP observations across the five gap-filled data sets, respectively.

DCT (Section 3.5), and the simple LIP, were also calculated. The LIP SM product was generated by conducting a linear temporal interpolation for the gaps among the available SMAP retrievals. Regarding the DCT SM data set, only the SMAP retrievals from the neighboring grids surrounding the targeted pixel (as a center of a 9×9 matrix) have been considered, rather than including the entire three-dimensional SMAP array within the study domain.

The PQF, DCT, and LIP display comparable ubRMSE and *R* medians (Table 2). The minor degradation of PQD relative to PQF could be resulted from the smaller number of SMAP retrievals used during its simulation. Given this, a sufficient number of satellite-based SM observations are necessary. Consistent with the results of Section 4.1, the distinct inferiority of the PLO could be largely attributed to the different regression techniques used for the loss estimations.

Compared to the metric scores of the SMAP observations, the gap-filled portions show comparable performances (Table 2). Although the ubRMSE values of SMAP observations slightly outperform the filled-in SM (when more digits are examined), the *R* values of filled-in SM of the DCT and LIP products are unexpectedly higher than those of the SMAP data (Table 2). Such small discrepancies between the performances of the SMAP retrievals and those exclusively filled data are adequate to demonstrate the validity of the proposed scheme and the derived ancillary parameters. It should be noted that the average number of filled-in SM data was around two times relative to that of the SMAP available samples.

Moreover, the consistency of rainfall-induced SM peaks between each gap-filled data set and in situ measurements was analyzed. It should be noted that rainfall events identified by the NLDAS precipitation product have been considered in evaluating the peak-capturing capacity, independent from the GPM precipitation used in deriving the PLO, PQD, and PQF. Those peaks caused by heavy rainfall were majorly investigated here given their importance. As shown in Section 3.3, the SM peak consistency was measured by three categorical metrics: POD, FAR, and CSI. The POD refers to the ratios of rainfall-induced SM peaks successfully detected by SMAP-derived products to all the rainfall-induced SM peaks caught by in situ measurements in the context of the NLDAS precipitation, and the optimum POD is 1. Figure 5a shows that the POD medians of the PQF, PQD, and PLO are near 1, suggesting that these products could capture almost all heavy rainfall-induced SM peaks. The FAR (optimal FAR is 0) represents the fractions of peaks identified by the SMAP-derived data sets but not observed by in situ measurements, and it is described in Figure 5b. In contrast, it reveals that the PLO, PQD, and PQF hold excessive peaks. Such frequent peak occurrences can be attributed to the forward simulation procedure. Generally, a precipitation event is bound to produce one immediate SM peak at the end of the precipitation event (Equation 6). Furthermore, the CSI (optimal CSI is 1) reflects an integrated ability of the simulation products to capture rainfall-induced SM peaks. The PLO, PQD, and PQF have better CSI values than the DCT and LIP and the PQF obtains the highest score (Figure 5c). When considering the categoric metrics of capturing SM peaks induced by all the rainfall, the advantages of the PQF over the DCT and LIP can be still observed (Figure S5 in Supporting Information S1) in a similar manner shown in Figures 5a and 5c.

A detailed investigation was conducted to understand the contrasting performance rankings of the PQF, DCT, and LIP in terms of validation metrics and their ability to capture rainfall-induced SM peaks. The SM time series of these products were compared against in situ measurements from three ground stations, focusing on a 4-month period with abundant observations over each station. The in situ SM measurements show timely responses to

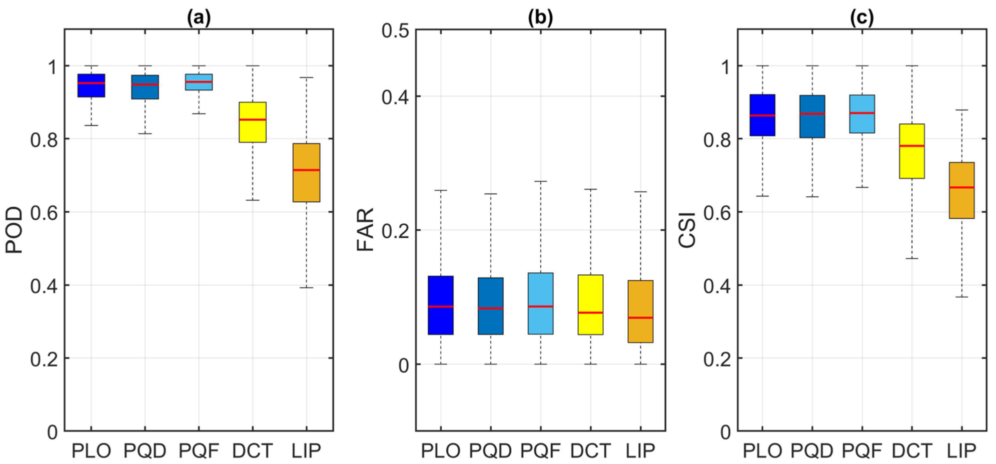

Figure 5. Boxplots of the categorical performance metrics of (a) probability of detection, (b) false alarm ratio and (c) critical success index for five different Soil Moisture Active Passive-based gap-filled soil moisture (SM) data sets in capturing SM peaks caused by the heavy rainfall events (exceeding 80% locally non-zero 12-hr precipitation volumes). Those SM peaks observed by in situ SM measurements are used as benchmarks.

notable precipitation events recorded by the GPM product, resulting in SM peaks (Figures 6a–6c). The PQF product closely follows the temporal fluctuations of in situ SM measurements, capturing rainfall-induced SM peaks that were not observed by the DCT and LIP. However, there are some differences in SM magnitudes between the PQF and in situ data. Figure 6c shows a systematic bias between the SMAP and in situ SM time series, indicating discrepancies between the two data sets. Several factors could contribute to these differences. First, the use of a uniform pixel-scale parameter set across all periods may not accurately capture the local variability of SM.

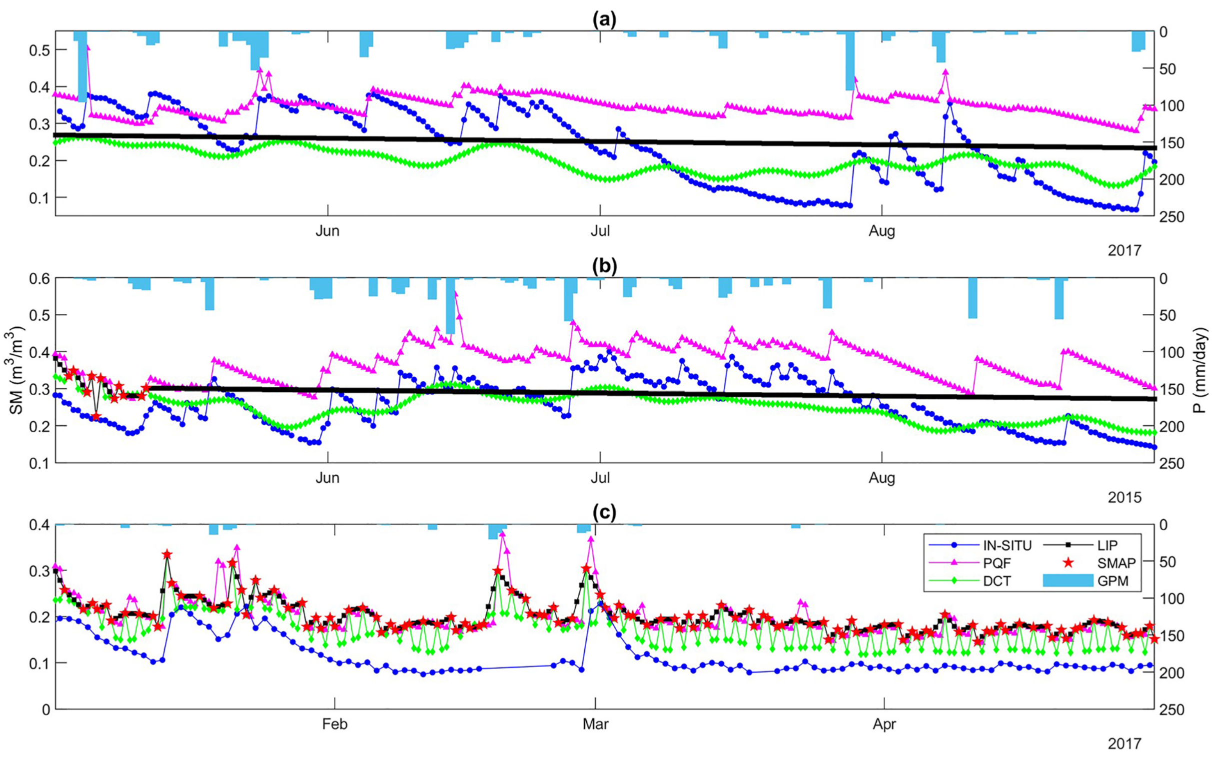

Figure 6. Four-month soil moisture time series from the PQF, discrete cosine transform, LIP, Soil Moisture Active Passive, and in situ measurements over three International Soil Moisture Network sites: (a) SCAN-Shenandoah (37.9°N, 79.2°W) (b) USCRN-Ithaca_13_E (42.4°N, 76.2°W) and (c) PBO_H2O-SPRECKLESS (32.9°N, 115.6°W).

Second, there are inherent discrepancies between SMAP retrievals and in situ measurements. Additionally, the point-scale in situ SM data may not fully represent the variability at the satellite-based 9 km grids. Moreover, there can be differences in the measuring depths between the passive microwave SM derivations and ground measurements, which can introduce biases in the validation scores (Crow et al., 2012).

In Figures 6a and 6b, the behaviors of SM estimates from the DCT and LIP are abnormal due to the sparse availability of SMAP retrievals. The LIP displays a linear decline at an extremely slow pace over the 4-month interval (Figure 6a) while the DCT shows a smooth sinusoidal variation, only capturing the general increasing trends of SM caused by heavy rainfall events (Figures 6a and 6b). Although Section 4.2 indicates that the low availability of SMAP retrievals during cold seasons could reduce the performance of PQD at those periods, Figure 6 exhibits that the DCT and LIP are more vulnerable to the sparse SMAP samples. The DCT SM product could exhibit unnatural fluctuations not observed in other products even when there are sufficient SMAP retrievals (Figure 6c). However, the impact of these abnormal fluctuations on the DCT's performance score was limited, partly due to the daily availability of in situ measurements.

Despite the slightly better validation results of the DCT and LIP compared to the PQF (Table 2), the detailed investigation on SM time series indicated that their performance might not be as good as the metric scores suggested. For example, the correlation between the flat line of LIP and in situ data from May to August 2017 is 0.78 (Table S3 in Supporting Information S1). This high correlation, however, does not necessarily indicate that the LIP accurately captures the SM dynamics. The LIP, with its linearly interpolated values, does not reflect the actual fluctuations in SM over this period. Despite this, it still maintains a high correlation with the in situ data because both data sets share a similar overall trend. This example illustrates that a high R does not always equate to an accurate representation of SM dynamics, particularly when one data set exhibits larger or smaller fluctuations than the other (Wilks, 2011). Additionally, it is important to note that the R values among the three products show a significant difference when only 4-month periods were considered (Table S3 in Supporting Information S1). This finding aligns with the work of Al-Yaari et al. (2019), which demonstrated the significant influence of data temporal sampling on performance metrics. Recognizing these constraints, the methodology presented in this study focuses on evaluating the capability to capture SM peaks induced by rainfall, offering a valuable additional tool for validating SM data sets and facilitating various related applications. For instance, this approach of evaluating peak-capturing performance could be utilized to preliminarily assess the linearly interpolated ASCAT SM time series employed in the SM2RAIN algorithm (Brocca et al., 2019). It is expected that areas frequently missing SM peaks may not yield high-quality precipitation estimates through SM2RAIN. While assessing peak-capturing capacity may not be highly suited for SM products derived from a single mission due to their coarse temporal resolution, it gains significant relevance with the increasing availability of merged SM products from multiple sensors, such as ESA CCI and SMOSSMAP-IB, and the advent of sophisticated gap-filling techniques (Dorigo et al., 2017; Li et al., 2022; K. Liu et al., 2023; Zheng et al., 2023). In these contexts, the temporal alignment of SM peaks with intense precipitation events can serve as robust evidence for the reliable performance of the SM data set under evaluation.

Additionally, the unique peak capturing capabilities of the PQD and PQF products offer a distinct advantage, facilitating the understanding of the hydrological processes. The residence time of SM is a vital indicator of the water dynamics at the land-atmosphere interface. To address this, the stored precipitation fraction (F_p), was introduced to reflect the averaged fraction of precipitation remaining in the surface soil layer after 3 days post-precipitation (McColl, Alemohammad, et al., 2017). However, F_p is an abstract and normalized indicator, likely disturbed by temporal mismatching between the precipitation and SM data sets. Since both PQD and PQF products are adequate to represent SMAP varying patterns, their SM peaks as well as the SM data captured 3 days post-peaks, can be used to derive a metric (analogous to F_p) that specifically gauges the quantity of precipitation inputs retained in the soil after a 3-day period. Furthermore, instead of the commonly used 50 mm depth, the active hydrological system's depth, ΔZ, appears to be a more suitable parameter to measure the SM memory of wet anomalies. Moreover, the peak capturing feature of the PQD and PQF extends its utility to estimating the response magnitude (i.e., the difference between the minimum and maximum SM during storm) and the peak-to-peak time (i.e., the time difference between precipitation peak and SM peak) (Singh et al., 2021). Correlating these metrics with topographic variables and vegetation indexes can unveil the primary factors driving SM response to rainfall. Furthermore, estimating lag times between SM and runoff peaks, when data are available, could enrich our grasp of SM-runoff process (Singh et al., 2021). Given the considerable influence of antecedent SM conditions on assessing flood severity, integrating SM-runoff peak lag times with the process-based water balance framework presented in this study—initialized with preceding observed SM and under diverse hypothetical precipitation scenarios—could offer a supplementary tool for flood prediction in regions susceptible to flooding (Haga et al., 2005; Sharma et al., 2018; Wasko & Nathan, 2019). In the context of climate change with increasing precipitation extremes, such studies are of great importance to identify flooding-prone regions and to advance water resources management.

5. Summary and Conclusion

In this study, a 12-hourly continuous SM product (PQF) was generated for the CONUS region from 2015 to 2021. Notably, regions identified as unsuitable for SM retrievals through satellite-based microwave observations have been excluded, according to the SMAP quality flag. This newly yielded data set is composed of SMAP retrievals and filled-in SM estimates derived through water balance budgeting. Gaps in the combined descending and ascending SMAP data set were filled using SM simulations estimated using the precipitation and hydrologic loss in the preceding 12-hr slots. The integrated loss for each 12-hr interval was approximated using a loss function derived from SMAP dry-down SM. Here, a novel approach that combines quantile regression with a performance-driven parameter has been proposed to account for the wide range of losses at the high SM magnitudes. Compared to the PLD yielded using the LOWESS with a fixed smoother, the PQD product based on this new method noticeably improved the performance scores by comparing against the SMAP retrievals.

The PQD product, which only includes SMAP dry-down retrievals as reset points, was used to determine the optimal parameters for the generation of the final PQF data set. The PQD product exhibited similar features to SMAP SM retrievals, with median R of 0.69 and a median ubRMSE of 0.05 m^3/m^3. A seasonal analysis of the PQD product revealed its deteriorated performance during the cold seasons.

To evaluate the 9 km sub-daily PQD and PQF SM data sets, comparisons were made with in situ surface SM measurements over 526 CONUS pixels and with the SMAP-based 12-hr SM products through other filling approaches (i.e., DCT and LIP). The PQF, DCT, and LIP exhibited similar accuracy, with an ubRMSE of 0.06 m^3/m^3 and an R of 0.63. Notably, the exclusive filled-in SM estimates of PQF displayed a comparable performance with the SMAP retrievals, demonstrating the effectiveness of this process-based water balance scheme. In addition to the conventional validation metrics, the ability of the PQD, PQF, DCT, and LIP to capture rainfall-induced SM peaks was assessed where the rainfall-induced SM peaks observed by in situ measurements served as the reference. The precipitation-driven PQF data sets showed superior performance compared to other gap-filling approaches in capturing SM responses to heavy rainfall events.

Based on the available results, the PQF product is considered the continuous SMAP-based SM data set of the best quality. However, it's important to acknowledge the limitations of the proposed process-based water balance scheme. The abrupt varying patterns in the PQD and PQF compared to observed time series can be partly ascribed to the lack of SM diurnal cycle information, such as the nighttime re-moistening (O'Neill, Bindlish, et al., 2021). Inaccurate loss quantification is another crucial factor, as existing loss functions may not be able to reflect the scattering of observed losses for the same SM initial. Additionally, SM simulations tend to have a higher SM upper bound than observations primarily because of inadequate representation of drainage and runoff during heavy precipitation. Furthermore, SM simulations filling longer SMAP gaps tend to introduce greater uncertainty. As such, a flag file has been generated to mark filled-in SM values for long-term SMAP gaps and those with impractical values.

The incorporation of novel features and identification of deficiencies in this study offers valuable insights and directions for future research. As a single loss function seems inadequate to reflect various losses under the initial SM of similar magnitudes, it is essential to supplement uncertainty information for the derived losses and filled-in SM data. This could be achieved through the use of an error propagation method (e.g., Monte Carlo simulation), which would enhance the interpretation of variability and uncertainty associated with the derived loss and simulated SM. Additionally, given the significant influence of precipitation inputs on driving SM estimates in the proposed water balance model, it is crucial to conduct sensitivity tests using different precipitation data sets. These tests can provide valuable supplementary information for evaluating the performance and reliability of the precipitation data used in the model.

Data Availability Statement

The datasets utilized in this study come from various data sources. The SMAP soil moisture data is provided by O'Neill, Chan, et al. (2021). The ERA5 Land Data comes from Muñoz-Sabater (2019), while the GPM IMERG dataset is sourced from Huffman et al. (2019). For validation, in situ soil moisture data were sourced both from the International Soil Moisture Network at ISMN (2022) and from the TxSON Network as provided by Bongiovanni and Caldwell (2019). Additionally, the NLDAS hourly precipitation data is available from NLDAS project (2021).

Acknowledgments
This work was supported by the Korea Environment Industry & Technology Institute (KEITI) through Water Management Program for Drought, funded by Korea Ministry of Environment (MOE) (RS-2023-0023194). We also want to express our sincere appreciation to Dr. Randal Koster and two anonymous reviewers for their insightful comments and suggestions, which have significantly contributed to the improvement of this manuscript.

References

Akbar, R., Short Gianotti, D., McColl, K. A., Haghighi, E., Salvucci, G. D., & Entekhabi, D. (2018). Hydrological storage length scales represented by remote sensing estimates of soil moisture and precipitation. *Water Resources Research*, *54*(3), 1476–1492. https://doi.org/10.1002/2017wr021508

Almendra-Martín, L., Martínez-Fernández, J., Piles, M., & González-Zamora, Á. (2021). Comparison of gap-filling techniques applied to the CCI soil moisture database in Southern Europe. *Remote Sensing of Environment*, *258*, 112377. https://doi.org/10.1016/j.rse.2021.112377

Al-Yaari, A., Wigneron, J.-P., Dorigo, W., Colliander, A., Pellarin, T., Hahn, S., et al. (2019). Assessment and inter-comparison of recently developed/reprocessed microwave satellite soil moisture products using ISMN ground-based measurements. *Remote Sensing of Environment*, *224*, 289–303. https://doi.org/10.1016/j.rse.2019.02.008

Balenzano, A., Mattia, F., Satalino, G., Lovergine, F. P., Palmisano, D., Peng, J., et al. (2021). Sentinel-1 soil moisture at 1 km resolution: A validation study. *Remote Sensing of Environment*, *263*, 112554. https://doi.org/10.1016/j.rse.2021.112554

Beck, H. E., Pan, M., Roy, T., Weedon, G. P., Pappenberger, F., Van Dijk, A. I., et al. (2019). Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS. *Hydrology and Earth System Sciences*, *23*(1), 207–224. https://doi.org/10.5194/hess-23-207-2019

Bell, J. E., Palecki, M. A., Baker, C. B., Collins, W. G., Lawrimore, J. H., Leeper, R. D., et al. (2013). US Climate Reference Network soil moisture and temperature observations. *Journal of Hydrometeorology*, *14*(3), 977–988. https://doi.org/10.1175/jhm-d-12-0146.1

Bongiovanni, T., & Caldwell, T. G. (2019). Texas soil observation network (TxSON) [Dataset]. Texas Data Repository, V3. Retrieved from https://dataverse.tdl.org/dataset.xhtml?persistentId=doi:10.18738/T8/JJ16CF

Brocca, L., Ciabatta, L., Massari, C., Moramarco, T., Hahn, S., Hasenauer, S., et al. (2014). Soil as a natural rain gauge: Estimating global rainfall from satellite soil moisture data. *Journal of Geophysical Research: Atmospheres*, *119*(9), 5128–5141. https://doi.org/10.1002/2014jd021489

Brocca, L., Filippucci, P., Hahn, S., Ciabatta, L., Massari, C., Camici, S., et al. (2019). SM2RAIN–ASCAT (2007–2018): Global daily satellite rainfall data from ASCAT soil moisture observations. *Earth System Science Data*, *11*(4), 1583–1601. https://doi.org/10.5194/essd-11-1583-2019

Caldwell, T. G., Bongiovanni, T., Cosh, M. H., Jackson, T. J., Colliander, A., Abolt, C. J., et al. (2019). The Texas soil observation network: A comprehensive soil moisture dataset for remote sensing and land surface model validation. *Vadose Zone Journal*, *18*(1), 1–20. https://doi.org/10.2136/vzj2019.04.0034

Chan, S. K., Bindlish, R., O'Neill, P., Jackson, T., Njoku, E., Dunbar, S., et al. (2018). Development and assessment of the SMAP enhanced passive soil moisture product. *Remote Sensing of Environment*, *204*, 931–941. https://doi.org/10.1016/j.rse.2017.08.025

Colliander, A., Jackson, T. J., Bindlish, R., Chan, S., Das, N., Kim, S., et al. (2017). Validation of SMAP surface soil moisture products with core validation sites. *Remote Sensing of Environment*, *191*, 215–231. https://doi.org/10.1016/j.rse.2017.01.021

Cook, D. R. (2016). Soil temperature and moisture profile (STAMP) system handbook.

Crow, W. T., Berg, A. A., Cosh, M. H., Loew, A., Mohanty, B. P., Panciera, R., et al. (2012). Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products. *Reviews of Geophysics*, *50*(2), RG2002. https://doi.org/10.1029/2011rg000372

Cui, W., Dong, X., Xi, B., Feng, Z., & Fan, J. (2020). Can the GPM IMERG final product accurately represent MCSs' precipitation characteristics over the central and eastern United States? *Journal of Hydrometeorology*, *21*(1), 39–57. https://doi.org/10.1175/jhm-d-19-0123.1

Cui, Y., Zeng, C., Zhou, J., Xie, H., Wan, W., Hu, L., et al. (2019). A spatio-temporal continuous soil moisture dataset over the Tibet Plateau from 2002 to 2015. *Scientific Data*, *6*(1), 247. https://doi.org/10.1038/s41597-019-0228-x

Das, N. N., & O'Neill, P. (2020). *Soil moisture active passive (SMAP) ancillary data report, soil attributes, August 15, 2020, JPL D-53058, version B*. Jet Propulsion Laboratory.

Dorigo, W., Gruber, A., De Jeu, R., Wagner, W., Stacke, T., Loew, A., et al. (2015). Evaluation of the ESA CCI soil moisture product using ground-based observations. *Remote Sensing of Environment*, *162*, 380–395. https://doi.org/10.1016/j.rse.2014.07.023

Dorigo, W., Himmelbauer, I., Aberer, D., Schremmer, L., Petrakovic, I., Zappa, L., et al. (2021). The international soil moisture network: Serving earth system science for over a decade. *Hydrology and Earth System Sciences*, *25*(11), 5749–5804. https://doi.org/10.5194/hess-25-5749-2021

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., et al. (2017). ESA CCI soil moisture for improved earth system understanding: State-of-the art and future directions. *Remote Sensing of Environment*, *203*, 185–215. https://doi.org/10.1016/j.rse.2017.07.001

Dorigo, W., Xaver, A., Vreugdenhil, M., Gruber, A., Hegyiova, A., Sanchis-Dufau, A., et al. (2013). Global automated quality control of in situ soil moisture data from the International Soil Moisture Network. *Vadose Zone Journal*, *12*(3), 1–21. https://doi.org/10.2136/vzj2012.0097

ElSaadani, M., Habib, E., Abdelhameed, A. M., & Bayoumi, M. (2021). Assessment of a spatiotemporal deep learning approach for soil moisture prediction and filling the gaps in between soil moisture observations. *Frontiers in artificial intelligence*, *4*, 636234. https://doi.org/10.3389/frai.2021.636234

Entekhabi, D., Yueh, S., O'Neill, P., Kellogg, K., Allen, A., Bindlish, R., et al. (2014). SMAP handbook soil moisture active passive. In *Mapping soil moisture freeze/thaw from space, Pasadena, CA*.

Fang, B., Kansara, P., Dandridge, C., & Lakshmi, V. (2021). Drought monitoring using high spatial resolution soil moisture data over Australia in 2015–2019. *Journal of Hydrology*, *594*, 125960. https://doi.org/10.1016/j.jhydrol.2021.125960

Findell, K. L., Gentine, P., Lintner, B. R., & Kerr, C. (2011). Probability of afternoon precipitation in eastern United States and Mexico enhanced by high evaporation. *Nature Geoscience*, *4*(7), 434–439. https://doi.org/10.1038/ngeo1174

Garcia, D. (2010). Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational Statistics & Data Analysis*, *54*(4), 1167–1178. https://doi.org/10.1016/j.csda.2009.09.020

Ge, Y., Thomasson, J. A., & Sui, R. (2011). Remote sensing of soil properties in precision agriculture: A review. *Frontiers of Earth Science*, *5*(3), 229–238. https://doi.org/10.1007/s11707-011-0175-0

Haga, H., Matsumoto, Y., Matsutani, J., Fujita, M., Nishida, K., & Sakamoto, Y. (2005). Flow paths, rainfall properties, and antecedent soil moisture controlling lags to peak discharge in a granitic unchanneled catchment. *Water Resources Research*, *41*(12), W12410. https://doi.org/10.1029/2005wr004236

Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., et al. (2014). The global precipitation measurement mission. *Bulletin of the American Meteorological Society*, *95*(5), 701–722. https://doi.org/10.1175/bams-d-13-00164.1

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., & Tan, J. (2019). GPM IMERG final precipitation L3 half hourly 0.1 degree x 0.1 degree V06 [Dataset]. Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/GPM/IMERG/3B-HH/06

ISMN. (2022). International soil moisture network [Dataset]. ISMN. Retrieved from https://ismn.earth/en/dataviewer/

Jackson, R. D. (1973). Diurnal changes in soil water content during drying. *Field soil water regime*, *5*, 37–55.

Jackson, T. J., O'Neill, P. E., & Swift, C. T. (1997). Passive microwave observation of diurnal surface soil moisture. *IEEE Transactions on Geoscience and Remote Sensing*, *35*(5), 1210–1222. https://doi.org/10.1109/36.628788

Kim, D., Lee, J., Kim, H., & Choi, M. (2016). Spatial composition of AMSR2 soil moisture products by conditional merging technique with ground soil moisture data. *Stochastic Environmental Research and Risk Assessment*, *30*(8), 2109–2126. https://doi.org/10.1007/s00477-016-1300-0

Kim, H., & Crow, W. T. (2024). Interpreting effective hydrologic depth estimates derived from soil moisture remote sensing: A Bayesian non-linear modeling approach. *Science of The Total Environment*, *908*, 168067. https://doi.org/10.1016/j.scitotenv.2023.168067

Kim, H., & Lakshmi, V. (2018). Use of Cyclone global navigation satellite system (CYGNSS) observations for estimation of soil moisture. *Geophysical Research Letters*, *45*(16), 8272–8282. https://doi.org/10.1029/2018gl078923

Kim, S., Sharma, A., Liu, Y. Y., & Young, S. I. (2021). Rethinking satellite data merging: From averaging to SNR optimization. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–15. https://doi.org/10.1109/tgrs.2021.3107028

Koster, R. D., Crow, W. T., Reichle, R. H., & Mahanama, S. P. (2018). Estimating basin-scale water budgets with SMAP soil moisture data. *Water Resources Research*, *54*(7), 4228–4244. https://doi.org/10.1029/2018wr022669

Koster, R. D., Dirmeyer, P. A., Guo, Z., Bonan, G., Chan, E., Cox, P., et al. (2004). Regions of strong coupling between soil moisture and precipitation. *Science*, *305*(5687), 1138–1140. https://doi.org/10.1126/science.1100217

Koster, R. D., Mahanama, S., Yamada, T., Balsamo, G., Berg, A., Boisserie, M., et al. (2010). Contribution of land surface initialization to subseasonal forecast skill: First results from a multi-model experiment. *Geophysical Research Letters*, *37*(2), L02402. https://doi.org/10.1029/2009gl041677

Koster, R. D., Reichle, R. H., & Mahanama, S. P. (2017). A data-driven approach for daily real-time estimates and forecasts of near-surface soil moisture. *Journal of Hydrometeorology*, *18*(3), 837–843. https://doi.org/10.1175/jhm-d-16-0285.1

Laio, F., Porporato, A., Ridolfi, L., & Rodriguez-Iturbe, I. (2001). Plants in water-controlled ecosystems: Active role in hydrologic processes and response to water stress: II. Probabilistic soil moisture dynamics. *Advances in Water Resources*, *24*(7), 707–723. https://doi.org/10.1016/s0309-1708(01)00005-7

Larson, K. M., Small, E. E., Gutmann, E. D., Bilich, A. L., Braun, J. J., & Zavorotny, V. U. (2008). Use of GPS receivers as a soil moisture network for water cycle studies. *Geophysical Research Letters*, *35*(24), L24405. https://doi.org/10.1029/2008gl036013

Leavesley, G., David, O., Garen, D., Lea, J., Marron, J., Pagano, T., et al. (2008). A modeling framework for improved agricultural water supply forecasting. In *Paper presented at the AGU fall meeting abstracts*.

Li, X., Wigneron, J.-P., Frappart, F., De Lannoy, G., Fan, L., Zhao, T., et al. (2022). The first global soil moisture and vegetation optical depth product retrieved from fused SMOS and SMAP L-band observations. *Remote Sensing of Environment*, *282*, 113272. https://doi.org/10.1016/j.rse.2022.113272

Liu, K., Li, X., Wang, S., & Zhang, H. (2023). A robust gap-filling approach for European Space Agency Climate Change Initiative (ESA CCI) soil moisture integrating satellite observations, model-driven knowledge, and spatiotemporal machine learning. *Hydrology and Earth System Sciences*, *27*(2), 577–598. https://doi.org/10.5194/hess-27-577-2023

Liu, Y., Chen, R., Yuan, S., Ren, L., Zhang, X., Liu, C., & Ma, Q. (2022). Satellite soil moisture data reconstruction in the temporal and spatial domains: Latent error assessments and performances for tracing rainstorms and droughts. *Remote Sensing*, *14*(19), 4841. https://doi.org/10.3390/rs14194841

Llamas, R. M., Guevara, M., Rorabaugh, D., Taufer, M., & Vargas, R. (2020). Spatial gap-filling of ESA CCI satellite-derived soil moisture based on geostatistical techniques and multiple regression. *Remote Sensing*, *12*(4), 665. https://doi.org/10.3390/rs12040665

Ma, H., Zeng, J., Chen, N., Zhang, X., Cosh, M. H., & Wang, W. (2019). Satellite surface soil moisture from SMAP, SMOS, AMSR2 and ESA CCI: A comprehensive assessment using global ground-based observations. *Remote Sensing of Environment*, *231*, 111215. https://doi.org/10.1016/j.rse.2019.111215

Magan, B., Kim, S., Wasko, C., Barbero, R., Moron, V., Nathan, R., & Sharma, A. (2020). Impact of atmospheric circulation on the rainfall-temperature relationship in Australia. *Environmental Research Letters*, *15*(9), 094098. https://doi.org/10.1088/1748-9326/abab35

MathWorks. (2020). Findpeaks, signal processing toolbox documentation.

Mazzoglio, P., Laio, F., Balbo, S., Boccardo, P., & Disabato, F. (2019). Improving an extreme rainfall detection system with GPM IMERG data. *Remote Sensing*, *11*(6), 677. https://doi.org/10.3390/rs11060677

McColl, K. A., Alemohammad, S. H., Akbar, R., Konings, A. G., Yueh, S., & Entekhabi, D. (2017). The global distribution and dynamics of surface soil moisture. *Nature Geoscience*, *10*(2), 100–104. https://doi.org/10.1038/ngeo2868

McColl, K. A., Wang, W., Peng, B., Akbar, R., Short Gianotti, D. J., Lu, H., et al. (2017). Global characterization of surface soil moisture drydowns. *Geophysical Research Letters*, *44*(8), 3682–3690. https://doi.org/10.1002/2017gl072819

Mecklenburg, S., Drusch, M., Kaleschke, L., Rodriguez-Fernandez, N., Reul, N., Kerr, Y., et al. (2016). ESA's Soil Moisture and Ocean Salinity mission: From science to operational applications. *Remote Sensing of Environment*, *180*, 3–18. https://doi.org/10.1016/j.rse.2015.12.025

Miralles, D. G., Teuling, A. J., Van Heerwaarden, C. C., & Vilà-Guerau de Arellano, J. (2014). Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation. *Nature Geoscience*, *7*(5), 345–349. https://doi.org/10.1038/ngeo2141

Mironov, V., Kerr, Y., Wigneron, J.-P., Kosolapova, L., & Demontoux, F. (2012). Temperature-and texture-dependent dielectric model for moist soils at 1.4 GHz. *IEEE Geoscience and Remote Sensing Letters*, *10*(3), 419–423. https://doi.org/10.1109/lgrs.2012.2207878

Moghaddam, M., Entekhabi, D., Goykhman, Y., Li, K., Liu, M., Mahajan, A., et al. (2010). A wireless soil moisture smart sensor web using physics-based optimal control: Concept and initial demonstrations. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *3*(4), 522–535. https://doi.org/10.1109/jstars.2010.2052918

Muñoz-Sabater, J. (2019). ERA5-Land hourly data from 1950 to present [Dataset]. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.e2161bac

NLDAS project. (2021). NLDAS primary forcing data L4 hourly 0.125 × 0.125 degree V2.0, edited by David M. Mocko, NASA/GSFC/HSL [Dataset]. Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/THUF4J1RLSYG

Ojo, E. R., Bullock, P. R., L'Heureux, J., Powers, J., McNairn, H., & Pacheco, A. (2015). Calibration and evaluation of a frequency domain reflectometry sensor for real-time soil moisture monitoring. *Vadose Zone Journal*, *14*(3), 1–12. https://doi.org/10.2136/vzj2014.08.0114

O'Neill, P., Bindlish, R., Chan, S., Chaubell, J., Colliander, A., Njoku, E., & Jackson, T. (2021). Algorithm theoretical basis document level 2 & 3 soil moisture (passive) data products, revision G, SMAP project, Pasadena, CA.

O'Neill, P., Chan, S., Njoku, E., Jackson, T., Bindlish, R., Chaubell, J., & Colliander, A. (2021). SMAP enhanced L3 radiometer global and polar grid daily 9 km EASE-grid soil moisture, version 5. Soil moisture [Dataset]. NASA Snow and ICE Data Center Distributed Active Archive Center. https://doi.org/10.5067/4DQ54OUIJ9DL

Osenga, E., Arnott, J. C., Endsley, K. A., & Katzenberger, J. (2019). Bioclimatic and soil moisture monitoring across elevation in a mountain watershed: Opportunities for research and resource management. *Water Resources Research*, *55*(3), 2493–2503. https://doi.org/10.1029/2018wr023653

Peng, J., Albergel, C., Balenzano, A., Brocca, L., Cartus, O., Cosh, M. H., et al. (2021). A roadmap for high-resolution satellite soil moisture applications–confronting product characteristics with user requirements. *Remote Sensing of Environment*, *252*, 112162. https://doi.org/10.1016/j.rse.2020.112162

Petropoulos, G. (2013). *Remote sensing of energy fluxes and soil moisture content*. CRC Press.

Pham, H. T., Kim, S., Marshall, L., & Johnson, F. (2019). Using 3D robust smoothing to fill land surface temperature gaps at the continental scale. *International Journal of Applied Earth Observation and Geoinformation*, *82*, 101879. https://doi.org/10.1016/j.jag.2019.05.012

Pradhan, R. K., Markonis, Y., Godoy, M. R. V., Villalba-Pradas, A., Andreadis, K. M., Nikolopoulos, E. I., et al. (2022). Review of GPM IMERG performance: A global perspective. *Remote Sensing of Environment*, *268*, 112754. https://doi.org/10.1016/j.rse.2021.112754

Ruf, C. S., Chew, C., Lang, T., Morris, M. G., Nave, K., Ridley, A., & Balasubramaniam, R. (2018). A new paradigm in earth environmental monitoring with the CYGNSS small satellite constellation. *Scientific Reports*, *8*(1), 1–13. https://doi.org/10.1038/s41598-018-27127-4

Salvucci, G. D. (2001). Estimating the moisture dependence of root zone water loss using conditionally averaged precipitation. *Water Resources Research*, *37*(5), 1357–1365. https://doi.org/10.1029/2000wr900336

Schaefer, G. L., Cosh, M. H., & Jackson, T. J. (2007). The USDA natural resources conservation service soil climate analysis network (SCAN). *Journal of Atmospheric and Oceanic Technology*, *24*(12), 2073–2077. https://doi.org/10.1175/2007jtecha930.1

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., et al. (2010). Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews*, *99*(3–4), 125–161. https://doi.org/10.1016/j.earscirev.2010.02.004

Sharma, A., Wasko, C., & Lettenmaier, D. P. (2018). If precipitation extremes are increasing, why aren't floods? *Water Resources Research*, *54*(11), 8545–8551. https://doi.org/10.1029/2018wr023749

Singh, N. K., Emanuel, R. E., McGlynn, B. L., & Miniat, C. F. (2021). Soil moisture responses to rainfall: Implications for runoff generation. *Water Resources Research*, *57*(9), e2020WR028827. https://doi.org/10.1029/2020wr028827

Taylor, C. M., de Jeu, R. A., Guichard, F., Harris, P. P., & Dorigo, W. A. (2012). Afternoon rain more likely over drier soils. *Nature*, *489*(7416), 423–426. https://doi.org/10.1038/nature11377

Tong, C., Wang, H., Magagi, R., Goïta, K., & Wang, K. (2021). Spatial gap-filling of SMAP soil moisture pixels over Tibetan plateau via machine learning versus geostatistics. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 9899–9912. https://doi.org/10.1109/jstars.2021.3112623

Tran, T.-N.-D., Nguyen, B. Q., Zhang, R., Aryal, A., Grodzka-Łukaszewska, M., Sinicyn, G., & Lakshmi, V. (2023). Quantification of gridded precipitation products for the streamflow simulation on the Mekong river basin using rainfall assessment framework: A case study for the Srepok river Subbasin, central highland Vietnam. *Remote Sensing*, *15*(4), 1030. https://doi.org/10.3390/rs15041030

Wagner, W., Hahn, S., Kidd, R., Melzer, T., Bartalis, Z., Hasenauer, S., et al. (2013). The ASCAT soil moisture product: A review of its. *Meteorologische Zeitschrift*, *22*, 1–29.

Wang, G., Garcia, D., Liu, Y., De Jeu, R., & Dolman, A. J. (2012). A three-dimensional gap filling method for large geophysical datasets: Application to global satellite soil moisture observations. *Environmental Modelling & Software*, *30*, 139–142. https://doi.org/10.1016/j.envsoft.2011.10.015

Wang, K., & Dickinson, R. E. (2012). A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability. *Reviews of Geophysics*, *50*(2), RG2005. https://doi.org/10.1029/2011rg000373

Wasko, C., & Nathan, R. (2019). Influence of changes in rainfall and soil moisture on trends in flooding. *Journal of Hydrology*, *575*, 432–441. https://doi.org/10.1016/j.jhydrol.2019.05.054

Wasko, C., & Sharma, A. (2014). Quantile regression for investigating scaling of extreme precipitation with temperature. *Water Resources Research*, *50*(4), 3608–3614. https://doi.org/10.1002/2013wr015194

Wasko, C., & Sharma, A. (2017). Global assessment of flood and storm extremes with increased temperatures. *Scientific Reports*, *7*(1), 7945. https://doi.org/10.1038/s41598-017-08481-1

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (p. 100). Academic press.

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., et al. (2012). Continental-scale water and energy flux analysis and validation for North American land data assimilation system project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow. *Journal of Geophysical Research*, *117*(D3), D03110. https://doi.org/10.1029/2011jd016051

Xiao, Z., Jiang, L., Zhu, Z., Wang, J., & Du, J. (2016). Spatially and temporally complete satellite soil moisture data based on a data assimilation method. *Remote Sensing*, *8*(1), 49. https://doi.org/10.3390/rs8010049

Zhang, Q., Yuan, Q., Li, J., Wang, Y., Sun, F., & Zhang, L. (2021). Generating seamless global daily AMSR2 soil moisture (SGD-SM) long-term products for the years 2013–2019. *Earth System Science Data*, *13*(3), 1385–1401. https://doi.org/10.5194/essd-13-1385-2021

Zhang, R., Kim, S., & Sharma, A. (2019). A comprehensive validation of the SMAP Enhanced Level-3 Soil Moisture product using ground measurements over varied climates and landscapes. *Remote Sensing of Environment*, *223*, 82–94. https://doi.org/10.1016/j.rse.2019.01.015

Zhang, R., Kim, S., Sharma, A., & Lakshmi, V. (2021). Identifying relative strengths of SMAP, SMOS-IC, and ASCAT to capture temporal variability. *Remote Sensing of Environment*, *252*, 112126. https://doi.org/10.1016/j.rse.2020.112126

Zheng, C., Jia, L., & Zhao, T. (2023). A 21-year dataset (2000–2020) of gap-free global daily surface soil moisture at 1-km grid resolution. *Scientific Data*, *10*(1), 139. https://doi.org/10.1038/s41597-023-01991-w