High-resolution surface and rootzone soil moisture over US cropland: A novel framework assimilating multi-source remote sensing data, machine learning, and the Layered Green and Ampt Infiltration with Redistribution model

Shuohao Cai; Yijia Xu; Zhengwei Yang; Wade T. Crow; Zhou Zhang; Jiali Shang; Jiangui Liu; Peter La Follette; Chris Reberg-Horton; Harry Schomberg; et al.

Remote Sensing of Environment 334 (2026) 115167

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

High-resolution surface and rootzone soil moisture over US cropland: A novel framework assimilating multi-source remote sensing data, machine learning, and the Layered Green and Ampt Infiltration with Redistribution model

Shuohao Cai [a,*], Yijia Xu [b], Zhengwei Yang [c], Wade T. Crow [d], Zhou Zhang [b], Jiali Shang [e], Jiangui Liu [e], Peter La Follette [f], Chris Reberg-Horton [g], Harry Schomberg [h], Steven Mirsky [h], Brian Davis [g], Sarah Seehaver [g], Alexis Correira [i], Andrea Basche [j], Ashley Waggoner [a], Charles Ellis [k], Dara Park [l], Danielle D. Treadwell [m], David Campbell [m], Deann Presley [i], Esleyther L. Henriquez Inoa [n], Heather Darby [o], Jared Adam [p], Jarrod Miller [q], Joseph Haymaker [r,v], John Wallace [s], Julia Gaskin [t], Kipling S. Balkcom [u], Lindsey Ruhl [o], Mark Reiter [r,v], Matthew Ruark [a], Michael Flessner [v], Cynthia Sias [v], Payton Davis [l], Peter Tomlinson [i], Richard G. Smith [w], Nicholas D. Warren [w], Ryan Dierking [x], Shalamar Armstrong [y], Tauana Almeida [j], Jingyi Huang [a,*]

[a] Department of Soil and Environmental Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA
[b] Department of Biological Systems Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA
[c] Research and Development Division, National Agricultural Statistics Service, United States Department of Agriculture, Washington, DC 20250, USA
[d] USDA ARS Hydrology and Remote Sensing Laboratory, Beltsville, MD 20705, USA
[e] Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, ON K1A 0C6, Canada
[f] Lynker, Boulder, CO 80301, USA
[g] Department of Crop and Soil Sciences, North Carolina State University, 840 Oval Dr, Raleigh, NC 27606, USA
[h] USDA-ARS Sustainable Agricultural Systems Laboratory, Beltsville, MD 20705, USA
[i] Department of Agronomy, Kansas State University, Manhattan, KS 66506, USA
[j] Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68583, USA
[k] University of Missouri Extension, MO 63379, USA
[l] Department of Plant and Environmental Sciences, Clemson University, Clemson, SC 29634, USA
[m] Department of Horticultural Sciences, University of Florida, Gainesville, FL 32611, USA
[n] Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC 27695, USA
[o] Northwest Crops and Soil Program, University of Vermont Extension, Burlington, VT 05478, USA
[p] College of Agriculture, Land Resources and Environmental Sciences, Montana State University, Bozeman, MT 59718, USA
[q] Department of Plant and Soil Sciences, University of Delaware, Georgetown, DE 19947, USA
[r] Eastern Shore Agricultural Research and Extension Center, Virginia Tech, Painter, VA 23420, USA
[s] Plant Science Department, The Pennsylvania State University, University Park, PA 16802, USA
[t] Department of Crop and Soil Sciences, University of Georgia, Athens, GA 30602, USA
[u] USDA-ARS National Soil Dynamics Laboratory, Auburn, AL 36832, USA
[v] School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA 24061, USA
[w] Department of Natural Resources and the Environment, University of New Hampshire, Durham, NH 03824, USA
[x] EarthOptics formally Indigo Ag, Lafayette, IN 22202, USA
[y] School of Agronomy, Purdue University, West Lafayette, IN 47907, USA

ARTICLE INFO

Edited by: Dr. Jing M. Chen

ABSTRACT

Accurate and high spatiotemporal resolution soil moisture (SM) monitoring in cropland is important for water resource management, drought forecasting, and nutrient transport estimation at the field scale for sustainable

* Corresponding authors.
E-mail addresses: scai62@wisc.edu (S. Cai), jhuang426@wisc.edu (J. Huang).

https://doi.org/10.1016/j.rse.2025.115167
Received 14 April 2025; Received in revised form 28 October 2025; Accepted 30 November 2025
Available online 6 December 2025

Keywords:
Machine learning
LGAR model
Ensemble Kalman filter
Soil moisture
High resolution

crop production. Although recent research has applied machine learning (ML) to downscale coarse-resolution satellite SM products, most of this past work has focused only on surface SM estimation, and the performance of rootzone SM products has not been intensively evaluated in cropland. This study introduces a novel framework that integrates multi-source satellite-based ML models with the Layered Green and Ampt Infiltration with Redistribution (LGAR) model to produce high-resolution (100 m, hourly) SM products for both the surface layer (0–5 cm) and rootzone (0–100 cm) across cropland in the contiguous United States (CONUS). First, six ML models were trained using multiple high-resolution remote sensing datasets (Sentinel-1, Sentinel-2, and Landsat) to predict surface and rootzone SM. These ML predictions were then assimilated into the LGAR model using the ensemble Kalman filter (EnKF). The framework was developed and validated using an eight-fold cross-validation scheme with in-situ data from 431 cropland sites across CONUS, sourced from three networks (SCAN, USCRN, and PSA). The 100-m hourly SM data from this framework surpasses existing products (9-km SMAP L4, SMAP-based 1-km thermal hydraulic disaggregation of SM product) in spatial and temporal resolution and captures rootzone SM that is not available in the SMAP-HydroBlocks SM product. It achieves good performance, with median bias-corrected root mean squared error (ubRMSE) of 0.053 m^3/m^3 and median Kling-Gupta efficiency (KGE) of 0.379 in the surface layer, and median ubRMSE of 0.027 m^3/m^3 and median KGE of 0.302 in the rootzone. While the framework demonstrates strong performance, its accuracy varies across climatic regimes, with surface SM performing better in non-humid areas (median KGE = 0.375 versus median KGE = 0.416) and rootzone SM in humid regions (median KGE = 0.313 versus median KGE = 0.127). This high-resolution cropland SM product can potentially benefit multiple agricultural applications, such as irrigation management and nutrient leaching estimation, and provide valuable insights to support farmers and land managers in decision-making processes.

1. Introduction

As a key state variable in terrestrial ecosystems, soil moisture (SM) plays an important role in regulating the terrestrial water, carbon, and energy cycles. Multiple applications such as flood forecasting (Massari et al., 2014; Wang et al., 2023), wildfire modeling (Hou et al., 2020; Krueger et al., 2015), agricultural drought (Cai et al., 2021; Sheffield et al., 2009), and irrigation demand monitoring (Karthikeyan et al., 2020; Lawston et al., 2017) rely on high-resolution and spatiotemporal-continuous SM information. In agricultural management, high spatio-temporal SM with the ability to capture irrigation signals is required to monitor water status over cropland to estimate water demands and predict crop yield (Kumar et al., 2015; Lawston et al., 2017). Moreover, rootzone SM is increasingly desired in many agricultural applications (Li et al., 2021), and an accurate estimate of rootzone SM can extend agricultural applications of SM toward environmental monitoring, such as nitrate leaching estimation (Arriaga et al., 2009; Jyotiprava Dash et al., 2015).

Rapid development of remote sensing makes it possible to produce large-scale SM data sets across different scales, and many studies have already explored SM retrieval using a range of remote sensing approaches applied at various electromagnetic frequency bands (Li et al., 2021; Wang and Qu, 2009; Zhang and Zhou, 2016). Among these, the microwave L-band is recognized as the most suitable for SM retrieval because of its direct physical linkage with SM and its ability to penetrate both cloud and moderate vegetation canopy (Babaeian et al., 2019; Li et al., 2021). These capacities have led to the rapid development of SM retrieval algorithms and products based on microwave remote sensing measurements in the last two decades. For example, L-band passive microwave brightness temperature observations obtained from satellite missions such as the NASA Soil Moisture Active Passive (SMAP) and ESA Soil Moisture and Ocean Salinity (SMOS) missions have been applied to retrieve surface SM (Al Bitar et al., 2012; Chan et al., 2018; Entekhabi et al., 2010a; Kerr et al., 2012; Wigneron et al., 2017). However, these passive L-band products suffer from coarse spatial resolutions (generally >25 km) and penetrate only the first 5 cm of the soil column. As a result, they generally cannot meet the requirements of farm- or field-scale agricultural applications (Sabaghy et al., 2020).

A range of spatial downscaling approaches has been proposed to generate higher-resolution SM estimates, including statistical and machine learning (ML) approaches and physics-based disaggregation. ML-based approaches typically establish empirical relationships between high-resolution input features (e.g., optical and thermal remote sensing data) and coarse-resolution SM products, such as SMAP SM (Sishah et al., 2023; Xu et al., 2022), or SMAP, other high-resolution variables, and in-situ measurements (Peng et al., 2024). These approaches can potentially enhance data quality and provide higher spatial data resolution in comparison with the original microwave data (Liu et al., 2020; Srivastava et al., 2013). However, they depend heavily on the availability of high-resolution input features and in-situ SM networks, particularly under intensively managed landscapes (e.g., cropland), limiting their applicability when such data are scarce. Physics-based disaggregation methods, such as the Physical And Theoretical scale Change algorithm, leverage physical relationships between SM and auxiliary factors like vegetation, surface temperature, soil properties, and Digital Elevation Model (DEM) data to produce high-resolution SM products, such as SMOS-disaggregated 1-km SM (Molero et al., 2016) and SMAP-based 1-km thermal hydraulic disaggregation of soil moisture (ThySM) (Liu et al., 2022). Even though these approaches have been widely used to produce high spatial resolution SM, they have not yet been shown to add value to existing coarse-scale observations objectively (Crow et al., 2022). While physics-based disaggregation methods can preserve water balance, they often struggle in agricultural areas due to human-induced changes (e.g., irrigation), and they may also fail when critical auxiliary data (e.g., land surface temperature) are unavailable. Moreover, both types of downscaling methods face challenges in estimating rootzone SM due to the general lack of training data for deeper soil layers.

Process-based or land surface models provide an alternative approach to producing SM by simulating SM dynamics at high temporal resolution for both the surface layer and rootzone. However, these models require accurate meteorological forcing (e.g., precipitation, temperature, radiation) and detailed soil hydraulic property data, as well as precise parameterizations of physical processes at relevant scales (Koster et al., 2009; Seneviratne et al., 2010). Irrigation-parameterized process-based models/land surface models (Ozdogan et al., 2010; Pokhrel et al., 2016) are a promising way to produce SM products in agricultural areas. However, it is still a challenge to represent human-engineered modifications in land surface models (Kumar et al., 2015) - especially over a large spatial extent where irrigation practices are uncertain.

Integrating process-based or land surface models with remote sensing through land data assimilation systems or satellite-informed modeling frameworks can enhance signal depth and generate high-resolution, continuous SM estimates (Rebel et al., 2012; Reichle et al., 2019; Reichle and Koster, 2005; Rodell et al., 2004; Vergopolan et al., 2020). For example, the SMAP L4 product assimilates SMAP brightness temperature observations into a land surface model, providing surface and rootzone SM estimates at a spatiotemporal resolution of 3 h and 9 km (Dong et al., 2019; Reichle et al., 2019). SMAP-HydroBlocks (SMAP-HB), in another way, combines SMAP passive microwave observations (brightness temperature) with a hyper-resolution 3-dimensional distributed land surface model to produce high-resolution surface SM at 30-m resolution with a temporal resolution of approximately two days (due to the availability of SMAP passive microwave data, see Vergopolan et al., 2020). Despite their potential, both SMAP L4 and SMAP-HB are limited by the coarse spatial resolution of the SMAP observations and do not utilize high-resolution remote sensing observations (i.e., <100 m), which is essential for capturing fine-scale SM variability. Such high-resolution data, particularly from optical or thermal sensors, may provide valuable information for capturing fine-scale processes, such as small-patch irrigation signals in agricultural applications or the uneven distribution of SM due to micro-topography or soil heterogeneity, which are missing in current SM products (Lawston et al., 2017; Vergopolan et al., 2020, 2021).

To address these limitations, this study develops a novel data assimilation framework for generating high-resolution SM estimates. The framework integrates six ML models, each trained to combine ancillary environmental covariates (i.e., soil properties, topographical features, and land climate data) with distinct high-resolution remote sensing data sources (Sentinel-1, Sentinel-2, or Landsat), to predict SM for both the surface soil layer (0–5 cm) and the rootzone (0–100 cm). These ML predictions are then assimilated into a process-based model, the Layered Green and Ampt Infiltration with Redistribution (LGAR) model, using the Ensemble Kalman Filter (EnKF) algorithm. The LGAR model was selected for its computational efficiency, reliability, and robustness (La Follette et al., 2023), making it a practical choice to simulate high-resolution SM over large areas. A detailed description of this framework is provided below.

By fully leveraging the potential of high-resolution remote sensing data with ML, the proposed framework can map surface and rootzone SM at fine spatial and temporal resolutions within agricultural areas of the contiguous United States (CONUS), with the potential to be applied outside CONUS. The resulting SM system has the potential to offer near-real-time insight into soil water status over cropland and decision support for farm-level agricultural management.

2. Materials and methods

2.1. Study area

The CONUS is chosen as the study area. In total, 431 cropland in-situ SM monitoring sites from 3 networks were used in this study. This includes the Soil Climate Analysis Network (SCAN), the U.S. Climate Reference Network (USCRN), and the Precision Sustainable Agriculture (PSA) network.

2.2. Data

This study utilized a comprehensive dataset comprising in-situ SM observations, high-resolution remote sensing data (Sentinel-1, Sentinel-2, and HLSL30), ERA5-Land hourly climate data (precipitation, 2-m air temperature, actual evapotranspiration, 10-m wind speed, surface net solar radiation, surface net thermal radiation, 2-m dew point temperature, and atmospheric surface pressure), Probabilistic Remapping of SSURGO (POLARIS) soil properties (sand/clay/silt content, bulk density, and saturated hydraulic conductivity), and DEM and its derivatives (slope, aspect, topographic wetness index). The in-situ SM observations were used to train and validate the ML models and the data assimilation framework. Input features for the ML models include high-resolution remote sensing data, soil properties, and topographic features (DEM, slope, aspect, and topographic wetness index). The LGAR model was driven by precipitation and potential evapotranspiration (PET), the latter calculated using surface net solar radiation and 2-m air temperature from the EAR5-Land hourly dataset based on the Priestley-Taylor method. Soil texture for each layer at each site was determined using sand, clay, and silt content from the POLARIS dataset, classified according to the USDA soil texture triangle, and a lookup table embedded in the LGAR model provided the Van Genuchten (VG) parameters required by the model. For comparative analysis, SMAP L4 products and two other high-resolution state-of-the-art SM products, SMAP-HydroBlocks (Vergopolan et al., 2021) and SMAP-based 1-km THySM (Liu et al., 2022), were collected and evaluated alongside the SM estimates generated in this study.

2.2.1. In-situ SM observations

The in-situ SM observations of crop sites from SCAN and USCRN were collected from the International Soil Moisture Network (ISMN). The ISMN is an international cooperation platform that standardizes a global in-situ SM database assembled across a large number of regional soil monitoring networks. In this study, the hourly SM observations from 2016 to 2024 within the SCAN (Schaefer et al., 2007) and the USCRN (Bell et al., 2013) networks were collected. As shown in Fig. 1, 127 crop/hay/pasture sites were chosen within these networks based on land cover data (USGS NLCD; Dewitz, 2021).

The SM sensors used by SCAN include Hydraprobe Digital SDI-12 (2.5 Volt), n.s., Hydraprobe Analog (2.5 Volt), Hydraprobe Analog (5.0 Volt), and Hydraprobe Digital SDI-12 Thermistor (linear). The sensor used by USCRN is Stevens Hydraprobe II SDI-12. In this study, SM observations with the ISMN flag labeled as 'G' (good), 'C02' (SM > 0.6 m^3/m^3), 'C03' (SM > saturation point derived from Harmonized World Soil Database parameter values), or 'C02,C03' (SM > 0.6 m^3/m^3 and SM > saturation point) were used to filter good-quality observations (Dorigo et al., 2013). The 'C02' and 'C03'-labeled SM were considered valid values because the oversaturated conditions are physically possible (e.g., long-lasting heavy rainfall). The typical SM depths provided by these two networks include six layers: 5 cm, 10 cm, 20 cm, 30 cm, 50 cm, and 100 cm. This study focuses on the 0–5 cm surface layer and the 0–100 cm integrated rootzone. The rootzone SM observations were obtained by taking a depth-weighted average of SM values across six depths. To ensure the quality of rootzone SM, only the observation cases where the SMs of the surface layer (5 cm), the deepest layer (100 cm), and at least one middle layer were available were considered. Linear interpolation was applied to fill missing values before calculating the depth-weighted average of SM.

The PSA network, including 304 cropland sites, provides another set of in-situ SM observations (https://github.com/precision-sustainable-ag) (Thapa et al., 2022). This network provides SM monitoring of many eastern US states (as shown in Fig. 1). The sensors provided by the PSA network were installed at 5, 15, 45, and 80 cm in cultivated cropland and based on time-domain reflectometry (TDR). One variable named *is_vwc_outlier* was used to decide whether the SM data point is an outlier. The 5 cm SM was taken as surface SM, and the rootzone SM was calculated using the same method as proposed for SCAN and USCRN networks.

2.2.2. Remote sensing data

The remote sensing data used in this study include Sentinel-1 (VV, VH, and angle bands), Sentinel-2 (B2, B3, B4, B5, B6, B7, B8, B8A, B11, and B12), and HLSL30 (B2, B3, B4, B5, B6, B7, B10, and B11). The Harmonized Landsat Sentinel-2 (HLS) products provide compatible surface reflectance (SR) data from a virtual constellation of satellite sensors, including the Operational Land Imager (OLI) onboard Landsat-8/9 and the Multispectral Instrument (MSI) onboard Sentinel-2 satellites (Masek et al., 2021). This study used the HLSL30 product, which only contains Landsat 8/9 images as a combination of Landsat data.

In addition to the original bands from Sentinel-1, Sentinel-2, and HLSL30, several indices were derived to enhance feature representation.

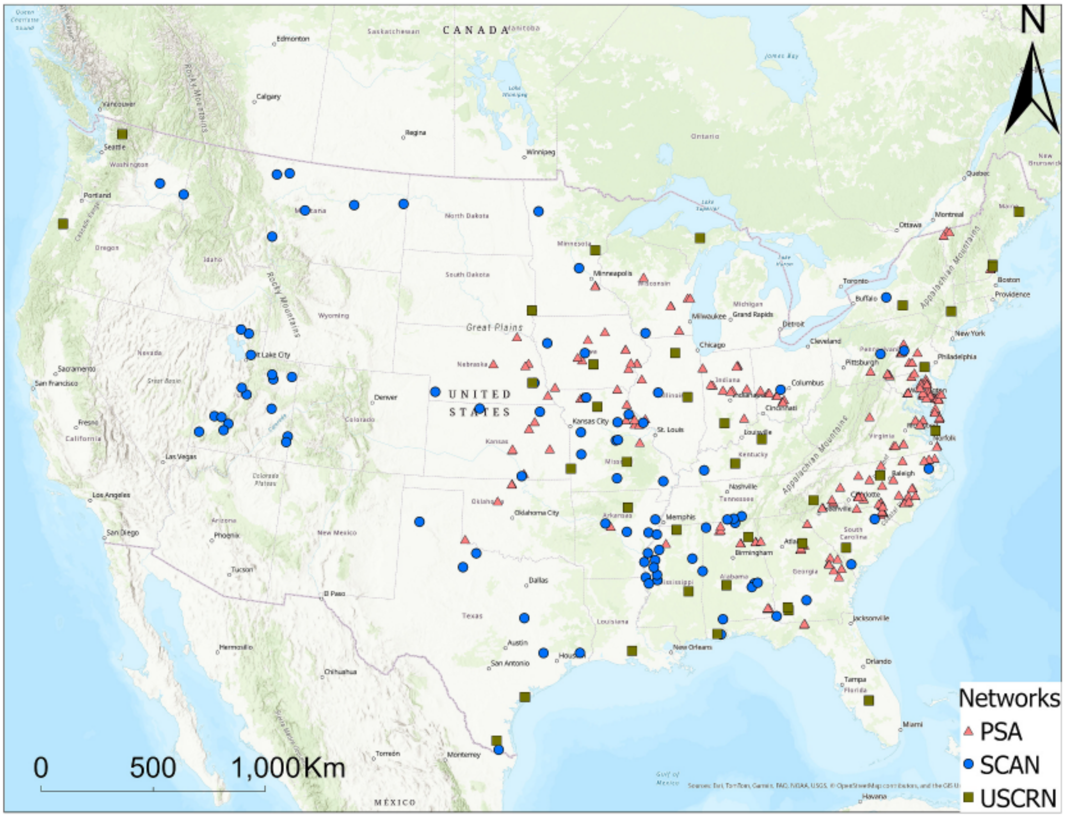


Fig. 1. Study area and the distribution of crop sites from three networks: Soil Climate Analysis Network (SCAN), U.S. Climate Reference Network (USCRN), and Precision Sustainable Agriculture (PSA) network.

From Sentinel-1 data, the cross-polarization ratio (CR), modified Dual Polarization SAR Vegetation Index (DPSVIm), and modified Radar Vegetation Index (Pol) were calculated, providing valuable information on vegetation structure and soil characteristics (Yu et al., 2024). From the surface reflectance bands of Sentinel-2 and HLSL30, the normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), normalized difference water index (NDWI), and tasseled cap transformation components (greenness, brightness, and wetness) were derived. Specifically, the formulas used to calculate the CR, DPSVIm, Pol, NDVI, EVI, and NDWI are shown as follows:

$$CR = \frac{VV}{VH} \tag{1}$$

$$DPSVIm = \frac{VV^2 + VH^2}{\sqrt{2}} \tag{2}$$

$$Pol = \frac{VV + VH}{VV - VH} \tag{3}$$

$$NDVI = \frac{NIR - R}{NIR + R} \tag{4}$$

$$EVI = 2.5\left(\frac{NIR - R}{NIR + 6{*}R - 7.5{*}B + 1}\right) \tag{5}$$

$$NDWI = \frac{G - NIR}{G + NIR} \tag{6}$$

where VV and VH are backscatter coefficients of Sentinel-1; NIR, R, G, and B are the near-infrared, red, green, and blue bands, and their corresponding band numbers of Sentinel-2 and HLSL-30 products are shown in Table 1.

Tasseled cap transformation (TCT) is a commonly used technique that can transform the reflectance data into a three-dimensional feature space composed of greenness, brightness, and wetness (Shi and Xu, 2019). However, the TCT coefficients are sensor-specific. Table 1 shows the TCT coefficients used in this study for Sentinel-2 and HLSL30, derived by Shi and Xu (2019) and Baig et al. (2014), respectively.

The remote sensing data was preprocessed and downloaded from the Google Earth Engine (GEE) platform. The images were first resampled to a spatial resolution of 100 m using the nearest interpolation method, and values corresponding to study sites were extracted. These high-resolution remote sensing data were used as input features for ML models.

2.2.3. *Meteorological data*

The meteorological data employed in this study includes precipitation (Prec), actual evapotranspiration (AET), 10-m wind speed (vs), surface net solar radiation (SSR), surface net thermal radiation (STR), 2-m dew point temperature (T_{dew_2m}), 2-m air temperature (T_{air_2m}), and atmospheric surface pressure (Pressure) collected from ERA5-Land hourly data (Muñoz-Sabater et al., 2021). In addition, potential evapotranspiration (PET) was calculated using SSR and T_{air_2m} based on the Priestley-Taylor method. The ERA5-Land hourly data was selected due to its availability on the Google Earth Engine (GEE) platform, which facilitates both preprocessing and downloading for model training. All the meteorological variables were used as input features for ML models. In addition, Prec and PET were two dynamic inputs used to drive the LGAR model.

2.2.4. *POLARIS soil properties*

The POLARIS soil properties dataset is a database of 30-m probabilistic soil property maps over the CONUS (Chaney et al., 2019). Here, POLARIS estimates of soil texture (sand, clay, and silt content), saturated hydraulic conductivity (K_{sat}), and bulk density (D_b) from this dataset were utilized. The clay and sand content, K_{sat}, and D_b were used

Table 1
TCT coefficients for Sentinel-2 and HLSL30.

Sensor	TCT	B	G	R	RE 1	RE 2	RE 3	RE 4	NIR	SWIR1	SWIR2	TIRS1	TIRS2
Sentinel-2		B2	B3	B4	B5	B6	B7	B8A	B8	B11	B12		
	brightness	0.3510	0.3813	0.3437	–	–	–	–	0.7196	0.2396	0.1949		
	greenness	0.3599	0.3533	0.4734	–	–	–	–	0.6633	0.0087	0.2856		
	wetness	0.2578	0.2305	0.0883	–	–	–	–	0.1071	0.7611	0.5308		
HLSL30		B2	B3	B4					B5	B6	B7	B10	B11
	brightness	0.3029	0.2786	0.4733					0.5599	0.508	0.1872	–	–
	greenness	0.2941	0.243	0.5424					0.7276	0.0713	- 0.1608	–	–
	wetness	0.1511	0.1973	0.3283					0.3407	- 0.7117	- 0.4559	–	–

Notes: B: Blue; G: Green; R: Red; RE: Red Edge; NIR: Near infrared; SWIR: Short wave infrared; TIRS: Thermal infrared; TCT: tasseled cap transformation components.

as input features for ML models. To estimate rootzone clay and sand content as well as bulk density (D_b), depth-weighted averages were calculated based on soil properties across multiple layers. The effective K_{sat} of the rootzone is determined using the following equation (Jury and Horton, 2004):

$$K_{sat,eff} = \frac{\sum_{J=1}^{N} L_J}{\sum_{J=1}^{N} L_J / K_{sat,J}} \tag{7}$$

where the $K_{sat,J}$ and $K_{sat,eff}$ are the saturated hydraulic conductivity of the J-th soil layer and the effective saturated hydraulic conductivity of the rootzone ($J = 1, 2, \ldots, 5$); and L_J is the depth of the J-th soil layer.

The soil names were determined by classifying sand, silt, and clay content using the USDA soil texture triangle, which was used to extract soil Van Genuchten (VG) parameters from the Lookup table in the LGAR model.

2.2.5. Topography and land cover data

In addition to time-varying remote sensing data and constant soil properties, other time-constant variables, including a DEM (USGS 3DEP DEM) and DEM-derived properties (slope, aspect, and topographic wetness index), were also collected and calculated as inputs into ML models. The land cover (USGS NLCD: Dewitz, 2021) data, which has an original spatial resolution of 30 m, provides a cropland mask for cropland SM mapping.

2.2.6. SM products

Several SM products were collected to compare with our SM-producing framework. They include the SMAP Level 4 (L4) Global 3-hourly 9 km EASE-Grid Surface and Root Zone SM product (Reichle et al., 2018), SMAP-HydroBlocks (Vergopolan et al., 2021), and SMAP-based 1-km THySM (Liu et al., 2022).

The SMAP L4 is an analysis product based on the assimilation of the SMAP L1_C surface brightness temperature data into the Catchment Land Surface model using an EnKF. It provides SM in both the surface layer (0–5 cm) and rootzone (0–100 cm) with a temporal resolution of 3 h. Based on extensive validation at SMAP core validation sites, the bias-corrected root mean squared error of SMAP L4 surface and rootzone SM products is 0.039 m^3/m^3 and 0.029 m^3/m^3, respectively (Reichle et al., 2018). The SMAP data were preprocessed and downloaded from the GEE platform.

SMAP-HydroBlocks and SMAP-based 1-km THySM are two cutting-edge SM products. They were collected and used as benchmarks for the performance of our SM estimates.

2.3. Methods

Here, we developed a new SM-producing framework by integrating ML results into a process-based model (i.e., the LGAR model) using the EnKF algorithm to produce high-resolution SM (100 m, hourly). To start, six ML models (ML-1S/R: Sentinel-1 with other features, ML-2S/R: Sentinel-2 with other features, and ML-3S/R: HLSL30 with other features) were developed for the surface layer (0–5 cm) and rootzone (0–100 cm) (see Section 2.3.1 for more details). Next, SM estimates from three ML models in the surface layer or rootzone were combined to provide SM estimates every 2 to 5 days, depending on the revisit time of Sentinel-1, Sentinel-2, or Landsat satellites and the quality of data they collected. Finally, the ML estimates were incorporated into the LGAR model using the EnKF algorithm to produce high-resolution and temporally continuous SM. The overall flowchart of this study is shown in Fig. 2.

2.3.1. Machine learning models

ML algorithms are widely used for SM assessment due to their ability to model the nonlinear relationship between variables and SM. In this study, three commonly-used ML algorithms, including Support Vector Regression (SVR), Random Forest (RF) regression, and Light Gradient Boosting (LGB) Machine, were pre-selected to estimate SM (Rani et al., 2022; Teshome et al., 2024).

For each ML algorithm, six separate ML models were developed, designated as ML-1S, ML-1R, ML-2S, ML-2R, ML-3S, and ML-3R, where "S" and "R" refer to the surface layer (0–5 cm) and rootzone (0–100 cm), respectively. Three pairs of models (ML-1S/1R, ML-2S/2R, and ML-3S/3R) for the surface layer and rootzone are distinguished by the use of different high-resolution remote sensing data sources: Sentinel-1 for ML-1S/1R, Sentinel-2 for ML-2S/2R, and HLSL30 for ML-3S/3R. The input features used by these models are listed in Table 2. Satellite overpass times were used to synchronize instantaneous remote sensing data with hourly input features (e.g., meteorological data) and target SM observations for model training. This alignment ensures the reliability of the ML models by matching data acquisition times at an hourly resolution. To enhance the performance of the ML models in capturing extreme SM conditions, SM observations falling within the upper 80th percentile and lower 20th percentile of SM were given higher weights during model training. Specifically, these observations were assigned a weight of 2.0 using the *sample_weight* parameter when fitting the models.

The following procedures were applied to determine the best ML algorithm and the optimal ML models. First, each algorithm was associated with a predefined hyperparameter search space, as detailed in Table S1. The *Optimize* function from the *optuna* (Version: 4.3.0) package was used to search for the optimal parameters for each ML model. During the hyperparameter tuning process, five-fold spatial cross-validation was employed by randomly dividing the sites into five spatially independent groups to ensure that the optimized parameters enhanced spatial transferability. The tuned parameters were listed in Table S2. Secondly, after parameter selection, model performance was further evaluated using an eight-fold spatial cross-validation scheme. Specifically, an eight-fold cross-validation approach was implemented to assess the performance of each ML model. The 431 sites distributed across the CONUS were partitioned into eight spatially independent groups. During each iteration, 7/8th of the groups was utilized for model training, while the remaining 1/8th served as the validation set. This spatial cross-validation design was employed to rigorously evaluate the spatial transferability of these models. The best ML algorithm was

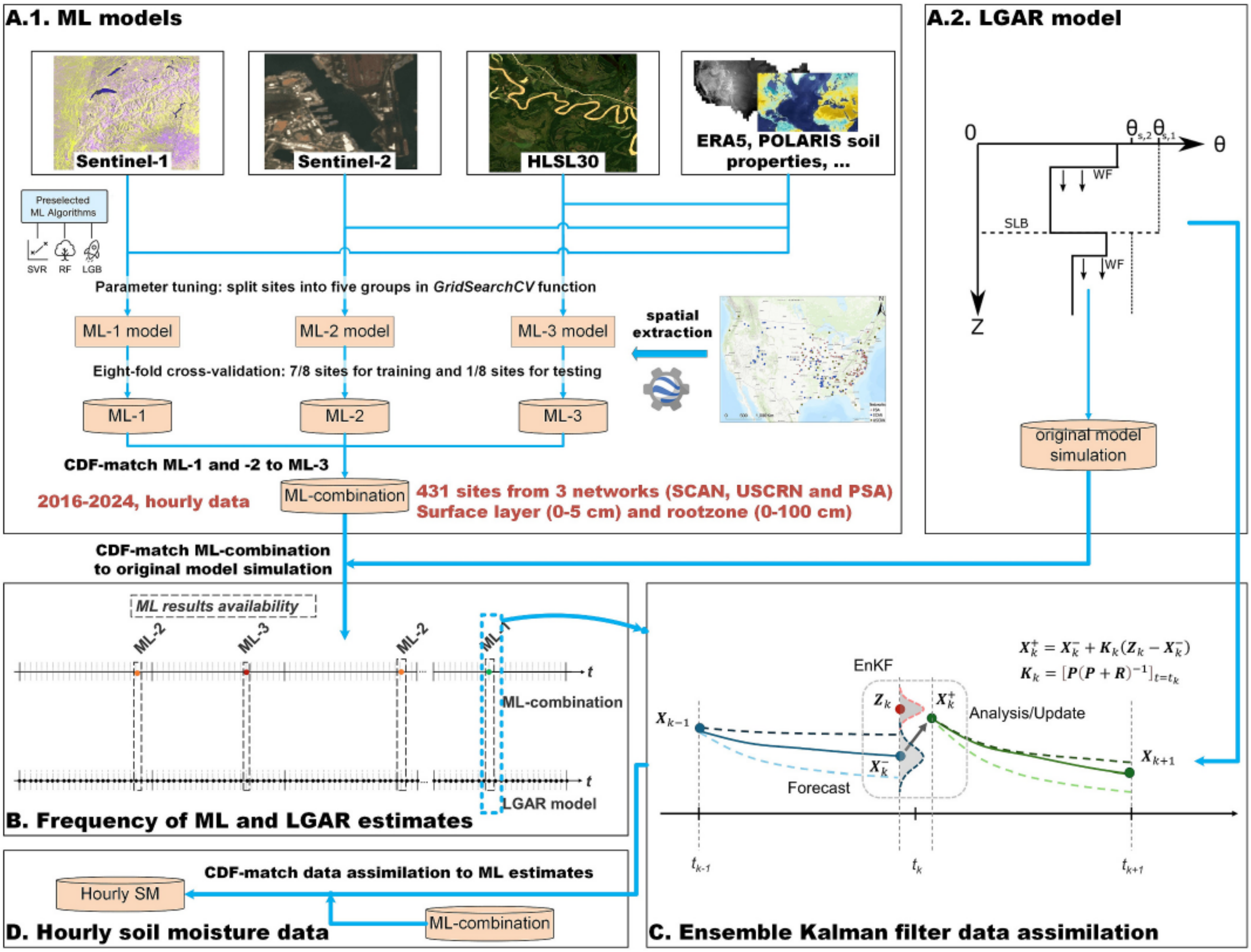


Fig. 2. Flowchart of the study. Subplot A.1 shows the development process of the ML models. Subplot A.2 demonstrates the mechanistic diagram of the LGAR model (see Section 2.3.2 for more details). Subplot B indicates the hourly availability of estimates from the ML models (available every 2 to 5 days, depending on the revisit time and data quality of high-resolution remote sensing datasets) and the LGAR model (available every hour). Subplot C shows the application of the EnKF to integrate ML-estimated SM into the LGAR model. The data assimilation results are finally CDF-matched to ML estimations as shown in Subplot D.

determined based on the validation results. The validation results from all iterations were subsequently aggregated to provide a comprehensive assessment of the performance of ML models and provide ML estimates as observations for the data assimilation system.

To facilitate data assimilation, the SM estimates from these ML models were first fused using a CDF (Cumulative Distribution Function) -matching approach (Eq. 8). This fusion process also enables the resulting ML-based estimates to function as a standalone product with high spatial (100 m) and temporal (2–5 days) resolution.

$$SM'_{i,j} = PPF_{3,j}\left(CDF_{i,j}\left(SM_{i,j}\right)\right) \tag{8}$$

where $SM_{i,j}$ refers to SM estimates produced by ML-i ($i = 1, 2$, and each with two dimensions: the surface layer and the rootzone SM) at site j; $PPF_{3,j}$ is the empirical Percentage Point Function (i.e., the inverse CDF) of ML-3 at site j. In this way, the climatology of ML-1 and ML-2 estimates can be matched to that of ML-3 estimates for both the surface layer and rootzone. The ML combination can then be defined as the union of three sets:

$$SM_{ML,j} = SM'_{1,j} \cup SM'_{2,j} \cup SM_{3,j} \tag{9}$$

If multiple ML models estimate SM at the same hour, the one with the lowest uncertainty is retained, though this condition is rare.

To estimate the uncertainty of ML estimates, which is an essential parameter for the application of EnKF algorithms, the quantile regression models were also fitted using the function *ConformalizedQuantileRegressor* from the *mapie* package (Version: 1.0.1) in Python. Quantile regression can model the relationship between a set of predictors and specific quantiles (or percentiles) of the target variable. In this study, the 0.95 ($Q_{0.95}$) and 0.05 ($Q_{0.05}$) quantiles of SM were predicted to calculate the uncertainty. The uncertainty (standard deviation) of ML estimates was then calculated using the following formula, given the quantiles.

$$\sigma_{ML} = \frac{Q_{0.95} - Q_{0.05}}{2Z_{0.9}} = \frac{Q_{0.95} - Q_{0.05}}{3.29} \tag{10}$$

Here, $Z_{0.9}$ is the z score for a 90 % confidence interval (i.e., 1.645).

2.3.2. *Layered green and Ampt infiltration with redistribution model*

The LGAR model is a newly developed process-based model that extends the capabilities of the earlier Green and Ampt infiltration (GA) and Green and Ampt infiltration with Redistribution (GAR) models, as illustrated in Fig. 3. In the original GA model, a single wetting front, representing saturated soil above and unsaturated soil below, is formed following a rainfall event and subsequently advances downward. The GAR model generalizes this concept by allowing for multiple,

Table 2
Input features for six ML models. ML-1, ML-2, and ML-3 are ML models trained using different variables indicated by different high-resolution remote sensing features, and "S" and "R" following the numbers denote the surface layer and rootzone, respectively.

Data sources	Variables	ML-1S/R	ML-2S/R	ML-3S/R
USGS 3DEP DEM	DEM, slope, aspect, TWI	x	x	x
POLARIS Soil properties*	K_{sat}, sand, clay, D_b	x	x	x
ERA5-Land	Prec, AET, vs, SSR, STR, T_{dew_2m}, T_{air_2m}, Pressure	x	x	x
Others	Day of year (DoY), coordinates (Latitude, Longitude)	x	x	x
Sentinel-1	VV, VH, angle, CR, DPSVIm, Pol	x		
Sentinel-2	B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12, NDVI, EVI, NDWI, greenness, wetness, brightness		x	
HLSL30	B2, B3, B4, B5, B6, B7, B10, B11, NDVI, EVI, NDWI, greenness, wetness, brightness			x

Notes: for POLARIS soil properties, we are using different depths of these variables to build ML models for the surface layer and rootzone. For example, the surface layer K_{sat} (0–5 cm) from the POLARIS dataset was used as an input to train the surface layer models—ML-1S, ML-2S, and ML-3S. In contrast, the rootzone effective K_{sat} (0–100 cm), derived using Eq. 7, served as an input for the corresponding rootzone models—ML-1R, ML-2R, and ML-3R. Meteorological variables were categorized into three groups based on their temporal influence: short-term (vs, T_{dew_2m}, T_{air_2m}, and Pressure), middle-term (SSR and STR), and long-term effective variables (Prec, AET). For short-term variables, weighted mean values over the preceding 24 h were computed, assigning greater weight to values closer to the current hour. For mid-term variables, weighted mean values were calculated over the previous 72 h with greater weight for values closer to the current hour. For long-term variables, the total sums over the past 120 h were used as inputs for the ML models with greater weight for values closer to the current hour.

unsaturated wetting fronts, but it remains limited to homogeneous soil profiles. The LGAR model builds upon these foundations by enabling simulation across multiple soil layers with varying soil properties, thereby allowing for the modeling of infiltration into stratified soils and the vertical redistribution of water throughout the soil column (La Follette et al., 2023). Rather than solving the full Richards Equation, this approach employs the Green-Ampt formulation to simulate infiltration and models the redistribution of wetting fronts based on water potential gradients and a water balance framework. This simplification significantly reduces computational demands relative to more complex land surface models (e.g., Noah-MP). Its combination of computational efficiency, reliability, and robustness makes it well-suited for simulating SM dynamics at fine spatial scales and hourly time steps. A comprehensive description of the LGAR model is available in La Follette et al. (2023).

2.3.3. Ensemble Kalman filter

Before applying the EnKF algorithm, the ML estimates (SM_{ML}) were first CDF-matched to the original LGAR model simulations to ensure no systematic bias exists between ML estimates and the background LGAR simulation.

$$SM'_{ML,j} = PPF_{LGAR,j}\left(CDF_{ML,j}\left(SM_{ML,j}\right)\right) \tag{11}$$

Here, $SM_{ML,j}$ are the SM estimates from the ML models at site j; and $CDF_{ML,j}$ is its corresponding empirical Cumulative Distribution Function, and $PPF_{LGAR,j}$ is the empirical Percentage Point Function (i.e., the inverse CDF) for LGAR simulations at site j.

Next, the EnKF algorithm is applied to develop a data assimilation framework to integrate the ML estimates into the LGAR model (Fig. 2).

Let $\mathscr{F}(\bullet)$ denote the nonlinear LGAR model, and then the model equation can be expressed as Eq. 12:

$$\frac{dX}{dt} = \mathscr{F}(Prec, PET; X) + w \tag{12}$$

Here, Prec (precipitation) and PET (potential evapotranspiration) are the two key dynamic inputs for the LGAR model; X is a state control vector with two members (surface layer and rootzone SM); X = $[SM_{LGAR\_sly}, SM_{LGAR\_rz}]$; and w represents the model structure error term.

At time t_k, the model state $X(t_k)$ can be converted into measurements using the following formula:

$$Z_k = \mathscr{M}[X(t_k)] + v_k \tag{13}$$

where Z_k is the measurement vector; the nonlinear operator $\mathscr{M}(\bullet)$ is used to relate the observations with measurements; v_k is the measurement error, which can be calculated from the standard deviation of ML results. Here, we treated ML results as measurements, $Z_k = [SM_{ML\_sly}, SM_{ML\_rz}]$, which have the same unit and dimension size as observations. Therefore, the nonlinear operator $\mathscr{M}(\bullet)$ is assumed to be the identity E, and the measurement process can be simply reduced to the following equation:

$$Z_k = E \bullet X(t_k) + v_k = X(t_k) + v_k \tag{14}$$

The EnKF uses an ensemble of model trajectories to calculate the covariance matrix for model forecasts at the time of an update. With an ensemble size of N_e, the forecast error covariance at the time t_k (k = 1, 2, …,) can be calculated as:

$$P = \frac{1}{N_e - 1}\sum \left(X^i - \overline{X}\right)\left(X^i - \overline{X}\right)^T \tag{15}$$

where $\overline{X}$ is the mean state of ensemble members, and X^i is the state of the i-th ($i = 1, 2, …, N_e$) ensemble member at the time t_k. In this study, three ensemble sizes (20, 50, and 100) were tested.

The forecast error covariance matrix considers the uncertainty of both inputs and the modeling process. Specifically, for the uncertainty of input variables, multiplicative perturbations were applied to forcing precipitation and potential evapotranspiration based on the scheme used by Crow et al. (2024): $Prec_i = Prec^* f_{Prec}$ and $PET_i = PET^* f_{PET}$, where f_{Prec} and f_{PET} follow mean-unity, lognormal distributions with

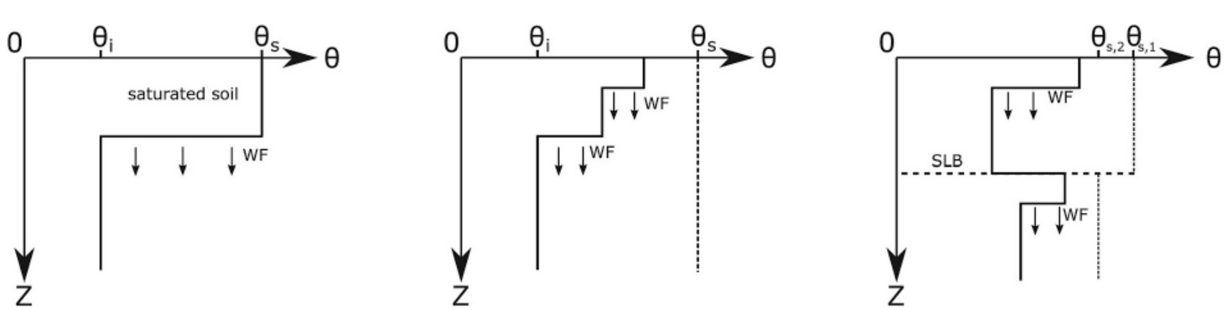


Fig. 3. Visual schematics for the Green and Ampt infiltration (GA) model (left), the Green and Ampt infiltration with Redistribution (GAR) model (center), and the LGAR model (right). "WF" and "SLB" indicate a wetting front and a soil layer boundary, respectively. Figure from: La Follette et al. (2023).

(unitless) standard deviations of 0.50 and 0.40, respectively. The covariance matrix of the modeling process error is: $w = diag\left(0.03^2, 0.02^2\right)$; That is, we assume that the standard deviations of modeling results caused by model structure are 0.03 m^3/m^3 and 0.02 m^3/m^3 for the surface layer and rootzone, respectively, and are mutually independent.

The covariance matrix of measurements (ML results) is:

$$R = diag\left(\sigma^2_{ML\_sly}, \sigma^2_{ML\_rz}\right) \tag{16}$$

In the equation, σ_{ML_sly} and σ_{ML_rz} are the standard deviation of the errors in ML estimates in the surface layer and rootzone SM at the time t_k, which are calculated from ML quantile estimates using Eq. 10.

The Kalman gain K_k is then calculated as:

$$K_k = \left[P(P+R)^{-1}\right]_{t=t_k} \tag{17}$$

At each measurement time t_k, when ML results are available, we update each ensemble member following:

$$X^i_+ = X^i + K_k\left(Z^i_k - X^i\right) \tag{18}$$

Here, X^i and X^i_+ are the state of the i-th (i = 1, 2, …, N_e) ensemble member at the time t_k before and after the update, respectively. Z^i_k is perturbed Z_k particular to each ensemble member ($Z^i_k = Z_k + v_k$), and v_k is a vector of Gaussian white noise with zero mean and standard deviations of σ_{ML_sly} and σ_{ML_rz}.

2.3.4. Model evaluation metrics

The metrics used to evaluate the performance of models include Pearson correlation coefficient (Corr), root mean squared error (RMSE), bias-corrected RMSE (i.e., ubRMSE), and the Kling-Gupta efficiency (KGE). RMSE and Corr are two commonly used metrics that capture the degree of mismatch between estimates and the true values of variables. Likewise, ubRMSE is a commonly used metric for SM estimation applications that reflects the common use of SM estimates for relative change and anomaly detection applications (Entekhabi et al., 2010b; Ma et al., 2024a, 2024b).

KGE is a comprehensive validation metric (Eq. 14) that combines the Corr, bias ratio (β), and variability ratio (γ) (Kling et al., 2012; Vergopolan et al., 2020).

$$KGE = 1 - \sqrt{(Corr - 1)^2 + (\beta - 1)^2 + (\gamma - 1)^2} \tag{19}$$

In this equation, β is the bias ratio, defined as the ratio of simulated and observed SM means, and γ is the variability ratio, defined as the ratio of simulated and observed SM variation coefficients.

3. Results

3.1. ML results

With the tuned hyperparameters (listed in Table S2), three ML algorithms (i.e., SVR, RF, and LGB) were evaluated, and the site-level statistics of evaluation metrics were summarized in Table S3. Based on the results, the LGB algorithm gives the best SM predictions both in the surface layer and rootzone with different high-resolution remote sensing data as inputs. Therefore, the LGB is selected as the ML algorithm in this SM-producing framework.

Table 3 illustrates the mean and median Corr, RMSE, ubRMSE, and KGE across all 431 crop sites for all ML models in the surface layer and rootzone built using the LGB algorithm. Within the surface layer, the median ubRMSE values for ML-1, ML-2, and ML-3 are 0.048 m^3/m^3, 0.049 m^3/m^3, and 0.047 m^3/m^3, respectively, which indicates that more than half of the sites achieve an ubRMSE below 0.049 m^3/m^3 for all three models thus highlighting their good predictive accuracy and reliable spatial transferability. The ML-2 model with the highest mean Corr (0.603) and KGE (0.320) gives the best performance among the three ML models in the surface layer. In the rootzone, from a site-level perspective, significant biases exist in the results of these three ML models with a site-level mean ubRMSE of 0.033 m^3/m^3, 0.033 m^3/m^3, and 0.034 m^3/m^3. Likewise, the three ML models achieve a median KGE of 0.334, 0.362, and 0.319, respectively, in the rootzone.

Density scatter plots of predictions versus observations (Fig. 4) further demonstrate the good predictive performance of the six ML models. Specifically, KGE values of the three ML models for the surface layer reach 0.614, 0.570, and 0.605, and those for the rootzone have values of 0.579, 0.569, and 0.545, respectively. Despite their overall good performance, a conditional bias is seen across all six models, characterized by the overestimation of SM at lower moisture levels and an underestimation at higher moisture levels. This pattern highlights a key limitation of the trained ML models for accurately capturing extreme SM values.

Fig. 5 illustrates the spatial distribution of ubRMSE and bias for SM estimates for all six ML models. The results reveal that all six ML models exhibit similar spatial patterns of bias across both layers. In the surface layer, sites with significant bias, both positive and negative, are predominantly located along the eastern coastal region and the lower Mississippi River basin. Similarly, in the rootzone, a concentration of sites with a negative bias, indicating SM underestimation, is observed in the lower Mississippi River basin. However, the ML estimates of the surface layer SM have a higher uncertainty (higher ubRMSE) than in the rootzone. Even in the surface layer, the three ML models exhibit different performance: ML-2 (Sentinel-2 with other features) performed best with smaller ubRMSE at most sites across the CONUS; ML-3 (HLSL30 with other features) has a significant uncertainty in the eastern areas of CONUS and much better in west regions.

3.2. EnKF results

To assess the performance of the EnKF method, SMAP L4 products (referred to as SMAP-L4 in Figures/Tables) were utilized as a benchmark for comparison. Fig. 6 presents the site-level statistics of metrics of the combined ML estimates (referred to as ML), original LGAR model, EnKF, and SMAP-L4 SM product on their overlapping dataset (to ensure a fair comparison, only instances where all four models and in-situ observations provided valid data are considered). For the EnKF method, an ensemble size of 50 is selected based on the performance (Table S4). The evaluation was conducted for both the surface layer and the rootzone.

These results show that, in the surface layer, the EnKF method achieves the highest KGE mean (0.411) and median (0.464) among all methods/products, indicating better overall balance of correlation, bias, and variability. Although SMAP-L4 has the highest correlation (mean = 0.680, median = 0.719), it exhibits a larger RMSE (0.081) and slightly lower KGE (0.355), suggesting more bias and variability. The ML estimates and original LGAR simulation show slightly lower performance than EnKF, with mean KGEs of 0.401 and 0.386, respectively. In the rootzone, EnKF again outperforms other methods/products in terms of KGE (mean = 0.316, median = 0.358), followed closely by ML (mean = 0.312) and LGAR (mean = 0.313). While SMAP-L4 has the highest correlation (mean = 0.686), it also has the largest RMSE (0.079), leading to a lower mean KGE (0.309) compared to EnKF. Overall, the EnKF consistently produced the most balanced and accurate SM estimates across both layers.

Fig. 7 displays comparisons of SM time series derived from the original LGAR simulation, ML predictions, and the EnKF analysis (including uncertainty bounds), alongside in-situ observations at three representative sites: Elsberry_PMC (SCAN), FSX (PSA), and Gadsden_19_N (USCRN). The three sites represent distinct soil textures: silt loam at Elsberry_PMC (3.78 % sand, 26.04 % clay), loam at FSX (35.30 % sand, 22.90 % clay), and sandy loam at Gadsden_19_N (52.96 % sand, 16.39 % clay), corresponding to different soil hydraulic properties. These comparisons are made at three temporal resolutions—weekly,

Table 3
Performance metrics for all six ML models within the surface layer (0–5 cm) and rootzone (0–100 cm) based on eight-fold cross-validation. The "mean" and "median" results summarize site-level statistics of these metrics. ML-1, ML-2, and ML-3 are ML models trained using different sets of variables (refer to Table 2 for more details), and the "S" and "R" notation following the numbers denote the surface layer and rootzone, respectively.

Models	Statistics	Metrics			
		Corr	RMSE (m^3/m^3)	ubRMSE (m^3/m^3)	KGE
ML-1S	mean	0.597	0.074	0.050	0.285
	median	0.680	0.067	0.048	0.412
ML-1R	mean	0.506	0.064	0.033	0.227
	median	0.583	0.055	0.030	0.334
ML-2S	mean	0.603	0.074	0.051	0.320
	median	0.676	0.069	0.049	0.432
ML-2R	mean	0.542	0.062	0.033	0.266
	median	0.607	0.054	0.029	0.362
ML-3S	mean	0.588	0.072	0.048	0.286
	median	0.697	0.065	0.047	0.418
ML-3R	mean	0.505	0.065	0.034	0.216
	median	0.597	0.057	0.029	0.319

daily, and hourly—for both the surface SM (SSM) and rootzone SM (RZSM) layers. Overall, the EnKF analysis exhibits consistently improved or comparable performance relative to both the standalone LGAR simulations and ML estimates across all sites and time scales. The benefits of data assimilation are particularly evident at the hourly scale, where the LGAR model provides continuous simulations but is periodically updated using ML-derived estimates through the EnKF. Although the ML estimates are available only every 2–5 days, their assimilation meaningfully corrects the LGAR state at those time steps, with the improvements propagating forward in time and enhancing the accuracy of subsequent predictions. This effect is especially notable in the rootzone layer at Elsberry_PMC, where EnKF achieves substantial performance gains over both LGAR and ML. At daily and weekly scales, the EnKF continues to perform well, and the smoothing effect of temporal aggregation helps to further reduce high-frequency variability and measurement noise. The EnKF maintains strong performance at these scales, capturing both short-term fluctuations and longer-term seasonal dynamics.

3.3. Comparison between ML, LGAR, EnKF, SMAP-L4, SMAP-1 K and SMAP-HB

Table 4 summarizes comparisons between the EnKF, ML models, LGAR model, SMAP-L4, SMAP-based 1-km ThySM (referred to as SMAP-1 K hereinafter), and SMAP-HydroBlocks (referred to as SMAP-HB hereinafter) products, and the site-level mean and median values of validation metrics and their spatiotemporal resolutions. Note that SMAP-1 K and SMAP-HB products only provide surface layer SM estimates.

Among the evaluated models and products, the ML estimates show the highest overall performance for surface SM, achieving a median correlation (Corr) of 0.698, a median ubRMSE of 0.056 m^3/m^3, and a median KGE of 0.518, at a spatial resolution of 100 m and a temporal frequency of every 2–5 days. The SMAP-HB product also performs well in the surface layer, with the highest median Corr of 0.701 and median KGE of 0.521, but with a lower mean KGE of 0.236 than ML estimates. The EnKF method ranks closely behind, achieving a median Corr of 0.604, a median ubRMSE of 0.053 m^3/m^3, and a median KGE of 0.379, while offering continuous hourly estimates at 100 m resolution. In contrast, SMAP-1 K products yield lower KGE and higher ubRMSE values for surface SM.

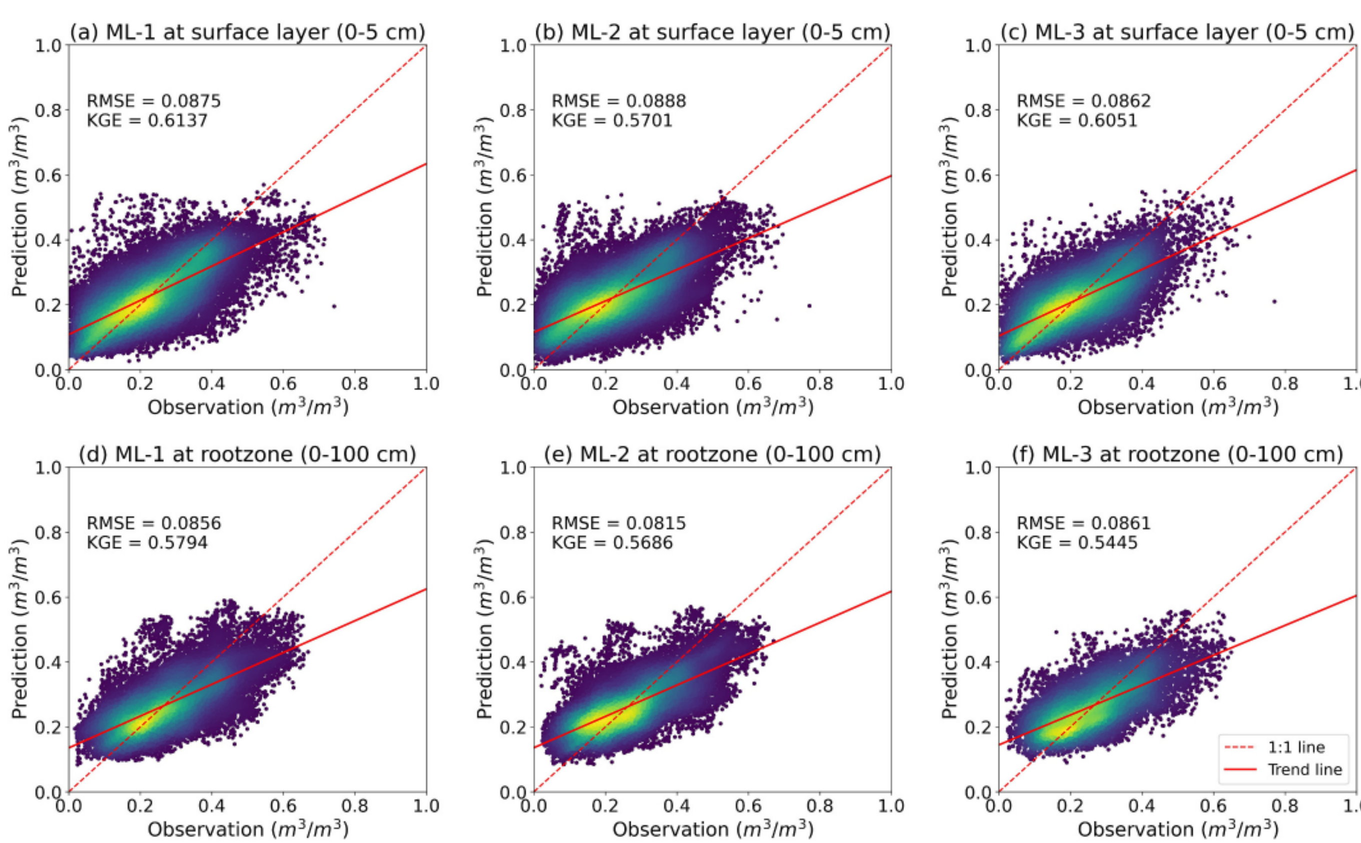


Fig. 4. Comparison between ML predicted and observed SM in the surface layer (0–5 cm) and rootzone (0–100 cm). ML-1 (a and d), ML-2 (b and e), and ML-3 (c and f) refer to ML models trained using different sets of variables (refer to Table 2 for more details), and the "S" (a, b and c) and "R" (d, e, and f) following the numbers denote the surface layer and rootzone, respectively.

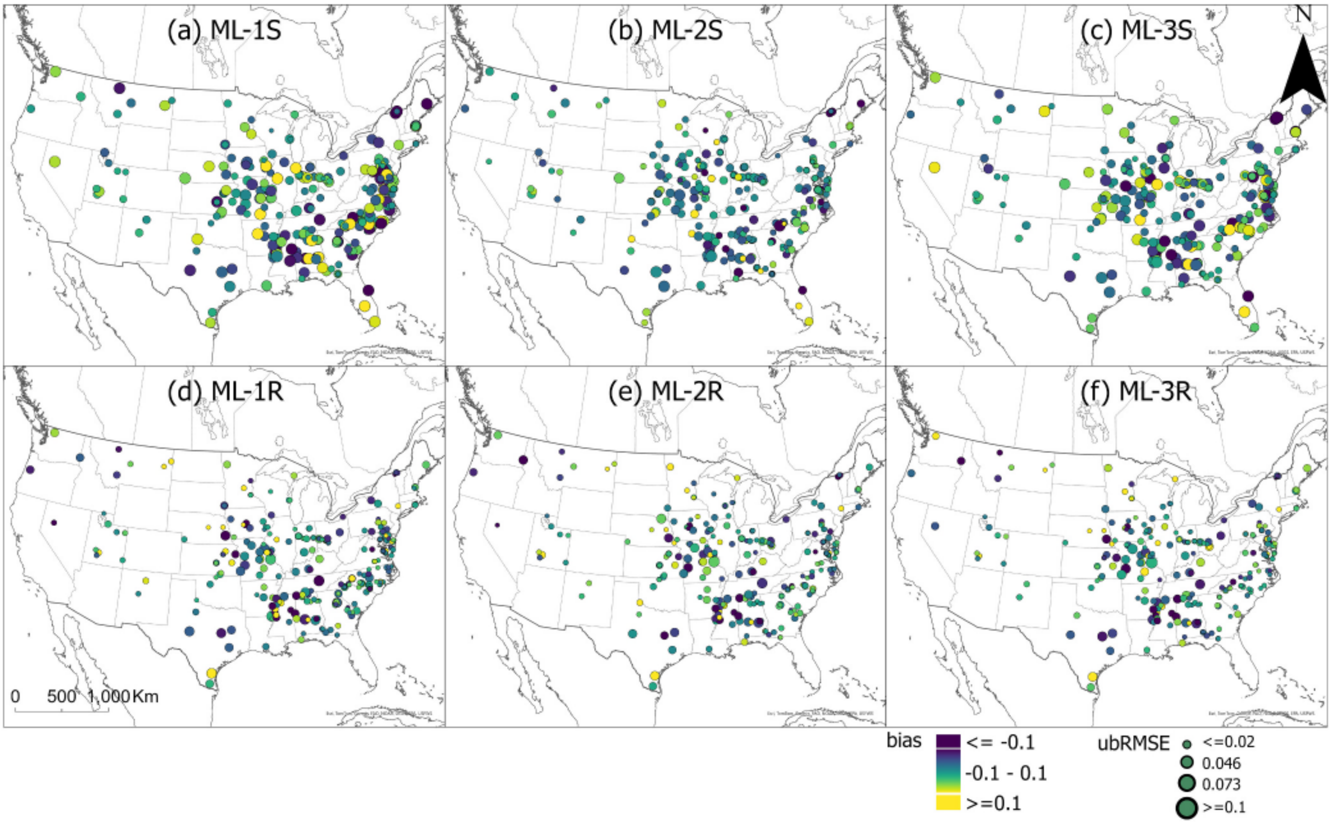


Fig. 5. The spatial distribution of the bias and ubRMSE of our six ML models in the surface layer and rootzone. ML-1 (a and d), ML-2 (b and e), and ML-3 (c and f) represent ML models trained with different sets of variables (refer to Table 2 for more details), with "S" (a, b and c) and "R" (d, e, and f) indicating the surface layer and rootzone, respectively.

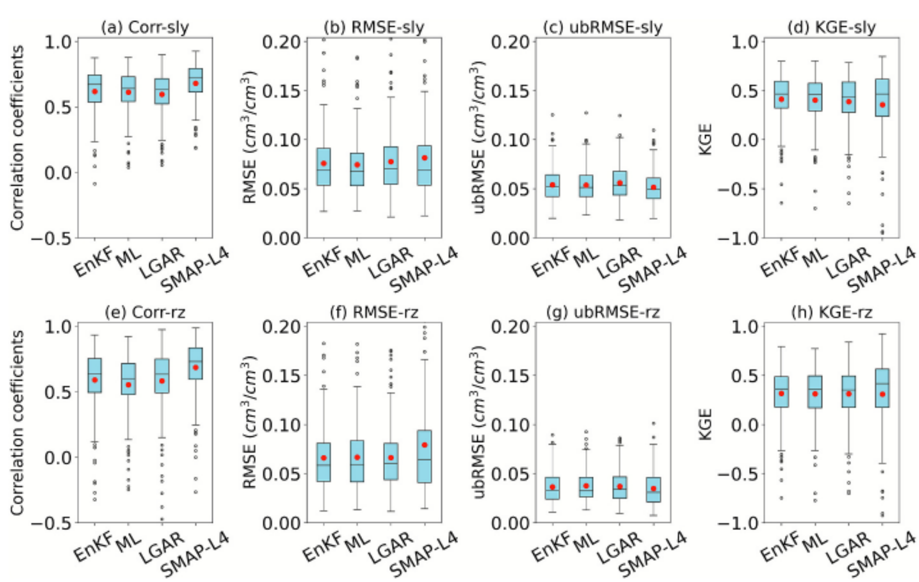


Fig. 6. Site-level boxplots summarizing validation metrics for the EnKF, ML models, LGAR model, and SMAP-L4 product on their overlapping dataset. The "sly" (a-d) and "rz" (e-h) denote the surface layer (0–5 cm) and rootzone (0–100 cm), respectively. Red points indicate the mean values of metrics. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In the rootzone, SMAP-L4 stands out with the highest performance, achieving a median Corr of 0.710, a median ubRMSE of 0.026 m^3/m^3, and a median KGE of 0.370, at a spatiotemporal resolution of 9 km every 3 h. The ML estimates follow, with a median Corr of 0.600 and a median KGE of 0.374, though they exhibit slightly higher ubRMSE (0.039 m^3/m^3). EnKF also performs well in the rootzone, with a median Corr of 0.595 and a median KGE of 0.302, surpassing the original LGAR model in all metrics and providing high-frequency (hourly) SM estimates at 100 m resolution. While ML slightly outperforms EnKF in terms of median KGE, EnKF's lower ubRMSE and higher mean KGE suggest more consistent improvements across sites. These results highlight the relative strengths of EnKF, which provides a valuable balance of spatial and

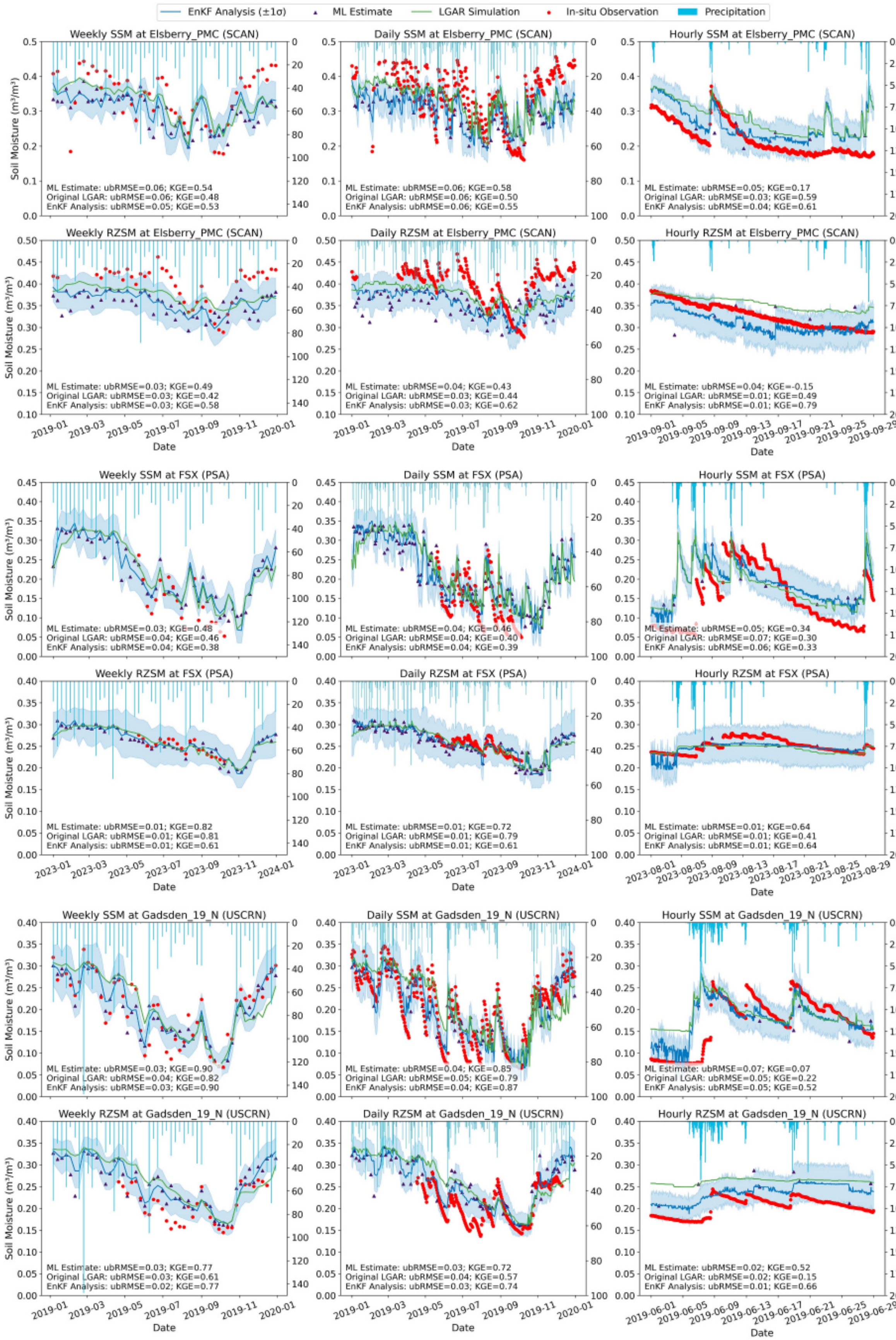


Fig. 7. Time series comparisons of EnKF analysis results, ML estimates, original LGAR simulations, and in-situ observations at three sites, evaluated across three temporal resolutions (weekly, daily, and hourly) for both the surface layer and rootzone, with precipitation displayed on the right inverted y-axis. Subplot titles specify the time scale, SM layer (SSM: surface SM; RZSM: rootzone SM), and site name (including the network). ubRMSE is the unbiased RMSE with a unit of m^3/m^3.

Table 4
Comparison between the EnKF, ML models, LGAR model, SMAP-L4, SMAP-1 K, and SMAP-HB products. The metrics for each model are calculated using its full dataset.

Soil layers	Models	Statistics	Metrics				Spatiotemporal resolution
			Corr	RMSE (m^3/m^3)	ubRMSE (m^3/m^3)	KGE	
Surface layer	EnKF	mean	0.559	0.077	0.055	0.298	Hourly/ 100 m
		median	0.604	0.069	0.053	0.379	
	ML	mean	0.647	0.080	0.058	0.419	2–5 days/ 100 m
		median	0.698	0.074	0.056	0.518	
	LGAR	mean	0.556	0.079	0.056	0.295	Hourly/ 100 m
		median	0.596	0.071	0.054	0.393	
	SMAP-L4	mean	0.639	0.088	0.051	0.240	3 h/ 9 km
		median	0.682	0.074	0.048	0.392	
	SMAP-1 K	mean	0.563	0.113	0.058	0.032	Daily/ 1 km
		median	0.630	0.106	0.054	0.227	
	SMAP-HB	mean	0.617	0.084	0.057	0.236	~2 days/ 30 m
		median	0.701	0.080	0.051	0.521	
rootzone	EnKF	mean	0.523	0.060	0.031	0.232	Hourly/ 100 m
		median	0.595	0.054	0.027	0.302	
	ML	mean	0.561	0.076	0.043	0.307	2–5 days/ 100 m
		median	0.600	0.066	0.039	0.374	
	LGAR	mean	0.523	0.062	0.033	0.213	Hourly/ 100 m
		median	0.597	0.056	0.028	0.277	
	SMAP-L4	mean	0.649	0.079	0.030	0.257	3 h/ 9 km
		median	0.710	0.059	0.026	0.370	

temporal resolution with competitive accuracy, particularly where data assimilation enhances the dynamic representation of SM processes.

The site-level statistics are further summarized in Fig. 8 (surface layer) and Fig. 9 (rootzone) by dividing the sites into humid and non-humid sites, and therefore providing a more interpretable perspective. All sites are categorized into humid and non-humid sites based on the aridity index (AI), a metric calculated as the ratio of mean annual precipitation to mean annual potential evapotranspiration (Zomer et al., 2022). The non-humid sites have an AI $\leq$0.65, encompassing dry sub-humid, semi-arid, arid, and hyper-arid regions, and humid sites have an AI >0.65. This division allowed for a more nuanced evaluation of model performance across different climatic conditions.

In the surface layer (Fig. 8), the ML estimates show consistent performance across the two climatic conditions, with no significant differences in results. In contrast, all other models perform better in non-humid areas (AI $\leq$0.65) compared to humid areas (AI >0.65). In non-humid regions, the SMAP-HB product demonstrates the best performance, achieving the highest mean and median KGE values of 0.464 and 0.620, and the highest mean and median Corr values of 0.686 and 0.729, respectively. This is followed by SMAP L4, ML, EnKF, SMAP-1 K, and the LGAR model. In humid regions, the ML models outperform others, with mean and median Corr values of 0.651 and 0.703, and mean and median KGE values of 0.417 and 0.524, respectively. The SMAP-HB product ranks second, with mean and median KGE values of 0.201 and 0.489. The EnKF model also performed well, with a mean KGE of 0.287, a median KGE of 0.375, and mean and median Corr values of 0.561 and 0.606, outperforming SMAP L4 (mean KGE = 0.208, median KGE = 0.364) and SMAP-1 K (mean KGE = 0.013, median KGE = 0.205).

Fig. 9 presents the performance of the ML models, EnKF method, LGAR model, and SMAP L4 product across two climatic conditions within the rootzone (noting that the SMAP-HB and SMAP-1 K products only provide surface layer SM). In the rootzone, all four models/products generally perform better in humid areas (AI >0.65) compared to non-humid areas (AI $\leq$0.65), although the performance differences for SMAP L4 between these two climate zones are less pronounced, with a higher mean KGE observed in non-humid areas. In humid regions, the ML models (mean KGE = 0.355, median KGE = 0.396) perform similarly to SMAP L4 (mean KGE = 0.263, median KGE = 0.394), followed by the EnKF method and the LGAR model. However, in non-humid areas, the SMAP L4 product significantly outperforms the other models/products. These results underscore the varying performance of the models under different climatic conditions and highlight the greater reliability of the EnKF model in humid regions in the rootzone.

3.4. Spatial mapping

Fig. 10 displays SM maps derived from the EnKF analysis and SMAP-HB over a cropland area in Grant County, USA, during the period from August 8 to August 22, 2019. This area includes both circular fields under center-pivot irrigation (e.g., croplands distributed in center-bottom and upper-left corners, highlighted using red frames in Fig. 10 subplot i) and rectangular fields that may be either irrigated or rainfed.

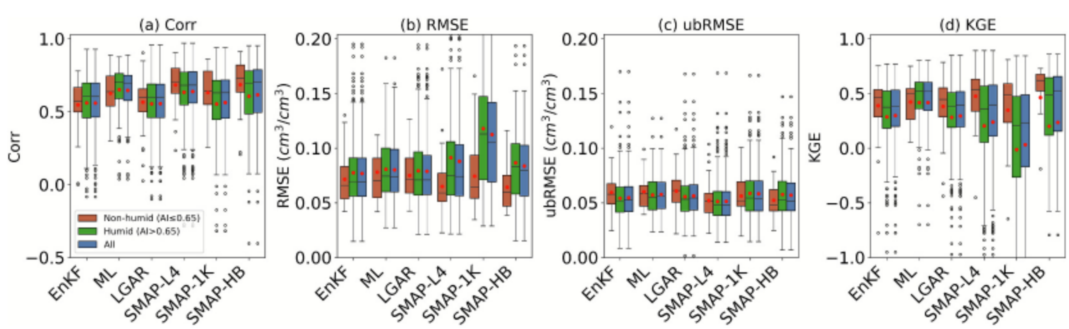


Fig. 8. Site-level performance metrics for the EnKF, ML models, LGAR model, SMAP-L4, SMAP-1 K, and SMAP-HB products in the surface layer (0–5 cm): (a) Corr; (b) RMSE; (c) ubRMSE; and (d) KGE. The sites are divided into two groups based on the AI (humid areas with AI >0.65 and non-humid areas with AI $\leq$0.65). Red points indicate the mean values of the metrics. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

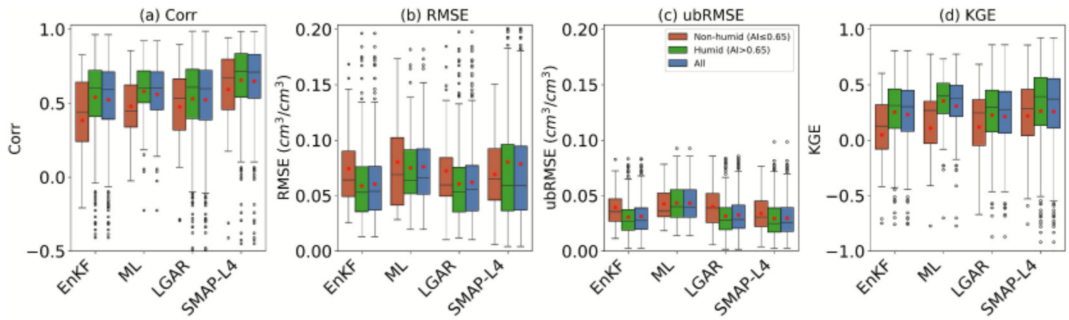


Fig. 9. Site-level performance metrics for the EnKF, ML models, LGAR model, and SMAP-L4 products within the rootzone (0–100 cm): (a) Corr; (b) RMSE; (c) ubRMSE; and (d) KGE. The sites are divided into two groups based on the AI (humid areas with AI >0.65 and non-humid areas with AI ≤0.65). Red points are the mean values of metrics. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Two precipitation events occurred during this period (August 10–12 and August 22), providing a basis for evaluating temporal SM responses. It is noteworthy that it had been a while before the first precipitation event.

Both EnKF and SMAP-HB exhibit strong spatial resolution and capture the general spatiotemporal dynamics of SM. Notably, SMAP-HB shows a marked response to the first rainfall event (August 10–12) but appears less responsive than EnKF to the second event (August 22). In most drier regions, EnKF displays similar spatial patterns to SMAP-HB, such as the persistently dry zone highlighted using blue boxes in maps (a) and (i). Despite the high (30 m) resolution of SMAP-HB, its pixels are aggregated into hydraulic response units defined by topographic, soil, land cover properties, etc., which means the pixels in the same unit have the same SM value. This aggregation can obscure fine-scale heterogeneity, particularly from human influences.

In contrast, the EnKF results exhibit finer spatial differentiation. For instance, areas identified with higher SM values (highlighted by red frames in map (i)) correspond to known irrigated croplands, underscoring the ability of this SM assimilation framework to capture human-driven heterogeneity in SM distribution. In the rootzone, spatial variation remains evident, although temporal fluctuations are less pronounced due to the dampened response of deeper soil layers to short-term rainfall events.

4. Discussion

4.1. Performance of ML models

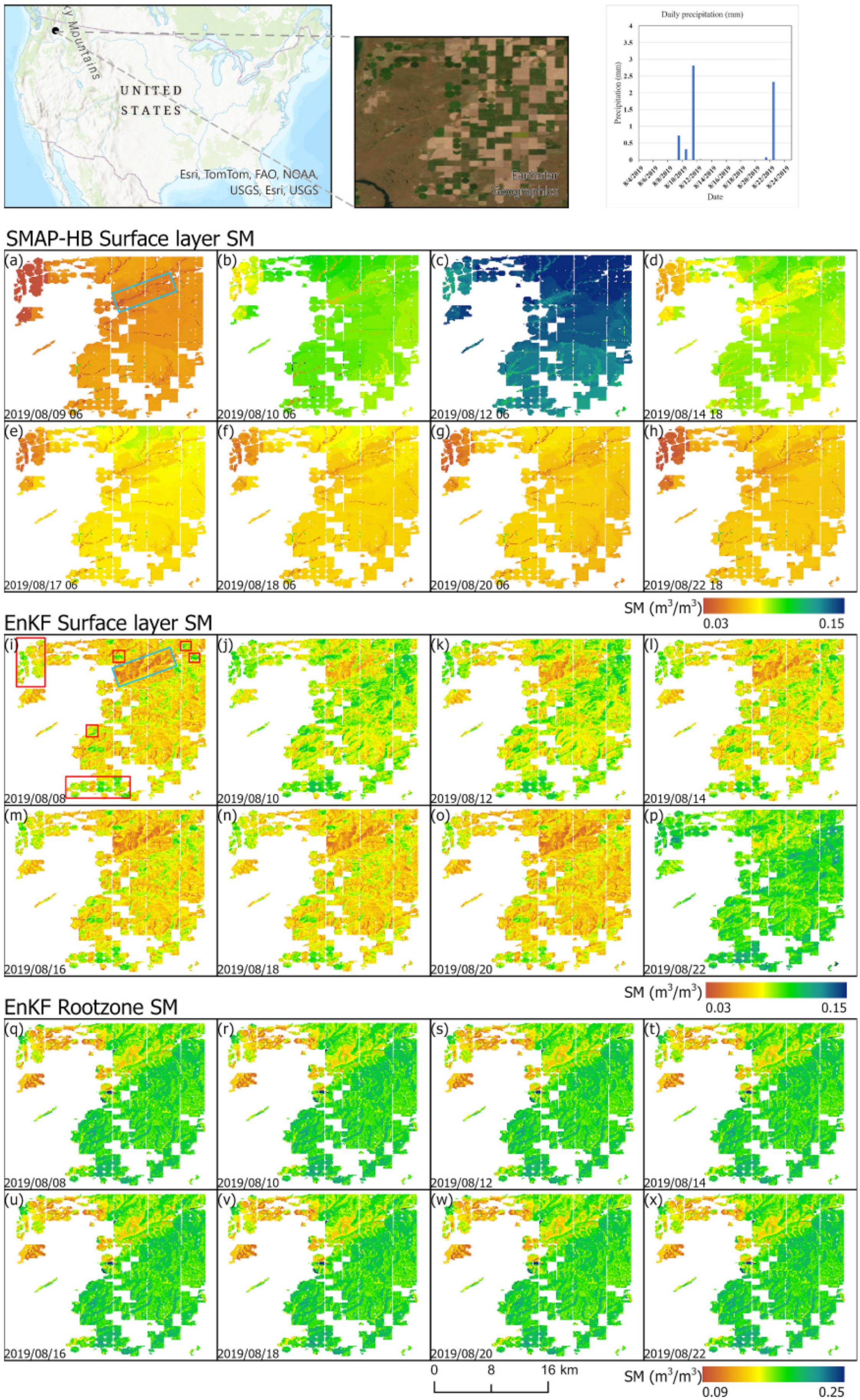
Because of their non-parametric nature and ability to capture complex and non-linear relationships (Padarian et al., 2020; Rani et al., 2022), Many ML algorithms have been widely used in SM modeling. In this study, three ML algorithms, including SVR, RF, and LGB, were tested. However, model training using limited input data often exhibits low spatial transferability due to either the different distribution of inputs or changes in output space (e.g., the trained data does not capture the entire range of SM) (Liu et al., 2021; Ma et al., 2024a, 2024b). To enhance the spatial transferability of ML models, the *Optimize* function from the *optuna* package in Python was used to optimize hyperparameters, and training data were manually partitioned into five spatially distinct groups for cross-validation thus ensuring robust parameter selection (the searching space and final selected parameters are shown in Tables S.1 and S.2). A spatially divided Eight-fold cross-validation was also performed to evaluate the performance of ML models (detailed results, including site-level distribution of multiple metrics, are shown in the Section 3.1).

In summary, the LGB ML algorithm gives the best performance among the three ML algorithms, and final training ML models demonstrate reliable performance, with mean and median KGEs of 0.419 and 0.518 in the surface layer, and 0.307 and 0.374 in the rootzone, as shown in Table 4. Nevertheless, the models exhibit certain limitations: they tend to underestimate high SM levels and overestimate low levels, particularly in the rootzone, indicating a reduced ability to capture extreme SM conditions accurately. This could be attributed to the limited observations at extreme conditions.

4.2. Performance of the LGAR model and the EnKF method

Based on our results (Table 4, Figs. 8 and 9), the LGAR model demonstrated robust performance in SM estimation (Note that the original LGAR simulation has been CDF-matched to ML estimates). For the surface layer, the LGAR model achieved a median Corr of 0.596 and a median ubRMSE of 0.054 m^3/m^3. For the rootzone, it attained a median Corr of 0.597, and a median ubRMSE of 0.028 m^3/m^3. These metrics underscore the reliability of the LGAR model in SM mapping. The original study noted that the LGAR model is only valid in regions where precipitation is less than potential evapotranspiration (i.e., AI >1) and where the groundwater table is deep due to the model's lack of groundwater linkage (La Follette et al., 2023). Interestingly, the results of this study show that in humid areas (AI >0.65), the LGAR model performed even better, with a median Corr and ubRMSE of 0.595 and 0.053 m^3/m^3 in the surface layer, and a median Corr and ubRMSE of 0.607 and 0.028 m^3/m^3 in the rootzone, compared to non-humid areas (the median Corr and ubRMSE are 0.597 and 0.061 m^3/m^3 in the surface layer, and the median Corr and ubRMSE are 0.532 and 0.039 m^3/m^3 in the rootzone). This suggests that the LGAR model has a broader application scope than previously acknowledged and is effective for SM mapping across most croplands (note that most sites in this study are characterized by AI<1), including humid regions.

The EnKF method improves the performance of the LGAR model by assimilating ML-derived estimates at specific time steps, thereby correcting the model states and enhancing the accuracy of subsequent forecasts through forward propagation. However, as with other land data assimilation approaches, EnKF is primarily designed to correct random errors and has limited capability in addressing systematic model errors (Crow et al., 2024; Kumar et al., 2015). As illustrated in Section 3.2 and Fig. 6, while EnKF outperforms the standalone LGAR model, it does not consistently exceed the performance of ML estimates—particularly in the rootzone layer. This limitation may stem from suboptimal assumptions regarding the input uncertainty and structural errors of the LGAR model. Moreover, parameter uncertainties within LGAR are not accounted for during the EnKF procedure, which may further contribute to residual systematic and random errors (Zhou et al., 2025). In this study, a final CDF-matching step was applied to the EnKF outputs using the ML estimates as a reference, which may help mitigate systematic biases caused by the process model but cannot eliminate random errors. Therefore, future research should focus on conducting additional experiments to identify optimal configurations for both the LGAR model and the EnKF method, thereby improving their integration and overall performance. Additionally, the temporal resolution of ML estimates,



(caption on next page)

Fig. 10. Spatial distribution of SM over a cropland area in Grant County, CONUS. Panels (a)–(h) show the SMAP-HB surface SM product; panels (i)–(p) present the surface SM generated from the EnKF data assimilation; and panels (q)–(x) display the rootzone SM from the EnKF. The red boxes in panel (i) highlight elevated SM values corresponding to circular fields typical of center pivot irrigation. The blue boxes in panels (a) and (i) indicate consistently dry areas observed in both the SMAP-HB product and the EnKF assimilation results. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

which is approximately every 2 to 5 days, is significantly less frequent compared to the hourly simulations generated by the LGAR model. Fig. 6 and Table 4 present the validation results of the EnKF approach at both the data assimilation steps (i.e., time steps overlapping with ML estimates) and across the entire simulation period. The findings reveal a notable decline in model performance when transitioning from assimilation steps to the full dataset, with mean and median KGE values dropping from 0.411 and 0.464 to 0.298 and 0.379 in the surface layer, and from 0.316 and 0.358 to 0.232 and 0.302 in the rootzone. These results suggest that increasing the temporal frequency of ML estimates could substantially enhance overall SM modeling performance. In the future, more high-resolution remote sensing data can be incorporated into this SM-producing framework, providing more high-resolution dynamic information.

4.3. Implications for agricultural applications

The SM generated by this framework achieves a high spatiotemporal resolution of 100 m and one hour. In the surface layer, the produced SM exhibits a median Corr of 0.604, a median RMSE of 0.069 m^3/m^3, and a median KGE of 0.379. In the rootzone, the median Corr is 0.595, the median RMSE is 0.027 m^3/m^3, and the median KGE is 0.302. This combination of resolution and accuracy demonstrates the product's potential to meet key requirements for most agricultural applications.

However, it is important to note that the performance of this SM retrieval framework varies between humid and non-humid areas (Figs. 8 and 9). For the EnKF-derived SM in the surface layer, the quality is superior in non-humid areas (AI $\leq$0.65), with a median Corr of 0.569, a median ubRMSE of 0.058 m^3/m^3, and a median KGE of 0.461, compared to humid areas (AI >0.65). Conversely, in the rootzone, the EnKF-derived SM shows better quality in humid areas, with a median Corr of 0.603, a median ubRMSE of 0.026 m^3/m^3, and a mean KGE of 0.313, outperforming its results in non-humid areas. These findings highlight the framework's sensitivity to climatic regimes, with performance varying significantly between humid and non-humid regions.

In agricultural applications, besides high resolution and spatiotemporal continuity, SM should also be characterized by a high potential to capture human-induced modifications such as irrigation. Among many existing SM products, the SMAP SM has been proven to have the ability to capture large-scale irrigation signals over semiarid regions in the western United States (Lawston et al., 2017), but it still cannot distinguish irrigated and rainfed croplands at the field scale due to its coarse footprint. Many SM products developed from SMAP have similar drawbacks, even though they have finer spatiotemporal resolutions, because there are no direct high-resolution observations as inputs in the SM retrieval algorithms. For example, SMAP-HB has a relatively high spatiotemporal resolution (30 m/ ~2 days) but may not be able to capture the impact of local-scale patchy irrigation because it lacks dynamic high-resolution data as input (Vergopolan et al., 2020, 2021). By integrating multiple high-resolution remote sensing datasets, the SM estimates generated by this framework have the potential to detect human-induced changes, including irrigation. Preliminary results (Fig. 10) indicate that the SM maps generated by this framework effectively capture both spatial and temporal heterogeneity and dynamics. Notably, the maps display elevated SM levels over circular croplands associated with center-pivot irrigation systems, demonstrating the framework's capability to detect human-induced spatial variability in agricultural fields at the field scale. Further research is needed to assess its capability to capture irrigation events using detailed irrigation information, which will be addressed in future work by incorporating an irrigation schedule dataset and/or using airborne-derived SM maps (Sabaghy et al., 2020).

5. Conclusions

This study proposed a novel SM estimation framework that integrates ML estimates into a process-based (LGAR) model using the EnKF data assimilation algorithm. By combining the strengths of data-driven and physically-based approaches, the framework effectively generates high-resolution (100 m, hourly) and spatiotemporally continuous SM across both the surface and rootzone layers over cropland in the CONUS.

Validation using in-situ observations from three networks, encompassing 431 cropland sites from 2016 to 2024, demonstrates that the framework achieves performance (median KGE = 0.379 in the surface layer and median KGE = 0.302 in the rootzone) comparable to SMAP L4 products (median KGE = 0.392 in the surface layer and median KGE = 0.370 in the rootzone). Notably, the framework enhances temporal consistency while retaining high spatial detail, which is a key advantage for agricultural and hydrological applications requiring field-scale monitoring.

The results also reveal spatial variability in performance across climate regimes: surface layer estimates are more accurate in non-humid regions, while rootzone estimates perform better in humid environments. These findings suggest that regional calibration or adaptive strategies may further enhance model performance.

Preliminary evaluations of the SM maps demonstrate the framework's ability to capture spatial heterogeneity and human-managed field-scale patterns, outperforming the SMAP-HydroBlocks product in representing fine-scale variability over cropland.

Overall, this study presents a scalable and adaptable framework for high-resolution SM retrieval by integrating remote sensing, ML, and process-based modeling. Future work will focus on optimizing the data assimilation process through targeted experiments aimed at refining the configuration of both the LGAR model and the EnKF algorithm. Additionally, the framework's ability to detect irrigation events could be further assessed using detailed irrigation records or drone-derived SM data.

CRediT authorship contribution statement

Shuohao Cai: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yijia Xu:** Writing – review & editing, Visualization, Methodology, Formal analysis, Data curation. **Zhengwei Yang:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition. **Wade T. Crow:** Writing – review & editing, Visualization, Validation, Methodology. **Zhou Zhang:** Writing – review & editing, Supervision, Methodology, Data curation. **Jiali Shang:** Writing – review & editing, Supervision, Methodology, Data curation. **Jiangui Liu:** Writing – review & editing, Supervision, Methodology, Data curation. **Peter La Follette:** Writing – review & editing, Methodology. **Chris Reberg-Horton:** Data curation. **Harry Schomberg:** Data curation. **Steven Mirsky:** Data curation. **Brian Davis:** Data curation. **Sarah Seehaver:** Data curation. **Alexis Correira:** Data curation. **Andrea Basche:** Data curation. **Ashley Waggoner:** Data curation. **Charles Ellis:** Data curation. **Dara Park:** Data curation. **Danielle D. Treadwell:** Data curation. **David Campbell:** Data curation. **Deann Presley:** Data curation. **Esleyther L. Henriquez Inoa:** Data curation.

Heather Darby: Data curation. **Jared Adam:** Data curation. **Jarrod Miller:** Data curation. **Joseph Haymaker:** Data curation. **John Wallace:** Data curation. **Julia Gaskin:** Data curation. **Kipling S. Balkcom:** Data curation. **Lindsey Ruhl:** Data curation. **Mark Reiter:** Data curation. **Matthew Ruark:** Data curation. **Michael Flessner:** Data curation. **Cynthia Sias:** Data curation. **Payton Davis:** Data curation. **Peter Tomlinson:** Data curation. **Richard G. Smith:** Data curation. **Nicholas D. Warren:** Data curation. **Ryan Dierking:** Data curation. **Shalamar Armstrong:** Data curation. **Tauana Almeida:** Data curation. **Jingyi Huang:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

None.

Acknowledgments

Support for this research was provided by the USDA Agriculture and Food Research Initiative Foundational Program (Grant No. 2023-67021-40007), the USDA Agriculture and Food Research Initiative Foundational Program Accession #1028199, and the University of Wisconsin–Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation. J. Huang was supported by the Marvin T. Beatty Professorship in Soil Science from the University of Wisconsin-Madison. We thank Aurelie Poncet from the University of Arkansas, Danielle Treadwell from the University of Florida, Samantha Wells from the University of Minnesota, and Stefan Gailans from the Practical Farmers of Iowa for their assistance in collecting the in-situ SM data from the PSA network. The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy. The United States Department of Agriculture is an equal opportunity employer.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2025.115167.

Data availability

All the data used in this study can be accessed freely online. The in-situ soil moisture data from the PSA network can be acquired upon request (visit their GitHub page: https://github.com/precision-sustainable-ag for more details). The code of this study can be accessed via this GitHub page: https://github.com/cshgiser/HRSM.

References

Al Bitar, A., Leroux, D., Kerr, Y.H., Merlin, O., Richaume, P., Sahoo, A., Wood, E.F., 2012. Evaluation of SMOS soil moisture products over continental US using the SCAN/SNOTEL network. IEEE Trans. Geosci. Remote Sens. 50, 1572–1586.

Arriaga, F.J., Lowery, B., Kelling, K.A., 2009. Surfactant impact on nitrogen utilization and leaching in potatoes. Am. J. Potato Res. 86, 383–390.

Babaeian, E., Sadeghi, M., Jones, S.B., Montzka, C., Vereecken, H., Tuller, M., 2019. Ground, proximal, and satellite remote sensing of soil moisture. Rev. Geophys. 57, 530–616.

Baig, M.H.A., Zhang, L., Shuai, T., Tong, Q., 2014. Derivation of a tasselled cap transformation based on Landsat 8 at-satellite reflectance. Remote Sens. Lett. 5, 423–431.

Bell, J.E., Palecki, M.A., Baker, C.B., Collins, W.G., Lawrimore, J.H., Leeper, R.D., Hall, M.E., Kochendorfer, J., Meyers, T.P., Wilson, T., et al., 2013. US Climate Reference Network soil moisture and temperature observations. J. Hydrometeorol. 14, 977–988.

Cai, S., Song, X., Hu, R., Leng, P., Li, X., Guo, D., Hao, Y., Wang, Y., et al., 2021. Spatiotemporal characteristics of agricultural droughts based on soil moisture data in Inner Mongolia from 1981 to 2019. J. Hydrol. 603, 127104.

Chan, S.K., Bindlish, R., O'Neill, P., Jackson, T., Njoku, E., Dunbar, S., Chaubell, J., Piepmeier, J., Yueh, S., Entekhabi, D., et al., 2018. Development and assessment of the SMAP enhanced passive soil moisture product. Remote Sens. Environ. 204, 931–941.

Chaney, N.W., Minasny, B., Herman, J.D., Nauman, T.W., Brungard, C.W., Morgan, C.L., McBratney, A.B., Wood, E.F., Yimam, Y., 2019. POLARIS soil properties: 30-m probabilistic maps of soil properties over the contiguous United States. Water Resour. Res. 55, 2916–2938.

Crow, W., Chen, F., Colliander, A., 2022. Benchmarking downscaled satellite-based soil moisture products using sparse, point-scale ground observations. Remote Sens. Environ. 283, 113300.

Crow, W.T., Kim, H., Kumar, S., 2024. Systematic modeling errors undermine the application of land data assimilation systems for hydrological and weather forecasting. J. Hydrometeorol. 25, 3–26.

Dewitz, J., 2021. National land cover database (NLCD) 2019 products. US Geol. Surv. 10, P9KZCM54.

Dong, J., Crow, W., Reichle, R., Liu, Q., Lei, F., Cosh, M.H., 2019. A global assessment of added value in the SMAP Level 4 soil moisture product relative to its baseline land surface model. Geophys. Res. Lett. 46, 6604–6613.

Dorigo, W., Xaver, A., Vreugdenhil, M., Gruber, A., Hegyiova, A., Sanchis-Dufau, A.D., Zamojski, D., Cordes, C., Wagner, W., Drusch, M., 2013. Global automated quality control of in situ soil moisture data from the International Soil Moisture Network. Vadose Zone J. 12, vzj2012-0097.

Entekhabi, D., Njoku, E.G., O'neill, P.E., Kellogg, K.H., Crow, W.T., Edelstein, W.N., Entin, J.K., Goodman, S.D., Jackson, T.J., Johnson, J., et al., 2010a. The soil moisture active passive (SMAP) mission. Proc. IEEE 98, 704–716.

Entekhabi, D., Reichle, R.H., Koster, R.D., Crow, W.T., 2010b. Performance metrics for soil moisture retrievals and application requirements. J. Hydrometeorol. 11, 832–840.

Hou, X., Orth, R., et al., 2020. Observational evidence of wildfire-promoting soil moisture anomalies. Sci. Rep. 10, 1–8.

Jury, W.A., Horton, R., 2004. Soil Physics. John Wiley & Sons.

Jyotiprava Dash, C., Sarangi, A., Singh, D., Singh, A., Adhikary, P.P., 2015. Prediction of root zone water and nitrogen balance in an irrigated rice field using a simulation model. Paddy Water Environ. 13, 281–290.

Karthikeyan, L., Chawla, I., Mishra, A.K., 2020. A review of remote sensing applications in agriculture for food security: crop growth and yield, irrigation, and crop losses. J. Hydrol. 586, 124905.

Kerr, Y.H., Waldteufel, P., Richaume, P., Wigneron, J.P., Ferrazzoli, P., Mahmoodi, A., Al Bitar, A., Cabot, F., Gruhier, C., Juglea, S.E., et al., 2012. The SMOS soil moisture retrieval algorithm. IEEE Trans. Geosci. Remote Sens. 50, 1384–1403.

Kling, H., Fuchs, M., Paulin, M., 2012. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. J. Hydrol. 424, 264–277.

Koster, R.D., Guo, Z., Yang, R., Dirmeyer, P.A., Mitchell, K., Puma, M.J., 2009. On the nature of soil moisture in land surface models. J. Climate 22, 4322–4335.

Krueger, E.S., Ochsner, T.E., Engle, D.M., Carlson, J., Twidwell, D., Fuhlendorf, S.D., 2015. Soil moisture affects growing-season wildfire size in the southern Great Plains. Soil Sci. Soc. Am. J. 79, 1567–1576.

Kumar, S., Peters-Lidard, C., Santanello, J., Reichle, R., Draper, C., Koster, R., Nearing, G., Jasinski, M., 2015. Evaluating the utility of satellite soil moisture retrievals over irrigated areas and the ability of land data assimilation methods to correct for unmodeled processes. Hydrol. Earth Syst. Sci. 19, 4463–4478.

La Follette, P., Ogden, F.L., Jan, A., 2023. Layered Green and Ampt infiltration with redistribution. Water Resour. Res. 59, e2022WR033742.

Lawston, P.M., Santanello Jr., J.A., Kumar, S.V., 2017. Irrigation signals detected from SMAP soil moisture retrievals. Geophys. Res. Lett. 44, 11–860.

Li, Z.-L., Leng, P., Zhou, C., Chen, K.-S., Zhou, F.-C., Shang, G.-F., 2021. Soil moisture retrieval from remote sensing measurements: current knowledge and directions for the future. Earth Sci. Rev. 218, 103673.

Liu, Y., Jing, W., Wang, Q., Xia, X., 2020. Generating high-resolution daily soil moisture by using spatial downscaling techniques: a comparison of six machine learning algorithms. Adv. Water Resour. 141, 103601.

Liu, J., Shen, Z., He, Y., Zhang, X., Xu, R., Yu, H., Cui, P., 2021. Towards Out-of-distribution Generalization: A Survey. arXiv preprint. arXiv:2108.13624.

Liu, P.-W., Bindlish, R., O'Neill, P., Fang, B., Lakshmi, V., Yang, Z., Cosh, M.H., Bongiovanni, T., Collins, C.H., Starks, P.J., et al., 2022. Thermal hydraulic disaggregation of SMAP soil moisture over the continental United States. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 15, 4072–4092.

Ma, Y., Chen, S., Ermon, S., Lobell, D.B., 2024a. Transfer learning in environmental remote sensing. Remote Sens. Environ. 301, 113924.

Ma, H., Zeng, J., Zhang, X., Peng, J., Li, X., Fu, P., Cosh, M.H., Letu, H., Wang, S., Chen, N., et al., 2024b. Surface soil moisture from combined active and passive microwave observations: integrating ASCAT and SMAP observations based on machine learning approaches. Remote Sens. Environ. 308, 114197.

Masek, J., Ju, J., Roger, J., Skakun, S., Vermote, E., Claverie, M., Dungan, J., Yin, Z., Freitag, B., Justice, C., 2021. HLS Operational Land Imager Surface Reflectance and TOA Brightness Daily Global 30 m v2. 0. NASA EOSDIS L. Process. DAAC.

Massari, C., Brocca, L., Moramarco, T., Tramblay, Y., Lescot, J.-F.D., 2014. Potential of soil moisture observations in flood modelling: estimating initial conditions and correcting rainfall. Adv. Water Resour. 74, 44–53.

Molero, B., Merlin, O., Malbéteau, Y., Al Bitar, A., Cabot, F., Stefan, V., Kerr, Y., Bacon, S., Cosh, M., Bindlish, R., et al., 2016. SMOS disaggregated soil moisture product at 1 km resolution: processor overview and first validation results. Remote Sens. Environ. 180, 361–376.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., et al., 2021. ERA5-Land: a

state-of-the-art global reanalysis dataset for land applications. Earth Syst. Sci. Data 13, 4349–4383.
Ozdogan, M., Rodell, M., Beaudoing, H.K., Toll, D.L., 2010. Simulating the effects of irrigation over the United States in a land surface model based on satellite-derived agricultural data. J. Hydrometeorol. 11, 171–184.
Padarian, J., Minasny, B., McBratney, A.B., 2020. Machine learning and soil sciences: a review aided by machine learning tools. Soil 6, 35–52.
Peng, Y., Yang, Z., Zhang, Z., Huang, J., 2024. A machine learning-based high-resolution soil moisture mapping and spatial–temporal analysis: the mlhrsm package. Agronomy 14, 421.
Pokhrel, Y.N., Hanasaki, N., Wada, Y., Kim, H., 2016. Recent progresses in incorporating human land–water management into global land surface models toward their integration into earth system models. Wiley Interdiscip. Rev.: Water 3, 548–574.
Rani, A., Kumar, N., Kumar, J., Sinha, N.K., 2022. Machine learning for soil moisture assessment. In: Deep Learning for Sustainable Agriculture. Elsevier, pp. 143–168.
Rebel, K., De Jeu, R., Ciais, P., Viovy, N., Piao, S., Kiely, G., Dolman, A., 2012. A global analysis of soil moisture derived from satellite observations and a land surface model. Hydrol. Earth Syst. Sci. 16, 833–847.
Reichle, R.H., Koster, R.D., 2005. Global assimilation of satellite surface soil moisture retrievals into the NASA Catchment land surface model. Geophys. Res. Lett. 32.
Reichle, R., De Lannoy, G., Koster, R., Crow, W., Kimball, J., Liu, Q., 2018. SMAP L4 Global 3-hourly 9 km EASE-grid Surface and Root Zone Soil Moisture Analysis Update, Version 4. NASA National Snow and Ice Data Center Distributed Active Archive Center.
Reichle, R.H., Liu, Q., Koster, R.D., Crow, W.T., De Lannoy, G.J., Kimball, J.S., Ardizzone, J.V., Bosch, D., Colliander, A., Cosh, M., et al., 2019. Version 4 of the SMAP level-4 soil moisture algorithm and data product. J. Adv. Model. Earth Syst. 11, 3106–3130.
Rodell, M., Houser, P., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., et al., 2004. The global land data assimilation system. Bull. Am. Meteorol. Soc. 85, 381–394.
Sabaghy, S., Walker, J.P., Renzullo, L.J., Akbar, R., Chan, S., Chaubell, J., Das, N., Dunbar, R.S., Entekhabi, D., Gevaert, A., et al., 2020. Comprehensive analysis of alternative downscaled soil moisture products. Remote Sens. Environ. 239, 111586.
Schaefer, G.L., Cosh, M.H., Jackson, T.J., 2007. The USDA natural resources conservation service soil climate analysis network (SCAN). J. Atmos. Oceanic Tech. 24, 2073–2077.
Seneviratne, S.I., Corti, T., Davin, E.L., Hirschi, M., Jaeger, E.B., Lehner, I., Orlowsky, B., Teuling, A.J., 2010. Investigating soil moisture–climate interactions in a changing climate: a review. Earth Sci. Rev. 99, 125–161.
Sheffield, J., Andreadis, K., Wood, E.F., Lettenmaier, D., 2009. Global and continental drought in the second half of the twentieth century: severity–area–duration analysis and temporal variability of large-scale events. J. Climate 22, 1962–1981.
Shi, T., Xu, H., 2019. Derivation of tasseled cap transformation coefficients for Sentinel-2 MSI at-sensor reflectance data. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 12, 4038–4048.
Sishah, S., Abrahem, T., Azene, G., Dessalew, A., Hundera, H., 2023. Downscaling and validating SMAP soil moisture using a machine learning algorithm over the Awash River basin, Ethiopia. PLoS One 18, e0279895.
Srivastava, P.K., Han, D., Ramirez, M.R., Islam, T., 2013. Machine learning techniques for downscaling SMOS satellite soil moisture using MODIS land surface temperature for hydrological application. Water Resour. Manag. 27, 3127–3144.
Teshome, F.T., Bayabil, H.K., Schaffer, B., Ampatzidis, Y., Hoogenboom, G., 2024. Improving soil moisture prediction with deep learning and machine learning models. Comput. Electron. Agric. 226, 109414.
Thapa, R., Tully, K.L., Reberg-Horton, C., Cabrera, M., Davis, B.W., Fleisher, D., Gaskin, J., Hitchcock, R., Poncet, A., Schomberg, H.H., et al., 2022. Cover crop residue decomposition in no-till cropping systems: insights from multi-state on-farm litter bag studies. Agr Ecosyst Environ 326, 107823.
Vergopolan, N., Chaney, N.W., Beck, H.E., Pan, M., Sheffield, J., Chan, S., Wood, E.F., 2020. Combining hyper-resolution land surface modeling with SMAP brightness temperatures to obtain 30-m soil moisture estimates. Remote Sens. Environ. 242, 111740.
Vergopolan, N., Chaney, N.W., Pan, M., Sheffield, J., Beck, H.E., Ferguson, C.R., Torres-Rojas, L., Sadri, S., Wood, E.F., 2021. SMAP-HydroBlocks, a 30-m satellite-based soil moisture dataset for the conterminous US. Sci. Data 8, 264.
Wang, L., Qu, J.J., 2009. Satellite remote sensing applications for surface soil moisture monitoring: a review. Front. Earth Sci. China 3, 237–247.
Wang, M., Wyatt, B.M., Ochsner, T.E., 2023. Accurate statistical seasonal streamflow forecasts developed by incorporating remote sensing soil moisture and terrestrial water storage anomaly information. J. Hydrol. 626, 130154.
Wigneron, J.-P., Jackson, T., O'neill, P., De Lannoy, G., de Rosnay, P., Walker, J., Ferrazzoli, P., Mironov, V., Bircher, S., Grant, J., et al., 2017. Modelling the passive microwave signature from land surfaces: a review of recent results and application to the L-band SMOS & SMAP soil moisture retrieval algorithms. Remote Sens. Environ. 192, 238–262.
Xu, M., Yao, N., Yang, H., Xu, J., Hu, A., de Goncalves, L.G.G., Liu, G., 2022. Downscaling SMAP soil moisture using a wide & deep learning method over the continental United States. J. Hydrol. 609, 127784.
Yu, T., Zhou, J., Ranjbar, S., Chen, J., Digman, M.F., Zhang, Z., 2024. Evaluation of the effect of Sentinel-1 SAR and Environmental Factors in Alfalfa yield and quality estimation. Agronomy 14, 859.
Zhang, D., Zhou, G., 2016. Estimation of soil moisture from optical and thermal remote sensing: a review. Sensors 16, 1308.
Zhou, J., Dong, J., Feng, H., Yang, K., Crow, W.T., Wu, Z., Tian, X., Tian, J., Ma, X., Jiang, Y., 2025. Can typical land surface model parameterizations support the expected soil moisture assimilation efficiency? Water Resour. Res. 61, e2024WR038702.
Zomer, R.J., Xu, J., Trabucco, A., 2022. Version 3 of the global aridity index and potential evapotranspiration database. Sci. Data 9, 409.