https://doi.org/10.5194/essd-2024-599
Preprint. Discussion started: 28 January 2025

A 1 km soil moisture data over eastern CONUS generated through assimilating SMAP data into the Noah-MP land surface model

Sheng-Lun Tai, Zhao Yang, Brian Gaudet, Koichi Sakaguchi, Larry Berg, Colleen Kaul, Yun Qian, Ye Liu, and Jerome Fast

Pacific Northwest National Laboratory, Richland, 99352, USA

Correspondence to: Sheng-Lun Tai (sheng-lun.tai@pnnl.gov)

Abstract. An improved fine-scale soil moisture (SM) dataset at 1-km grid spacing, covering much of the eastern continental U.S., was generated by assimilating 9-km SMAP SM data into the v4.0.1 Noah-MP land surface model. The assimilation, conducted using the Ensemble Kalman Filter algorithm within NASA's Land Information System, involved 12 ensemble members. The SM analysis for 2016 was fully validated against in-situ observations from four different networks and compared with four other existing datasets. Results indicate that this SM analysis surpasses other datasets in top-layer SM distribution, including a machine learning-based product, despite all SM estimates being less heterogeneous than observed. The analysis of anomalous errors suggests that large similarity in intrinsic errors is likely due to overlapping data sources among the selected SM datasets. By assessing the product using the ARM SGP data, we show that soil temperature and surface heat fluxes are concurrently simulated in good accuracy. A specific investigation into the 2016 southeastern U.S. drought response further indicates drier conditions and higher evapotranspiration estimates compared to GLEAMv4.1. Notably, large errors are associated with grids having clay soil textures, highlighting the need for refined model treatments for specific soil types to further improve SM estimates.

1 Introduction

Soil moisture (SM) is a critical component in the complex interactions between the land surface and the atmosphere, influencing a range of processes that are vital for weather and climate dynamics. More specifically, it plays a significant role in regulating surface energy fluxes by controlling the partitioning of incoming solar radiation into sensible and latent heat fluxes, thereby impacting atmospheric stability, boundary layer dynamics, and the initiation of convective systems (Dirmeyer et al. 2016; Ek and Holtslag 2004; Betts 2002; Taylor et al. 2011).

In addition to groundwater, precipitation falling onto ground surface contributes to the SM availability. Conversely, variations in SM heterogeneity can also influence the spatial and temporal distribution of precipitation through its effects on evapotranspiration rates and the atmospheric moisture and energy budgets (Katul et al. 2012; Hsu and Dirmeyer 2023). Hence, the feedback loop between SM and precipitation is crucial for understanding and predicting regional hydrological cycles, droughts, and flood events (Koster et al. 2004; Dirmeyer et al. 2016). Furthermore, SM conditions can impact weather extremes such as heatwaves by modulating the surface energy balance and the efficiency of heat exchange between the land surface and the atmosphere (Seneviratne et al. 2010). These interactions occur across various spatial and temporal scales, underscoring the need for accurately capturing the spatial and temporal variabilities of SM distribution.

A variety of sensors such as TDR (Time Domain Reflectometry), capacitance probes, and neutron probes have been used in in-situ (ground-based) SM measurements. These measurements provide direct assessments of SM content at specific locations with high temporal resolution and accuracy in the soil column and are most useful for validating remote sensing data and calibrating hydrological models (Robock et al. 2000; Rasheed et al. 2022). However, their relatively sparse distribution hinders their applicability in providing realistic local-to-regional SM variability in broader regions despite efforts made in expanding soil moisture observation networks (Diamond et al. 2013; Schaefer et al. 2007; Hawdon et al. 2014; Dorigo et al. 2021; McPherson et al. 2007; Wang et al. 2023). Conversely, remote sensing satellites such as the Soil Moisture Active Passive (SMAP) mission, Advanced Microwave Scanning Radiometer-EOS (AMSR-E), Soil Moisture and Ocean Salinity (SMOS), and Sentinel-1 (Entekhabi et al. 2010; Njoku et al. 2003; Kerr et al. 2001; Torres et al. 2012), provide nearly global coverage of soil moisture estimates measured by passive and active microwave sensors. Passive microwave sensors measure soil moisture based on microwave emissions from the Earth's surface, while active radar sensors use backscatter measurements to infer soil moisture levels (e.g., Kerr et al. 2001; Wagner et al. 2013). These satellite-based retrievals offer spatially extensive coverage and reasonable revisit times (1 – 3 days), contributing to large-scale hydrological and climate studies. Nevertheless, known uncertainties of satellite SM retrievals such as relatively coarse resolution (O (10 km)), limited accuracy (affected by vegetation, surface roughness, and temperature), shallow depth (only in depth of 0-5 cm is measured), and environmental interference by rain, cloud, and snow cover have posed challenges on their contributions to represent local-to-regional scale SM distribution (e.g., Colliander et al. 2017).

Land Surface Models (LSMs) can simulate soil moisture conditions for any region by representing the interactions among the atmosphere, vegetation, and the ground (Niu et al. 2011; Lawrence et al. 2019; Liang et al. 1994). The key processes such as precipitation, infiltration, lateral flow, evaporation, and plant transpiration, and groundwater table variations are parameterized in the LSMs. When precipitation occurs, water can infiltrate into the soil, accumulate, or run off, depending on soil characteristics and the rate of rainfall. Evaporation from the soil and transpiration from plants (collectively called evapotranspiration) reduce soil moisture, while infiltration and percolation move water downward through the soil profile. LSMs typically predict these processes to provide estimates of soil moisture at different depths over time. Various depths of soil layers can be configured to model the water movement between these layers in the soil column. A retrospective LSM simulation forced by observation-constrained surface atmospheric conditions (rainfall, temperature, wind, humidity, and radiation, etc.), land and soil properties (leaf area index (LAI), albedo, land cover, soil texture, and permeability, etc.) is commonly used to reproduce the soil conditions. Despite the advantages, state-of-the-art LSMs still contain uncertain, incomplete, and/or unresolved physical processes that may introduce biases into the simulated land surface properties.

As a way to mitigate such modelling issues, data assimilation (DA) techniques such as ensemble Kalman filter (EnKF), variational methods (e.g., 3DVar and 4DVar), and Bayesian approaches have been used to merge multiple sources of observational data (in-situ measurements and satellite retrievals) with LSM simulations to optimize soil moisture simulations through improving initial conditions and parameter estimates, enhancing the accuracy of soil moisture predictions and hydrological forecasts (e.g., Reichle et al. 2002; Crow and Wood 2003; Kumar et al. 2008; Chao et al. 2022; Martens et al. 2017). In any DA approach, the assimilation scheme must be coupled with an LSM. As such, the generated analysis consists of model states which are always physically balanced and can be directly used as the initial conditions of LSM. Some additional advantages of utilizing DA in generating high-resolution SM data include their flexibility in data resolution (output frequency, horizontal grid spacing, and vertical layers) and domain coverage, the possibility to incorporate any improvements in the coupled models and/or new observables, and the availability of the full suite of land surface properties relevant for studies of atmospheric boundary layer and hydraulic processes.

Earlier studies have explored the impact of SMAP soil moisture data assimilation soil moisture estimates, hydrological modelling, and drought monitoring across different regions of the globe. For example, studies have shown promising results by assimilating SMAP soil moisture data into the Noah-MP land surface model (e.g., Rouf et al. 2021; Ahmad et al. 2022). Research by Rouf et al. (2021) discussed how the spatial resolution of SMAP SM data (36-km versus 9-km) and the grid spacing of analysis (12.5 and 0.5 km) would impact SM estimation over Oklahoma using the framework of NASA's Land Information System (LIS). They showed the accuracy in SM analysis is enhanced when assimilating the 9-km SMAP data with 0.5 km LSM grid spacing. Likewise, Yin and Zhan (2020) showed positive influence of soil moisture data assimilation coupled with Noah-MP simulations in the continental U.S. (CONUS) and underscores the needs of fine-scale soil moisture data in achieving optimal result. Ahmad et al. (2022) further demonstrated the positive impact of SMAP DA on soil moisture estimate in South Asia along with its sensitivity to SMAP data bias correction settings. In Chakraborty et al. (2024), the improved soil moisture distribution over India was carried out by incorporating SMAP soil moisture into the Indian Land Data Assimilation System (ILDAS). Building upon these studies, we aim to improve local-to-regional soil moisture distributions over much of the east CONUS region by assimilating SMAP Level 3 (L3) 9-km soil moisture product into the Noah-MP LSM using a grid spacing of 1 km with optimized precipitation forcing. The goal of this study is to demonstrate the creation of a year-long soil moisture data set for the eastern U.S., which allows us to look at the importance of DA for improving soil moisture estimates on both seasonal and regional bases. We assess the performance of our data set over the full study domain, but also explore key regions in additional detail. Specifically, we evaluate the performance of our data set in a known “hotspot” of land-atmosphere coupling using the dense in-situ observations collected by the Oklahoma Mesonet and DOE’s Atmospheric Radiation Measurement (ARM) facility in Southern Great Plains (SGP), which were also used in the study of Rouf et al. (2021). Moreover, we examine our datasets characterization of the extreme drought conditions affecting the southeast U.S. over the fall and winter of 2016. Comparisons are made to alternative SM data sets, including datasets generated through machine learning approaches, to better understand the value of DA incorporated with an LSM.

The remaining parts of this manuscript are organized as follows: the analysis domain and period are described in Section 2. The methodologies and datasets employed in this study are detailed in Section 3. The results of the impact of SM data assimilation and the evaluations of the generated SM estimate along with the other existing SM datasets are discussed in Section 4. Lastly, the summary and discussion are provided in Section 5.

2 Analysis domain and period

Our study domain encompasses a wide swath of the central and eastern CONUS (Figure 1). The time period for the analysis covers the entire year of 2016 from January 1 through December 31, 2016. This analysis period was selected in order to complement land-atmosphere coupled simulations associated with the 2016 Holistic Interactions of Shallow Clouds, Aerosols, and Land-Ecosystems (HI-SCALE) field campaign. The locations of in-situ measurements from the networks of United Sates Climate Reference Network (USCRN), Soil Climate Analysis Network (SCAN), Oklahoma Mesonet (OKMet), and ARM SGP are overlaid on the map in Figure 1a. The soil texture and land cover maps are given in Figure 1b and 1c, respectively. Table 1 summarizes the grid numbers and their percentages over the study domain, for each classified type of soil texture and land cover. The top three soil types (besides water) are silt loam (24.02%), loam (18.88%), and sandy loam (15.7%), whereas grassland, cropland, cropland/natural vegetation mosaics are top three land cover types with percentages of 22.2, 19.64, and 10.2%.

https://doi.org/10.5194/essd-2024-599
Preprint. Discussion started: 28 January 2025

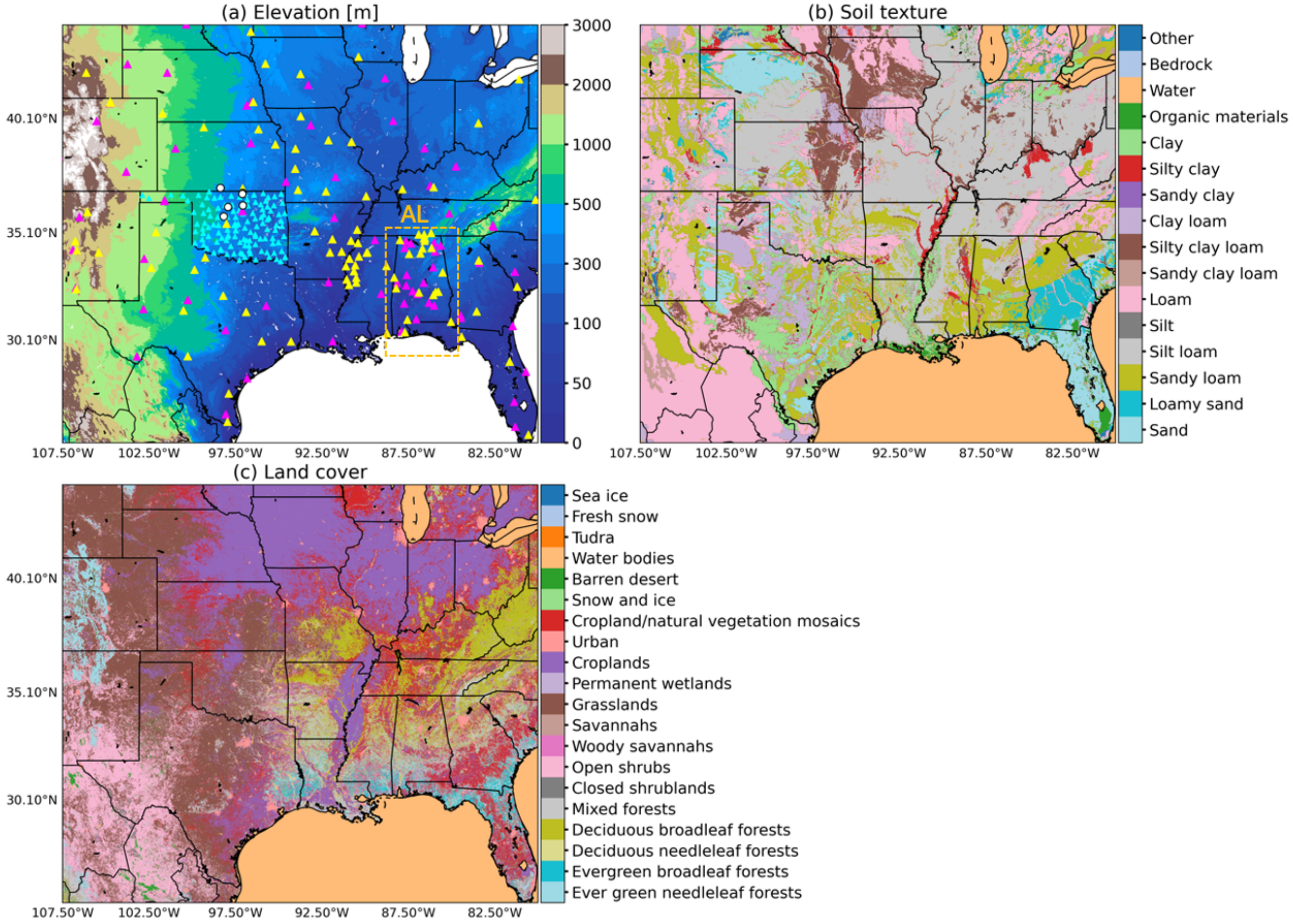

Figure 1 Maps illustrating the study domain over eastern CONUS. The yellow, magenta, and cyan triangles denote the stations of SCAN, USCRN, OKMet observational networks, respectively. The white circles mark the locations of selected ARM SGP sites. The domain soil texture was categorized into 14 soil types (**c**) according to the NCEP/STATSGO+FAO classification. The domain land cover comprised 18 main types based on the MODIS-derived IGBP classification. The subdomain AL is represented by the orange box (dashed line) in (a).

Table 1 Summary of grid number and percentage of total grids for the soil texture/land cover types.

Soil texture			Land cover		
Class	# of grids	Percentage of total grids [%]	Class	# of grids	Percentage of total grids [%]
Sand	370,227	7.12	Evergreen needleleaf forests	120,816	2.32

Loamy sand	140,072	2.69	Evergreen broadleaf forests	66,193	1.27
Sandy loam	816,849	15.70	Deciduous needleleaf forests	344	0.007
Silt loam	1249,730	24.02	Deciduous broadleaf forests	373,716	7.18
Loam	982,206	18.88	Mixed forests	332,114	6.38
Sandy clay loam	41,522	0.80	Closed shrublands	23,319	0.45
Silty clay loam	240,916	4.63	Open shrubs	514,667	9.89
Clay loam	187,450	3.60	Woody savannahs	100,390	1.93
Silty clay	63,818	1.23	Savannahs	27,746	0.53
Clay	205,676	3.95	Grasslands	1154,805	22.20
Organic materials	39,598	0.76	Permanent wetlands	9,591	0.18
Water	842,008	16.19	Cropland	1021,681	19.64
Other	22,069	0.42	Urban	64,997	1.25
			Cropland/Natural vegetation mosaics	521,029	10.02
			Snow and ice	33	0.0006
			Barren desert	28692	0.55
			Water bodies	842,008	16.19

3 Methodology and datasets

3.1 NASA Land Information System and Noah-MP land surface model

The NASA Land Information System (LIS) is an advanced modelling and data assimilation framework designed to better simulate land surface processes and improve our understanding of terrestrial hydrology, biogeochemistry, and climate interactions (Kumar et al. 2006; Peters-Lidard et al. 2007). LIS incorporates multiple hydrological and LSMs and data assimilation techniques to optimize the representation of land surface processes. This model-observation integration enhances the accuracy and reliability of simulations by leveraging the strengths of different models and observational datasets. It is functionable in assimilating satellite-derived observations of soil moisture, vegetation dynamics, and other land surface variables to improve the initialization and calibration of model simulations. Its versatility and scalability make it suitable for both research and operational uses. Given the above, LIS is primarily used in this study to generate realistic representation in soil states through assimilation of SMAP soil moisture retrievals into Noah-MP land surface model.

The version 4.0.1 Noah-MP LSM (Ek et al. 2003; Niu et al. 2011; Yang et al. 2011) was run within LIS to simulate the relevant land surface processes across the study domain. The Noah-MP model was run with a 0.01° by 0.01° horizontal grid spacing and using a 15-min time step. The specific model configurations utilized are detailed in Table 2. Each soil column within the study region is represented by four layers with depths of 10, 30, 60, and 100 cm below the ground surface. The surface soil moisture updates are transmitted to deeper layers according to model formulations in water diffusivity and hydraulic conductivity. More specifically, while moisture fluxes between successive layers controls how water moves within each soil column, excess water above saturation in any layer will be transferred to the next unsaturated layer downward. The Noah-MP LSM can be driven by many sources of meteorological forcing data as desired. Note that external irrigation and groundwater extraction were not explicitly simulated in Noah-MP and these processes might be important for certain locations (Yang et al. 2020, 2021).

Table 2 Selected parameters, parameterizations, and forcing data used in the configured Noah-MP LSM.

LSM parameter/parameterization/forcing data	
Land cover	MODIS (IGBP-NCEP) (Friedl et al. 2002)
Elevation, slope, and aspect	SRTM30-v2.0 (Farr et al. 2007)
Greenness	National Center for Environmental Prediction (Gutman and Ignatov 1998)
Vegetation	Dynamic vegetation option
Maximum albedo	National Center for Environmental Prediction (Robinson and Kukla 1985)
Canopy stomatal resistance	Ball-Berry method (Ball et al. 1987)
Snow surface albedo	Canadian land surface scheme (Verseghy 1991)
Runoff and groundwater	Simple groundwater model, SIMGM (Niu et al. 2007)
Surface-layer drag coefficient	General Monin-Obukov similarity theory (Brutsaert 1982)
Snow and soil temperature	Semi-implicit option
Partitioning of rain and snowfall	Jordan91(Jordan 1991)
Lower boundary of soil temperature	Noah native option
Supercooled liquid water and frozen soil permeability	NY06 (Niu et al. 2007)
Surface meteorological forcing	NLDAS-2 and Stage IV QPE (precipitation)

3.2 Datasets

The datasets employed in this study include the forcing data that drive the Noah-MP LSM (section 3.2.1 – 3.2.3), multiple in-situ observations (section 3.2.4) used as the benchmarks for intercomparison among our SM estimate and the other existing SM datasets (section 3.2.5 – 3.2.7).

3.2.1 Enhanced SMAP Level 3 soil moisture data

The Soil Moisture Active-Passive (SMAP) uses passive (radiometer) L-band microwave remote sensing to estimate land surface soil moisture and freeze/thaw state (O'Neill et al., 2014). The L-band radiometry offers all-weather, diurnal sensing of the surface dielectric properties which are a function of the near-surface soil moisture. The SMAP has a 2- to 3- day revisit frequency and two overpasses (morning and afternoon) at local time 6 a.m. and 6 p.m., respectively. One of the SMAP products, the enhanced SMAP Level 3 soil moisture product (SPL3SMP_E; O'Neill et al., 2020), is primarily used for assimilation in this study. It consists of daily estimates of global soil moisture within the top soil layer (~ 5 cm depth) on a cylindrical 9-km Equal-Area Scalable Earth Grid (https://nsidc.org/data/spl3smp_e/versions/6), spanning from 31 March 2015 to present.

3.2.2 North America Land Data Assimilation System Phase 2 (NLDAS-2)

The NLDAS-2 (Xia et al. 2012) aims to provide high-resolution, near-real-time and retrospective datasets that integrate land surface model outputs with observations to monitor and simulate land surface conditions across North America. It is available at hourly intervals and on a 12.5-km spatial grid from January 1979 to present. A wide range of land surface variables such as soil moisture, soil temperature, snow cover, evapotranspiration, and runoff are provided. Meteorological forcing variables such as precipitation, temperature, wind speed, and solar radiation are also included. The NLDAS-2 is used in this study as the meteorological forcing data to drive the Noah-MP LSM.

3.2.3 NCEP Stage IV Quantitative Precipitation Estimate

The NCEP Stage IV Quantitative Precipitation Estimate (QPE) (Lin and Mitchell (2005)) is a high-resolution, quality-controlled dataset produced by the National Centers for Environmental Prediction (NCEP). It integrates precipitation data from multiple sources, including NEXRAD radar, rain gauges, and satellite observations, to provide accurate and detailed precipitation estimates across the contiguous United States. With a grid spacing of 4 km at hourly intervals, Stage IV QPE is widely used in meteorology, hydrology, and climate research for tasks such as weather forecasting, flood modelling, and studying precipitation trends. We replace the precipitation data in the NLDAS-2 by the Stage IV QPE data as it provides a higher-resolution and more realistic precipitation forcing over the CONUS region.

3.2.4 In-situ measurements

In-situ soil moisture observations used in this study were obtained from the 1) U.S. Climate Reference Network (USCRN); 2) Soil Climate Analysis Network (SCAN); 3) Oklahoma Mesonet (OKMet, McPherson et al. 2007); 4) ARM SGP (Sisterson et al. 2016). The USCRN and SCAN data are acquired from the International Soil Moisture Network (Dorigo et al. 2021). The four networks are selected as the benchmarks of our SM analysis due to either their relatively wide spatial coverages or preferred site locations. Besides atmospheric and environmental parameters such as air temperature, humidity, and wind conditions, both SCAN and USCRN stations are equipped with sensors that measure critical soil parameters, including soil moisture and temperature at the depths of 5, 10, 20, 50, and 100 cm. The USCRN and SCAN are superior among available soil moisture networks as many of their stations (112 and 91 sites from USCRN and SCAN, respectively) are uniformly distributed over the study domain (Figure 1). They are used to evaluate our SM analysis along with other existing SM datasets (Table 3). The OKMet and ARM SGP observations are adopted as their site locations are densely distributed (average distance between any two stations is shorter than 30 km) over a portion of the Southern Great Plains (SGP) region which is one of the hotspots with strong land-atmosphere coupling (e.g., Fast et al. 2018; Sakaguchi et al. 2022). In addition to SM, the soil temperature observations and the latent and sensible heat fluxes measured by the Soil Temperature and Moisture Profiles (STAMP) and Eddy Correlation Flux Measurement System (ECOR) deployed by the ARM SGP facility, are also used to concurrently assess the simulated soil properties and surface heat fluxes. Note soil moisture (temperature) measured at a depth of 5 cm below ground surface was primarily used to compare with the model-estimated surface soil moisture (soil layer depth = 0 to 10 cm).

3.2.5 ERA5-Land reanalysis

The ERA5-Land (Muñoz-Sabater et al. 2021) is a global reanalysis dataset that provides essential land variables with a grid spacing of 0.1 degree and is valid at hourly frequency, spanning from January 1950 to present. It is continuously produced by rerunning the land component (Tiled ECMWF Scheme for Surface Exchanges over Land incorporating land surface hydrology (H-TESSEL)) of the ECMWF ERA5 climate reanalysis that sequentially assimilates available meteorological observations (Hersbach et al. 2020). Despite model uncertainties due in part to imperfect atmospheric forcing, unresolved physical processes, and lack of observational constraint, the spatiotemporal coverages of ERA5-Land dataset have been advantageous in many land surface applications including flood or drought monitoring and forecasting. It is thus employed in this study as one of the SM reference data, providing more insights though the comparison.

3.2.6 Global Land Surface Satellite soil moisture (GLASS SM)

The global, daily 1-km GLASS soil moisture product (GLASS SM; Zhang et al. 2023) was derived using an ensemble learning model (eXtreme Gradient Boosting – XGBoost) that integrates multiple datasets as the machine learning (ML) model's inputs, including the remotely sensed Global Land Surface Satellite (GLASS) products (Liang et al. 2021), ERA5-Land reanalysis products (Muñoz-Sabater et al. 2021), and static auxiliary datasets (e.g., Multi-Error-Removed Improved-Terrain (MERIT) and Global gridded soil information (SoilGrids; Poggio et al. 2021). The ground-based soil moisture archived by the International Soil Moisture Network (ISMN) and the 0.25 deg grid spacing combined soil moisture data of European Space Agency's Climate Change Initiative (ESA CCI; Dorigo et al. (2017)) are collectively used as the target data of training in ML. The validations carried out for the GLASS SM product in Zhang et al. (2023) demonstrated its capability in capturing temporal dynamics of measured soil moisture. Hence, given its novelty in the methodology and high spatial resolution (1km), the GLASS SM data is used as one of the benchmarks in this study.

3.2.7 Global Land Evaporation Amsterdam Model (GLEAM)

The GLEAM (Global Land Evaporation Amsterdam Model; Miralles et al., 2011) is a state-of-the-art dataset that provides global estimates of soil moisture, terrestrial evaporation (or evapotranspiration), and related hydrological components. GLEAM soil moisture data is derived from satellite observations and model simulations. It integrates a variety of satellite observations and meteorological data, such as soil moisture from microwave remote sensing, vegetation indices, and meteorological data of precipitation, air temperature, and radiation. The version 4.1 of GLEAM (Miralles et al., in review) is used in our analysis, which is available at 0.1-degree resolution between the period of 1980 to 2023.

Table 3 Soil moisture estimates analyzed in this study.

Soil moisture product	Grid spacing	Spatial coverage	Temporal resolution	Temporal coverage	References
SPL3SMP_E	9 km	Global	Daily	31 March 2015 - present	O'Neill et al. (2020)
ERA5-Land	0.1°	Global	Hourly	1950 - present	Muñoz-Sabater et al. (2021)
GLASS SM	1 km	Global	Daily	2000 - 2020	Zhang et al. (2023)
GLEAM v4.1	0.1°	Global	Daily	1980 - 2023	Miralles, D.G., Koppa, A., Baez-Villanueva, O.M., Tronquo, E., Bonte, O., Zhong, F., Beck, H.E., Hulsman, P., Haghdoost, S., Dorigo, W.A. GLEAM4: global evaporation and soil moisture datasets at 0.1° resolution from 1980 to near present, *in review*
SMAPDA	1 km	East CONUS	Hourly	2016	-

3.3 Data Assimilation (DA) simulation

Since the SMAP SM data is only available from March 31, 2015 onwards, the DA simulation started on 00 UTC of April 1, 2015 and ended on 00 UTC of January 1, 2017. The ensemble Kalman filter (EnKF) assimilation algorithm implemented in the LIS is utilized to assimilate the SMAP SM retrievals into the Noah-MP-modelled estimates. The EnKF's sequential assimilation algorithms including two main steps (model propagation and data assimilation update) are coupled with model integration and executed recursively. Here, the Noah-MP is the nonlinear forward model to advance the propagation step and generate the prognostic state vector forward in time. The update step occurs whenever any observations are valid, and the update of prognostic state variable can be described by the equation below:

$$\hat{x}_{k+1}^{a} = \hat{x}_{k+1}^{b} + K\left(y_{k+1} - H_{k+1}(\hat{x}_{k+1}^{b})\right). \quad (1)$$

Where $\hat{x}_{k+1}^{a}$ stands for analyzed (updated) state of variable x at time step $k+1$. $\hat{x}_{k+1}^{b}$ represents the background state of variable x integrated from time step k. The Kalman gain matrix K and the innovation vector $\left(y_{k+1} - H_{k+1}(\hat{x}_{k+1}^{b})\right)$ are required when updating the background state. Here, the y_{k+1} denotes the observations valid at time step $k+1$ and H_{k+1} is the observation operator that applies conversion and interpolation in time and space to the model state variable in order to conform with the observable.

The ensemble simulations are required at each propagation step to provide an estimate on the model spread (uncertainty). Here, the NASA Land Data Toolkit (LDT; Arsenault et al. (2018)) is used to initialize the ensemble simulations based on the open loop (OL) simulation restart output file at 2345 UTC on March 31, 2015. The open loop (OL) simulation starts from January 1, 2015 and refers to the integration of Noah-MP land surface model without any assimilation of external observations. The initial conditions of those ensemble members are obtained by perturbing atmospheric forcing variables as listed in Table 4. Perturbation type is grouped as either multiplicative (M), sampled from a log-normal distribution or additive (A) which is sampled from a normal distribution.

Table 4 Description of parameters used in meteorological forcing perturbations for the ensemble simulations

Perturbed meteorological forcing	Perturbation type	Standard deviation	Cross-correlations with perturbations			
			SW	LW	P	T_{air}
Shortwave radiation (SW)	M	0.2 W m^{-2}	1.0	-0.3	-0.5	0.3
Longwave radiation (LW)	A	30 W m^{-2}	-0.3	1.0	0.5	0.6
Precipitation (P)	M	0.5 mm	-0.5	0.5	1.0	-0.1
Near-surface air temperature (T_{air})	A	0.5 K	0.3	-0.6	-0.1	1.0

According to the sensitivity study regarding the impact of ensemble size in Ahmad et al. (2022), the ensemble spread (measured by standard deviation across all members) may be flattened when the number of replicates increases beyond 15.

We experimented with 12 and 24 ensemble members, and the result suggested minor difference is demonstrated in terms of soil moisture representation. Hence, the DA experiment we show in the study has an ensemble size of 12. The model and SMAP soil moisture retrieval error standard deviations are set as 0.04 $m^3\ m^{-3}$. Due to the existence of relative systematic difference between SMAP and modelled SM, the cumulative distribution function (CDF) matching technique (Reichle and Koster 2004) is used for bias correction of the SMAP soil moisture retrievals using Noah-MP model data as the reference. Monthly CDFs of the SMAP soil moisture retrievals and the Noah-MP-simulated soil moisture were both generated using the NASA LDT and used to map the SMAP SM retrievals into the Noah-MP-modelled soil moisture space prior to assimilation. Since the SMAP SM data is representative of the top soil layer (~5 cm deep from surface), the topmost soil layer soil moisture is employed as the model state variable during assimilation. The DA simulation as well as its SM data are abbreviated as "SMAPDA" hereafter. More detailed discussion regarding its performance in estimated SM is covered in the Section 4.

3.4 Metrics for DA impact measuring and evaluation

3.4.1. Soil moisture analysis increment

To assess the impact of the SMAP SM data assimilation on the soil moisture estimates, we analyze the soil moisture analysis increments generated from the DA experiment (SMAPDA). The analysis increment refers to the difference between the analysis (optimized estimate of the state after DA) and the background forecast (model state before DA). It is a measure of how much the model state has been corrected (updated) by incorporating new observations, which is not only related to the deviation from model background to observation, but also modulated by observation and model errors. In the EnKF approach, the model error varies in time and space and is estimated using the ensemble spread (standard deviation of ensemble simulations). We use the cumulative number and temporal mean of soil moisture analysis increments to indicate the spatial distribution of observational constraint by the SMAP_L3_E data and highlight the areas that experience an overall wetting or drying due to the cycling of assimilation. Note the SMAP_L3_E data was processed into hourly subset and assimilated when it matches the model time step.

3.4.2. Evaluation against in-situ measurements

The soil moisture estimates generated through different approaches are evaluated against in-situ measurements using the metrics of correlation coefficient (CC), root-mean-square error (RMSE), and Bias defined as follows:

$$\mathrm{CC} = \frac{\sum_{i=1}^{n}(P_i - \bar{P})(M_i - \bar{M})}{\sqrt{\sum_{i=1}^{n}(P_i - \bar{P})^2 \sum_{i=1}^{n}(M_i - \bar{M})^2}} \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(P_i - M_i)^2} \tag{3}$$

$$\text{Bias} = \frac{1}{N}\sum_{i=1}^{n}(P_i - M_i) \tag{4}$$

Where P represents the estimated top layer soil moisture, and M stands for in-situ soil moisture measurement. N is the total number of samples. The CC, ranging from -1 to 1, is essentially computed as the Pearson correlation coefficient using the modeled and in situ soil moisture time series in each location. It mainly quantifies the skill in capturing soil moisture temporal variations across all time scales. While RMSE is used to measure the mean difference between the modeled and in-situ SM, Bias is computed as the overall deviation (including the signs) of the modeled SM from in-situ SM observations. In addition, the standard deviation (STD) is also calculated for each SM dataset to quantify the spatial heterogeneity in SM across the given sites at different locations:

$$\text{STD} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(S_i - \bar{S})} \tag{5}$$

Here, S_i refers to individual SM data points and $\bar{S}$ stands for mean over the entire dataset.

4 Results

4.1 Impact of SMAP soil moisture data assimilation

To gauge how much observational information was effectively assimilated into the model, we examined the outputs of SM analysis increments at the top layer (5-cm depth). Figure 2 illustrates the maps of cumulative number (hours) of SM analysis increment over each of the three-month long periods. Overall, the SMAP SM data assimilation is more effective in spring, summer, and fall (Figures 2b, 2c, and 2d) as opposed to winter (Jan-Feb-Mar; Figures 2a). The relatively small number of analysis increment shown in the Jan-Feb-Mar period (Figure 2a) is likely due to the increased uncertainty in L-band microwave radiometer SM retrieval as a result of snow cover and frozen ground in the cold season (e.g., Liu et al. 2021). While analysis increments are distributed over the majority of domain, there are grids received zero update, especially in the eastern part of the domain. Nevertheless, despite generally less effective assimilation over this region, a few spots in Florida and partially Georgia and South Carolina show most frequent updates from DA across the entire domain.

Figure 3 demonstrates the spatial distribution of mean SM analysis increments over the four seasons. The calculation of mean increment only includes samples with non-zero increments. While consistently positive increments are shown in Texas and northern Mexico throughout the year, seasonal variations are evident in portions of the Great Plains. For instance, in Kansas, more negative (positive) increments are seen for Jan-Feb-Mar/Apr-May-Jun (Figures 3a and 3b) and Jul-Aug-Sep/Oct-Nov-Dec (Figures 3c and 3d), respectively. This suggests that compared to SMAP observations, the model most likely has a

290 consistent dry bias over part of Texas and the adjacent Mexican territory, and the biases are more variable temporally in other parts of the domain including the northern SGP.

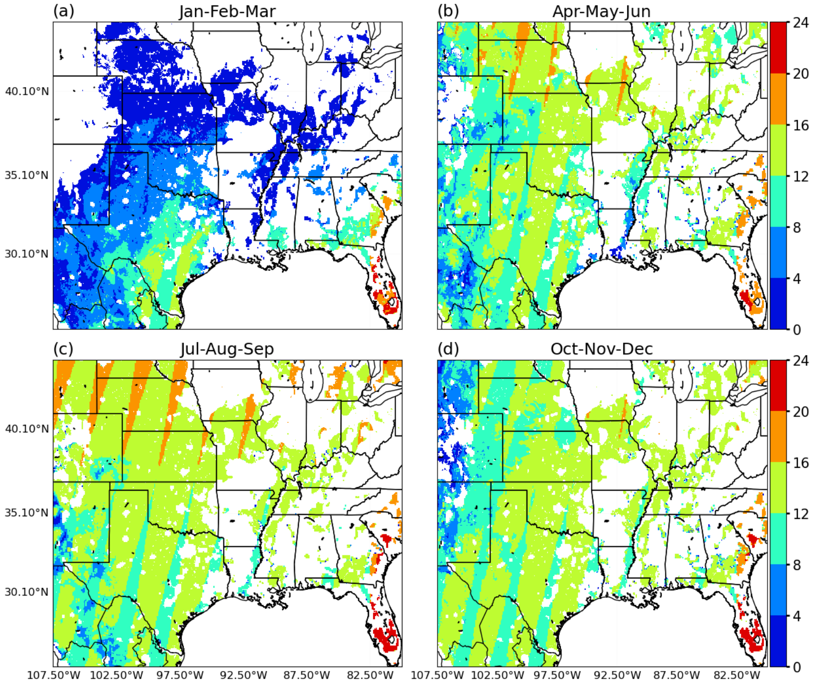

Figure 2 Maps of cumulative number of DA SM increments computed for the periods of (a) Jan-Feb-Mar, (b) Apr-May-Jun, (c) Jul-Aug-Sep, and (d) Oct-Nov-Dec in 2016.

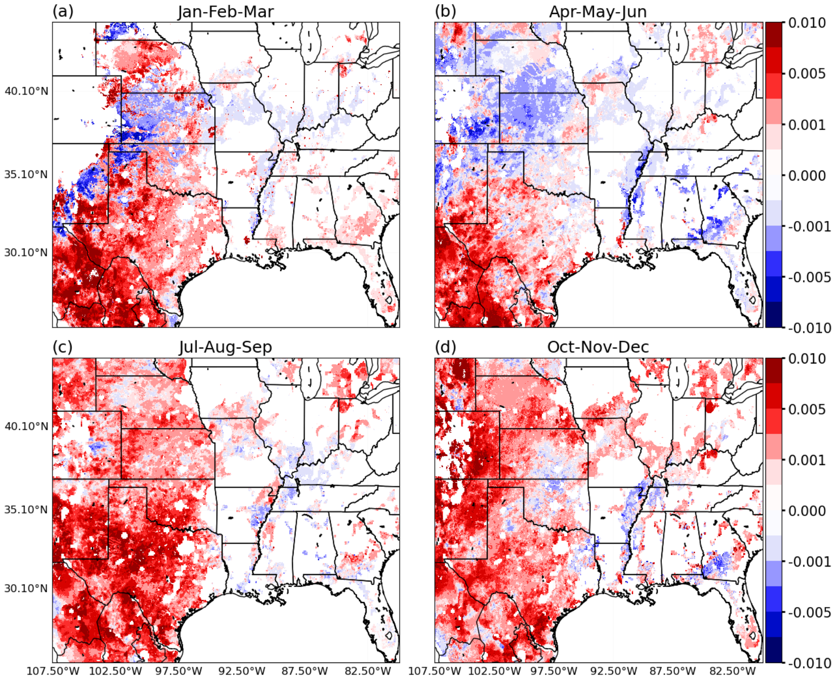

Figure 3 Similar to Figure 2, but for mean SM increments.

4.2 Comparison with existing surface SM datasets

To assess the performance of our SM analysis (SMAPDA) along with other existing SM products, we conduct a comprehensive intercomparison among all derived datasets (Table 3) against a collection of in-situ measurements from four observational networks (USCRN, SCAN, OKMet, and ARM SGP). Note the assessments were conducted separately against the USCRN and the SCAN datasets despite both have well-distributed site locations over the study domain. This was carried out purposely in order to verify whether any inconsistency between their instruments and/or measurements may alternatively bias the validation results. The following sub-sections discuss the evaluation results referenced by using the observations from each network.

4.2.1 Evaluation using USCRN soil moisture observations

SM estimates from SMAPDA, GLASS SM, ERA5-Land, GLEAMv4.1, and SMAP AM (the morning overpass of SMAP_L3_E) are first evaluated using the in-situ observations from the USCRN (Figure 1). The metrics described in Section 3.4.2 are computed accordingly. Since only SMAPDA and ERA5-Land consist of SM representations through the entire soil column, surface (top-layer) SM representations are primarily assessed here. To perform one-to-one comparisons with in-situ data, for each SM product, the daily SM timeseries data at the grid cells closest to the observational site locations are extracted.

The 2-D histograms as given in Figure 4 are illustrated to visualize the differences between the observation and the estimates and depict the contrasts among the datasets. All scatter points are grouped by 50 bins (2-D pixels) and the contours are smoothed using the gaussian filter for an improved visualization. The more samples concentrated along the diagonal line, the better estimate it would be considered.

The results indicate SMAPDA has the highest correlation coefficient (~ 0.8) among all estimates, but its RMSE and Bias (0.085 and -0.005 m^3 m^{-3}) are slightly larger than what GLASS SM has (0.083 and 0.004 m^3 m^{-3}). Since the GLASS SM uses in-situ data including those from USCRN as the target when training the ML model (i.e., not independent), it is not surprising the GLASS SM magnitudes better align with the USCRN data in general. The SMAPDA estimate tends to produce slightly smaller (larger) SM when observed SM is less than (above) 0.3 m^3m^{-3}. However, it also shows GLASS SM has more off-diagonal samples than SMAPDA, which degrades its overall performance. These two 1-km grid spacing products significantly outperform others. Constructed in 0.1-degree grid spacing, both ERA5-Land and GLEAMv4.1 may partially suffer from a relative coarser resolution. The SMAP AM has even worse skill in SM estimation given its highly scattered samples in the 2-D histogram despite relatively low bias. As a result, their CCs lie in a range from ~0.49 to ~0.66, and RMSEs are all greater than 0.1 m^3 m^{-3}. Meanwhile, relatively larger biases are also computed (0.017 and -0.006 m^3 m^{-3}).

From Figure 4, we also notice the cut-in (smallest) values of surface SM vary notably across the SM products. For example, the GLEAMv4.1 and SMAPDA have relative larger cut-in values of ~0.05 to 0.06 m^3m^{-3}. Whereas the ERA5-Land and SMAP AM are valid above approximately 0.02 m^3 m^{-3}. The GLASS SM has negligible limit on the smallest SM value. The differences in these cut-in SM values may be associated with either the formulations of land surface models or the observational sensitivities and could at least partially affect how good each estimate agrees with the observations.

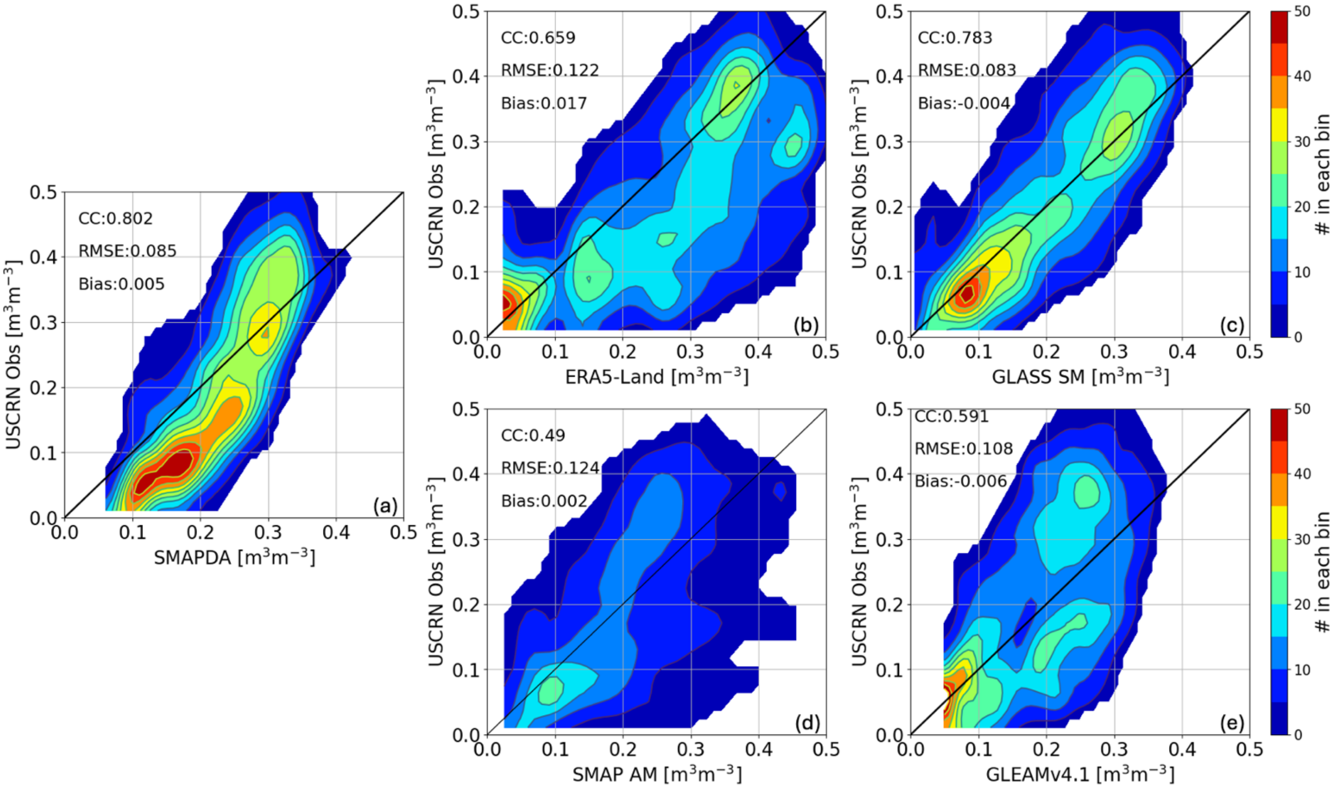

Figure 4 2-D histograms summarizing the evaluation results using the observational data measured by the USCRN network. 335 Panels (a) to (e) represent results of SMAPDA, ERA5-Land, GLASS SM, SMAP AM, and GLEAMv4.1. 50 bins are used to generate the 2D histograms. Correlation coefficient (CC), RMSE, and Bias are given in the upper left corner of each panel.

Figure 5 illustrates the disaggregated RMSEs at each USCRN site. While SMAPDA and GLASS SM have much smaller RMSEs (~ 0.08 $m^3 m^{-3}$) compared to other estimates in general (> 0.1 $m^3 m^{-3}$), SMAPDA has slightly smaller standard deviation of RMSEs than GLASS SM (0.035 versus 0.038 $m^3 m^{-3}$). This implies that SMAPDA performs more consistently across all 340 USCRN sites than GLASS SM. Conversely, GLEAMv4.1, ERA5-Land, and SMAP AM estimates have much larger errors especially for those sites in the southeast U.S., and coastal sites in Florida and Texas.

https://doi.org/10.5194/essd-2024-599
Preprint. Discussion started: 28 January 2025

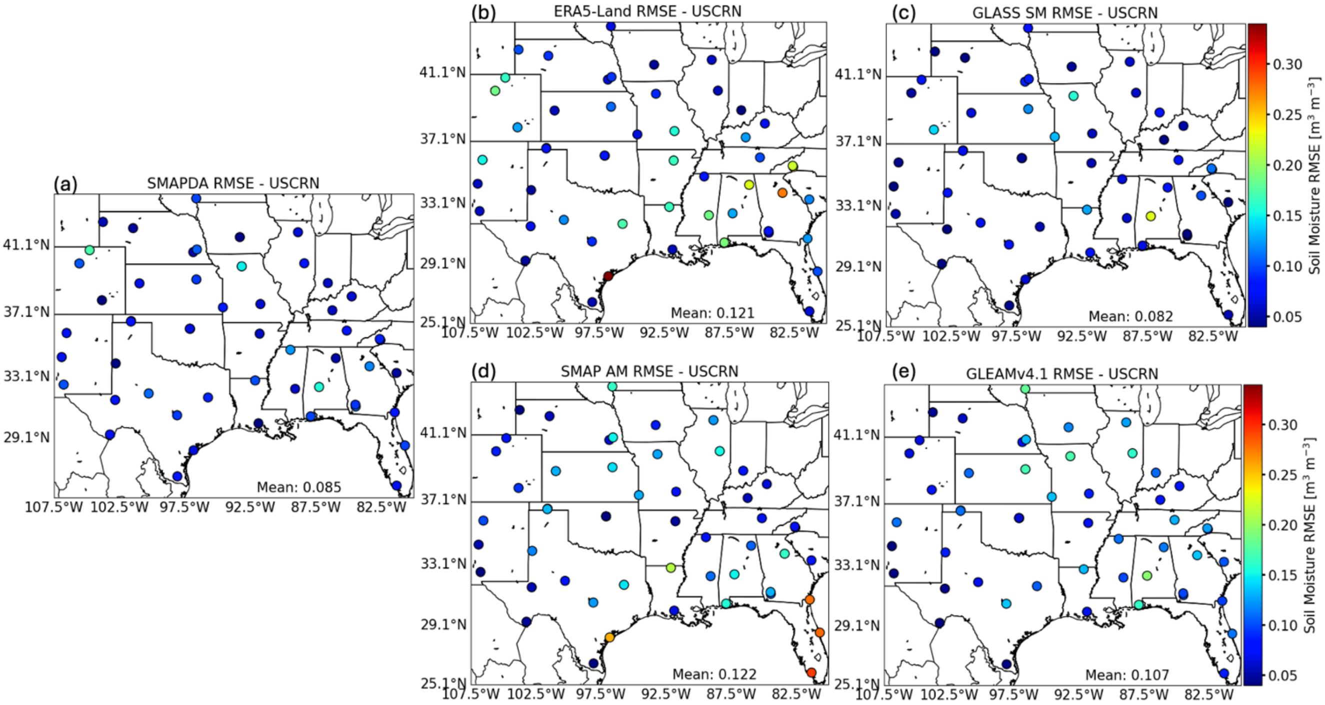

Figure 5 Site-wise RMSE computed using observations from the USCRN observations. Results for (a) SMAPDA, (b) ERA5-Land, (c) GLASS SM, (d) SMAP AM, and (e) GLEAMv4.1 are illustrated.

345 Likewise, the biases are displayed in Figure 6. In all SM datasets, wet bias is more evident in the southeastern U.S. sites than others, whereas dry bias is distinct across many sites in the northern and eastern Great Plains despite variability in their magnitudes. This consistent bias pattern implies that duplicate sources from observational data and/or treatments in the models may exist among those SM estimates.

https://doi.org/10.5194/essd-2024-599
Preprint. Discussion started: 28 January 2025

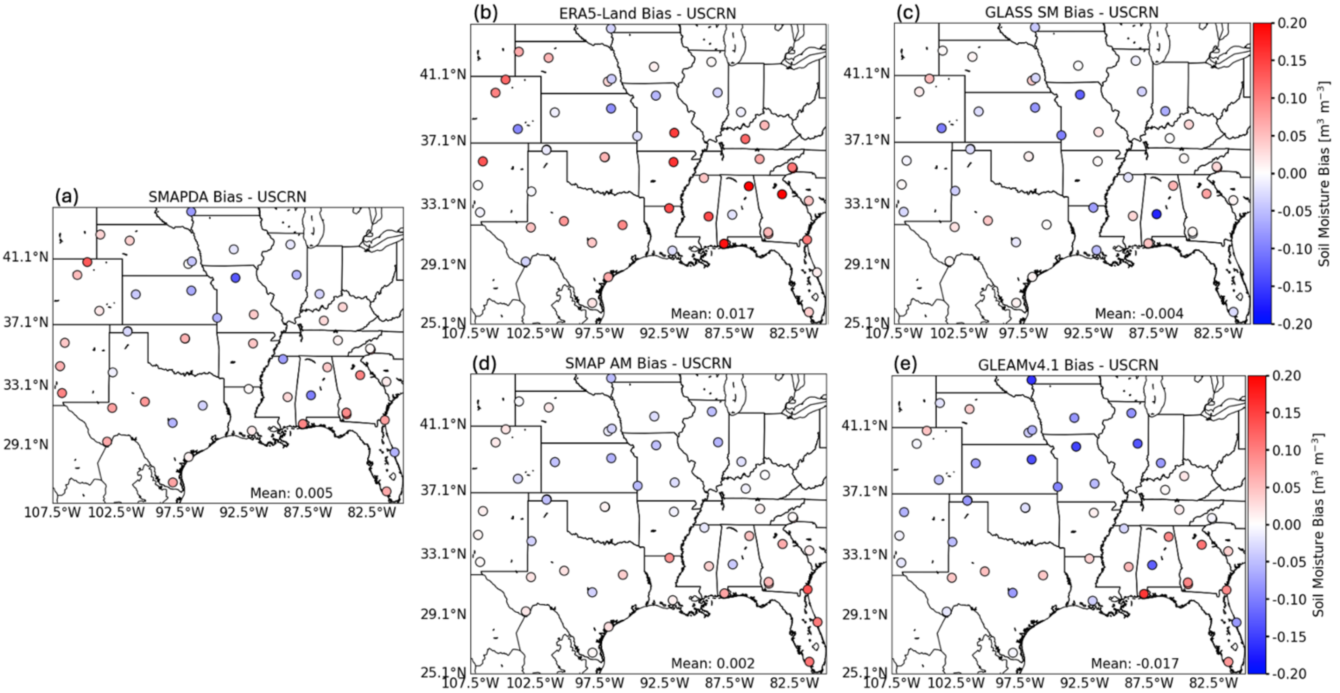

Figure 6 Similar to Figure 5, but for illustrations of Bias.

To further examine the potential errors in common among the five SM estimates, we calculated the RMSE anomaly for each dataset. The RMSE anomaly is obtained by subtracting annual mean RMSE (as shown in the panels of Figure 5) from the daily timeseries of each estimate. It extracts intrinsic variation in SM errors from the original SM timeseries and thus facilitate bias-free intercomparison. A diverse variation among the datasets is shown in Figure 7. Despite relatively large day-to-day variability in SMAP AM timeseries than other datasets, the multiday variability in SMAP AM is similar to GLEAMv4.1. For example, both of them show much larger SM errors from January to April and relatively smaller errors present in late spring and summer. The errors climb when it transitions into late fall and early winter. There is also much similarity between the ERA5-Land and GLASS SM timeseries. Despite minor discrepancies, compared to other datasets, they both show relatively smaller variation over the one-year period with slightly larger errors in April, June, and July. These results are not surprising as SMAP data is one of the ingredients of GLEAMv4.1 (Miralles et al., in review), whereas GLASS SM adopts ERA5-Land soil moisture as the SM input data in their ML model (Zhang et al. 2023). Figure 7 also indicates that SMAPDA demonstrates a unique trend with the smallest errors before June and peak errors occur in early July and late November.

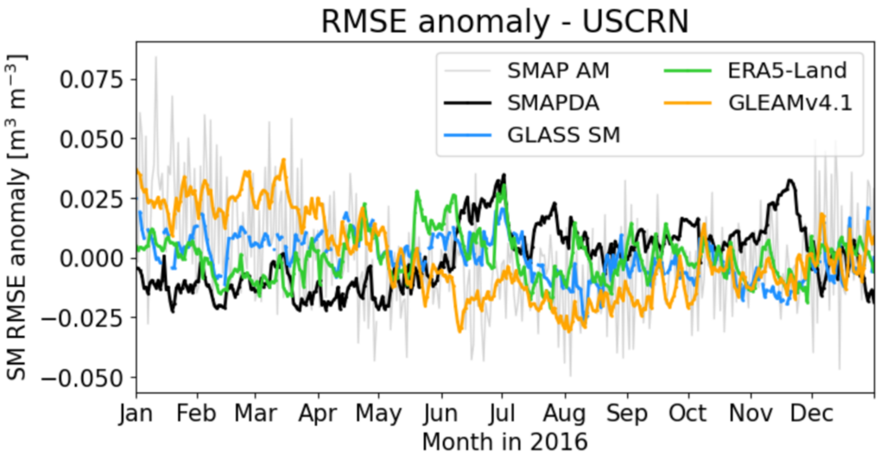

Figure 7 The SM RMSE anomaly timeseries computed against USCRN observations during 2016. Results of SMAP AM, ERA5-Land, SMAPDA, GLEAMv4.1, and GLASS SM are denoted by colored lines as indicated in the legend.

4.2.2 Evaluation using SCAN soil moisture observations

Along the same line as discussed in Section 4.2.1, we examined the SM 2-D histograms as referenced by the SCAN observations (Figure S1). Overall, similar conclusions can be drawn from the comparisons, implying that the evaluation is robust with very little dependence on target observations. For example, SMAPDA and GLASS SM remain the top two performers among the SM products. The CC, RMSE, and Bias for SMAPDA (GLASS SM) are 0.727 (0.687), 0.091 (0.095) m^3 m^{-3}, and -0.002 (0.005) m^3 m^{-3}. Again, despite better alignment with the diagonal in general, GLASS SM has more off-diagonal samples than the SMAPDA. Those samples are relatively fewer but carry large errors, therefore the skill scores for GLASS SM turn out to be slightly worse than SMAPDA. This is likely due in part to that GLASS SM uses ERA5-Land as the initial guess of SM. Since ERA5-Land has relatively scattered samples in the 2-D histogram (Figure S1) and ML algorithm does not overfit by design (Zhang et al. 2023), some pixels may receive less correction than others. The estimate from GLEAMv4.1 suffers from generally smaller SM estimates (capped around ~ 0.38 m^3 m^{-3}), which potentially causes severe underestimation. SMAP AM has the least bias among all estimates. However, it also owns many samples far off the diagonal, which lower the overall skill scores. As a result, the CCs for ERA5-Land, GLEAMv4.1, and SMAP_AM do not exceed 0.6. Moreover, their RMSEs are all greater than 0.11 m^3 m^{-3}, which is about 20% more than what is computed for SMAPDA.

The site-wise RMSEs given in Figure S2 confirm that SMAPDA is the top performer among the SM estimates in general as it shows consistency in producing small error across the SCAN sites. Specifically, the RMSEs of SMAPDA at all sites are below 0.15 m^3 m^{-3}, whereas that RMSE value is exceeded somewhere in all the other datasets. Excessive SM errors are found at several sites in the southeastern U.S. when evaluating the ERA5-Land and SMAP AM. Despite small contrast in RMSEs across the SCAN sites, the estimates from GLEAMv4.1 show extensively larger errors at more locations, leading to non-trivial mean RMSE of 0.11 m^3 m^{-3}. Figure S3 shows that, regardless of the in-situ observations, the mean bias and bias pattern of each SM estimate resembles those given in Figure 6, suggesting that the analyzed biases should be robust and representative.

Figure S4 shows that the SM RMSE anomaly of SMAPDA is very similar to what are obtained using GLASS SM and even SMAP AM (light gray) when assessed using the SCAN data. The ERA5-Land and GLEAMv4.1 exhibit trends partially different from the other three estimates. Specifically, ERA5-Land (GLEAMv4.1) has relatively smaller (larger) errors than the other three estimates from January to April and tends to produce rather larger (smaller) errors from October to December. Despite the results are not entirely identical with the comparison against USCRN (Figure 7), it remains clear that large similarity in RMSE anomalies is analyses. This is most likely due to duplicate of SM data source in different SM estimates even through different methods are employed.

4.2.3 Regional assessment over the Southern Great Plains

The Southern Great Plains (SGP), including Oklahoma, has been recognized as one of the hotspots for strong land-atmosphere coupling (LAC; Santanello et al. (2009); Tao et al. (2019)). Earlier studies revealed the key physical processes modulate the LAC strength and how it influences the lifecycle of convective clouds using the model and observational datasets generated for this region. For instance, Fast et al. (2018) investigated the impact of SM spatial heterogeneity on simulated convective clouds near the ARM SGP site using a Large-Eddy Simulation model for a selected event during the 2016 HI-SCALE field campaign. They found that the scales of SM gradient in the model can significantly affect the presence of simulated cloud populations even with identical atmospheric conditions. Sakaguchi et al. (2022) further analyzed the LES model data produced by Fast et al. (2019) using the spectrum analysis and demonstrated the SM spatial heterogeneity may strengthen secondary circulations and extend their spatial scales. Both studies concluded that a more realistic and high-resolution representation of SM is desired to better understand LAC at local-to-regional scales (~1 km and greater). This motivates us to examine how SMAPDA SM estimate perform in this region in comparison to other datasets and the evaluations are carried out by leveraging highly concentrated observations measured by the OKMet (Figure 1).

As shown in Figure 8, the performance of GLASS SM degrades when evaluated by the OKMet data. The 2D histogram shows most samples occur in the bins above the diagonal, meaning that GLASS SM generally underestimates SM (mean Bias: -0.039 m^3 m^{-3}). Whereas when comparing with data from USCRN and SCAN (Figures 4 and S1), GLASS SM has much better agreement with the observations. This is most likely due to the exclusion of OKMet data in their ML training process. Using a fully data-driven approach, the skill of ML-based SM estimate highly depends on the availability of in-situ observations. Conversely, SMAPDA exhibits the lowest annual mean RMSE (0.087 m^3 m^{-3}) in general in comparison to the other four datasets (0.105, 0.107, 0.103, and 0.112 m^3 m^{-3} for ERA5-Land, GLASS SM, SMAP_AM, and GLEAMv4.1, respectively). The annual mean RMSE stays close to the RMSEs obtained when comparing against the USCRN and SCAN observations

(0.085 and 0.091 m^3 m^{-3}, respectively). This demonstrates that a physically constrained model may perform more consistent and better mitigate biased soil moisture estimates despite uncertain and unresolved physical processes.

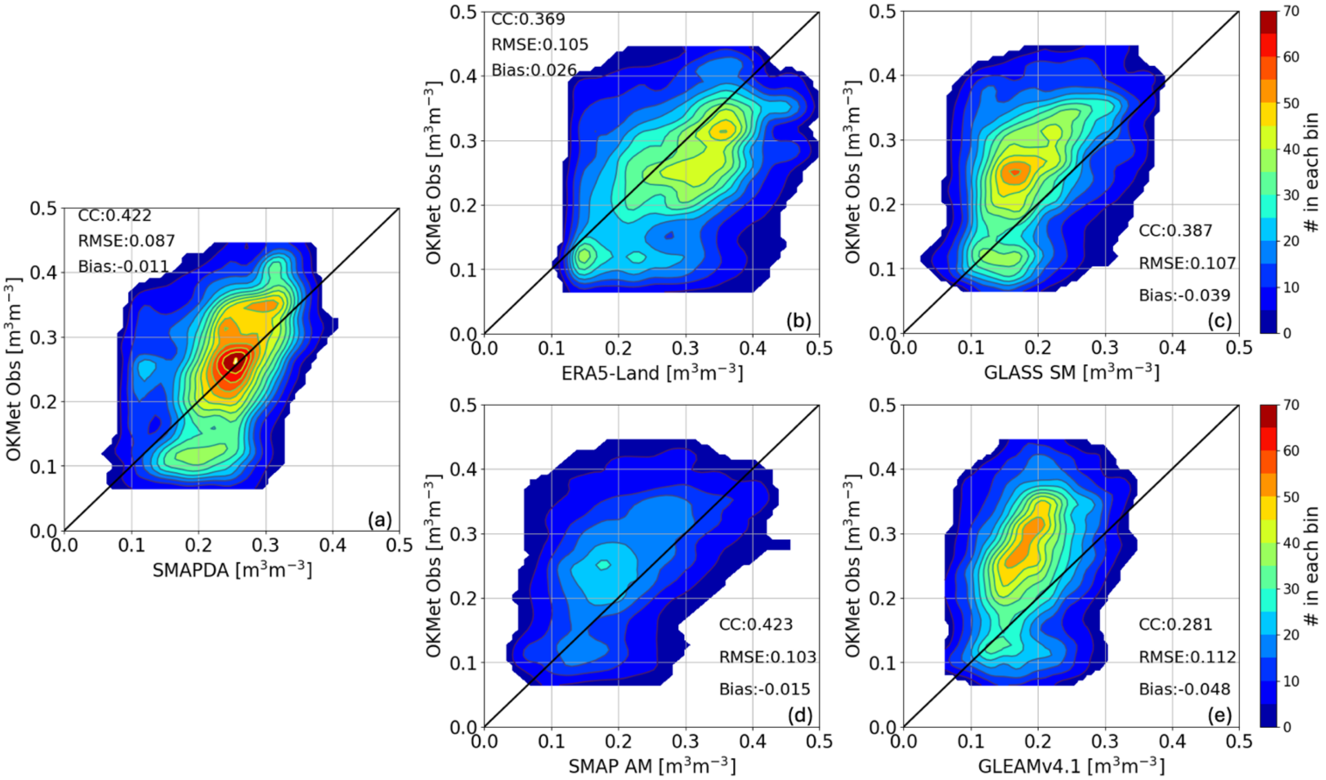

Figure 8 Similar to Figures 4 and S1, but for results computed against the OKMet observations.

The relatively dense spatial distribution of the OKMet sites enables further investigation into the realism of estimated SM spatial heterogeneity. We computed daily standard deviation (STD) across all OKMet sites for each SM estimate as a way to quantify the spatial SM heterogeneity (meaning how spread the SM values are in space). Figure 9 shows that observed STD (magenta) is mostly larger than what is estimated by any of the derived SM approaches over the year despite notable day-to-day variations. Even though SMAPDA and GLASS SM top the others in SM estimates based on the evaluations shown earlier, they both underestimate the SM spatial heterogeneity with an averaged STD ~ 0.6 m^3 m^{-3}, which is about 25% less than observed. GLEAMv4.1 and SMAP AM have even smaller STDs over the period. While ERA5-Land tends to have larger and more comparable variances as observed, it does not accurately distribute those SM values in space (Figure S5).

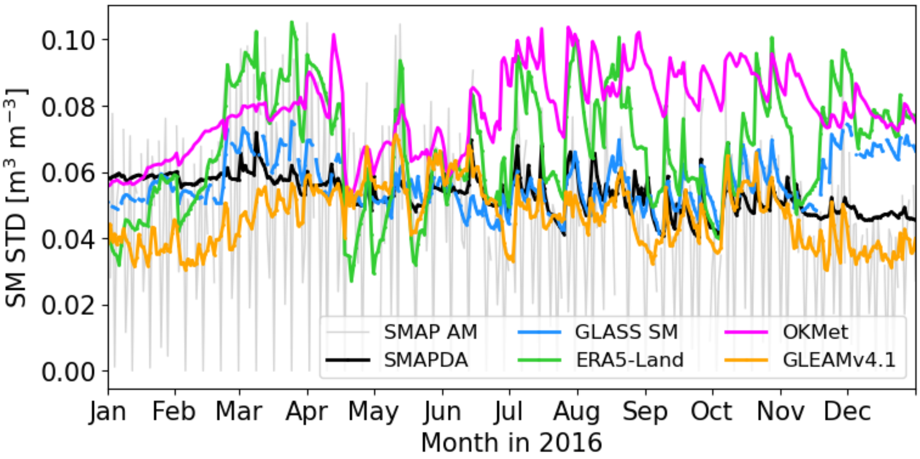

Figure 9 Daily timeseries of SM standard deviation across OKMet sites computed for each SM estimate.

In addition to soil moisture, the ARM SGP facility (through instruments of STAMP and ECOR) has collected soil temperature, surface heat fluxes (latent and sensible) that are critical for land-atmosphere coupling research and directly modulate the strength of turbulent mixing in the atmospheric boundary layer. Here we primarily assess how SMAPDA represents SM, soil temperature (ST), latent heat flux (LHF), and sensible heat flux (SHF) using the concurrent measurements collected across six ARM SGP sites (E31, E33, E37, E38, E39, and E41) as denoted in Figure1a. Note SM and ST data from the STAMP are not valid in January.

Results show that SMAPDA reproduces the observed monthly trends in SM and slightly overestimates SM (annual mean model-observation difference of + 0.04 m^3 m^{-3}) with larger positive biases in winter months (Figure 10a). The annual mean ST is warmer in SMAPDA than observed (difference: + 0.84 K) which can be attributed to relatively distinct warm bias in summer months (June – September) (Figure 10b). While LHF has an annual mean difference of -12.14 W m^{-2} when compared to the observations (Figure 10c), it is considered minor as annual LSM error can be ~ - 20 W m^{-2} based on earlier studies (ARM 2014). Whereas in the case of SHF, SMAPDA tends to overestimate for most of the months (Figure 10d). This is likely due to consistent wet bias in SM throughout the year (Figure 10a), leading to increased energy partition in latent heat and alternatively reduces the component of sensible heat. Distinct positive biases even appear over summer months (June and July) despite higher ST is simulated (Figure 10b).

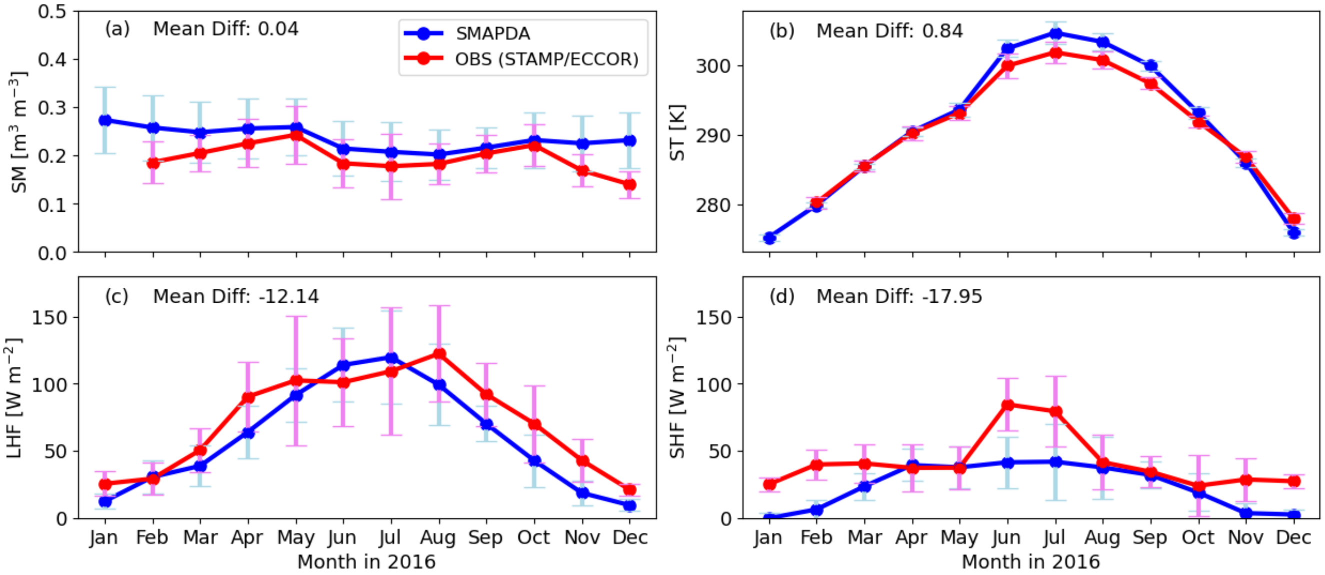

Figure 10 Timeseries of monthly mean of (a) SM, (b) ST, (c) LHF, and (d) SHF. Blue (red) line with circle represent results obtained from SMAPDA simulations (ARM SGP observations [STAMP/ECOR]) as computed across six ARM SGP sites (Figure 1a)

4.2.4 Regional analysis associated with 2016 drought in the Southeastern U.S.

The southeastern U.S. experienced one of the most significant drought events in the region during the fall of 2016 (peaked in October and November) based on historical record (Park Williams et al. 2017). It primarily affected parts of Georgia, Alabama, Tennessee, and the Carolinas. The drought reached extreme and exceptional levels, especially in northern Georgia and Alabama, where some areas experienced their driest conditions in history. A combination of factors, including below-average rainfall during the spring and summer months and unusually high temperatures led to the increased evaporation and reduced soil moisture and thereby the drought conditions in fall. The drought severely impacted agriculture, leading to reduced crop yields, and contributed to widespread wildfires in the Appalachian region. The strained water resources also posed a great challenge on water availability for communities and industries. Hence, we aim to explore the representativeness of the SMAPDA SM estimate under the extreme drought conditions in the southeastern U.S.

A subdomain “AL” which covers Alabama (as denoted in Figure 1a) was chosen for conducting the following analyses since the in-situ measurements from USCRN and SCAN are relatively denser in Alabama than in other areas in the southeastern U.S. (Figure 1a). In addition to the spatial variability of soil types (Figure 11a), the forest (Figure 11b) can further complicate how soil moisture is distributed through hydraulic processes such as evapotranspiration (ET), interception, infiltration, runoff, groundwater recharge, and hydraulic redistribution due to the presence of root systems and tree canopies. Here, we selectively examine the relationship between SM and ET under the drought conditions by comparing the SMAPDA output with the GLEAMv4.1 data.

Both SMAPDA and GLEAMv4.1 exhibit a decreasing trend in SM over the summer months (JJA) as well as a steeper decline in fall (SON) over the AL subdomain (Figure 12a). Except in winter months (DJF) where the SM estimate is slightly larger in SMAPDA than in GLEAMv4.1, soil conditions produced by SMAPDA are consistently drier than GLEAMv4.1. Both datasets suggest increases in ET before June with similar magnitudes (Figure 12c), mostly due to the seasonal increase in the solar insolation as well as the leaf area. However, in summer (JJA), SMAPDA produces much larger ET than GLEAMv4.1 does, which leads to much drier soil conditions concurrently. This then facilitates the intensification of drought conditions in the fall, leading to further reduction in water availability through the soil columns which significantly limits the amount of ET as opposed to GLEAMv4.1.

Although in-situ ET observations are not available through the USCRN and SCAN measurements, the SM observations (Figure 12b) suggest while SMAPDA overall captures how SM evolves over time, GLEAMv4.1 gives much weaker responses to the SM drying process than SMAPDA. This ultimately produces overall much larger wet bias in SM than SMAPDA in the fall. Whether data from all grid cells in the AL subdomain (Figures 12a and 12c) or only the 17 grid cells nearest to the in-situ measurements (Figures 12b and 12d) are used, the trends in both SM and ET are very similar. This suggests that the evaluation illustrated in Figures 12b and 12d are representative for the subdomain. As we investigate in more detail through comparison at each individual site (Figure 13), we find most of the large errors in SMAPDA's SM estimate can be attributed to sites' soil properties (Figure 13), specifically where the clay soil types are present (site #3 and 5: Clay; site# 7: Silty clay). At those sites, the enhanced temporal variability in SM is distinct but underestimated by both SMAPDA and GLEAMv4.1 estimates. This suggests both approaches are unable to capture the drastic changes in SM, likely due in part to the nature of clay soil texture. Similar to what was reported in Colliander et al. (2022), further model refinements may be needed to improve treatments in resolving hydraulic processes for the variants of clay soil.

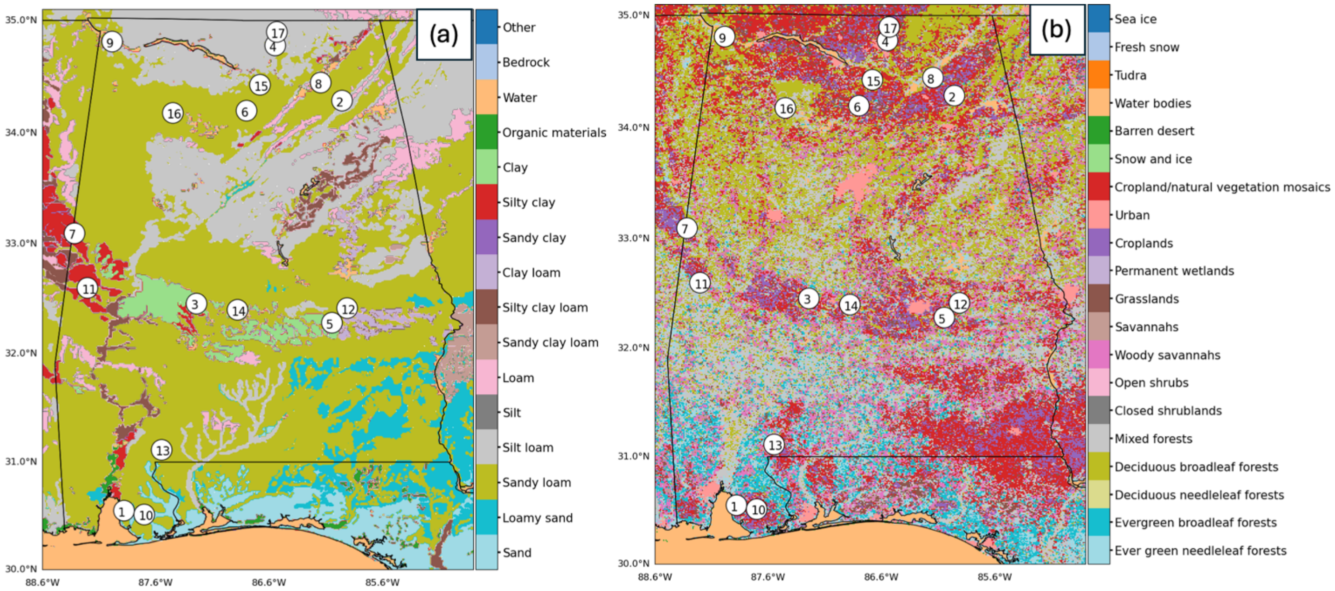

Figure 11 The zoomed-in maps of (a) soil and (b) land cover types over the AL subdomain as marked in Figure 1a. The locations of 17 valid observational sites from USCRN and SCAN are denoted by the white circles with numbers overlaid in correspondence of panels in Figure 13.

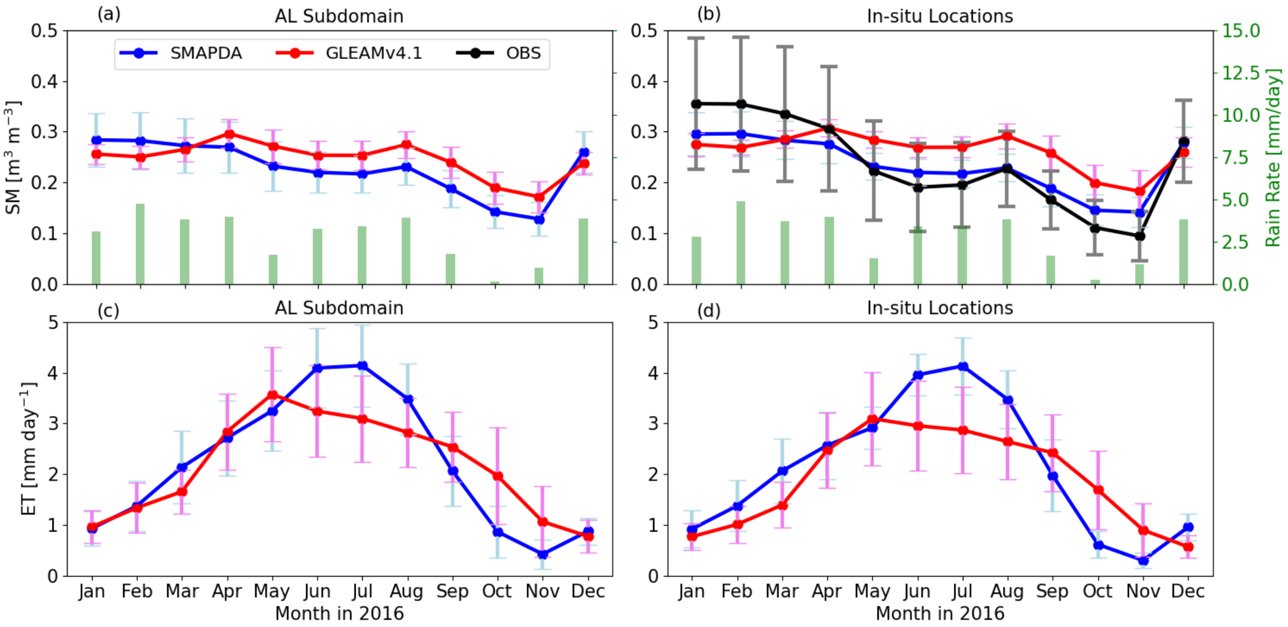

Figure 12 Monthly mean and standard deviation (denoted by error bars) of SM (a, b) and ET (c, d) over the AL subdomain and among 17 in-situ observation locations as denoted in Figure 11. The mean rain rate represented by green bars in (a) and (b) are computed from SMAPDA correspondingly.

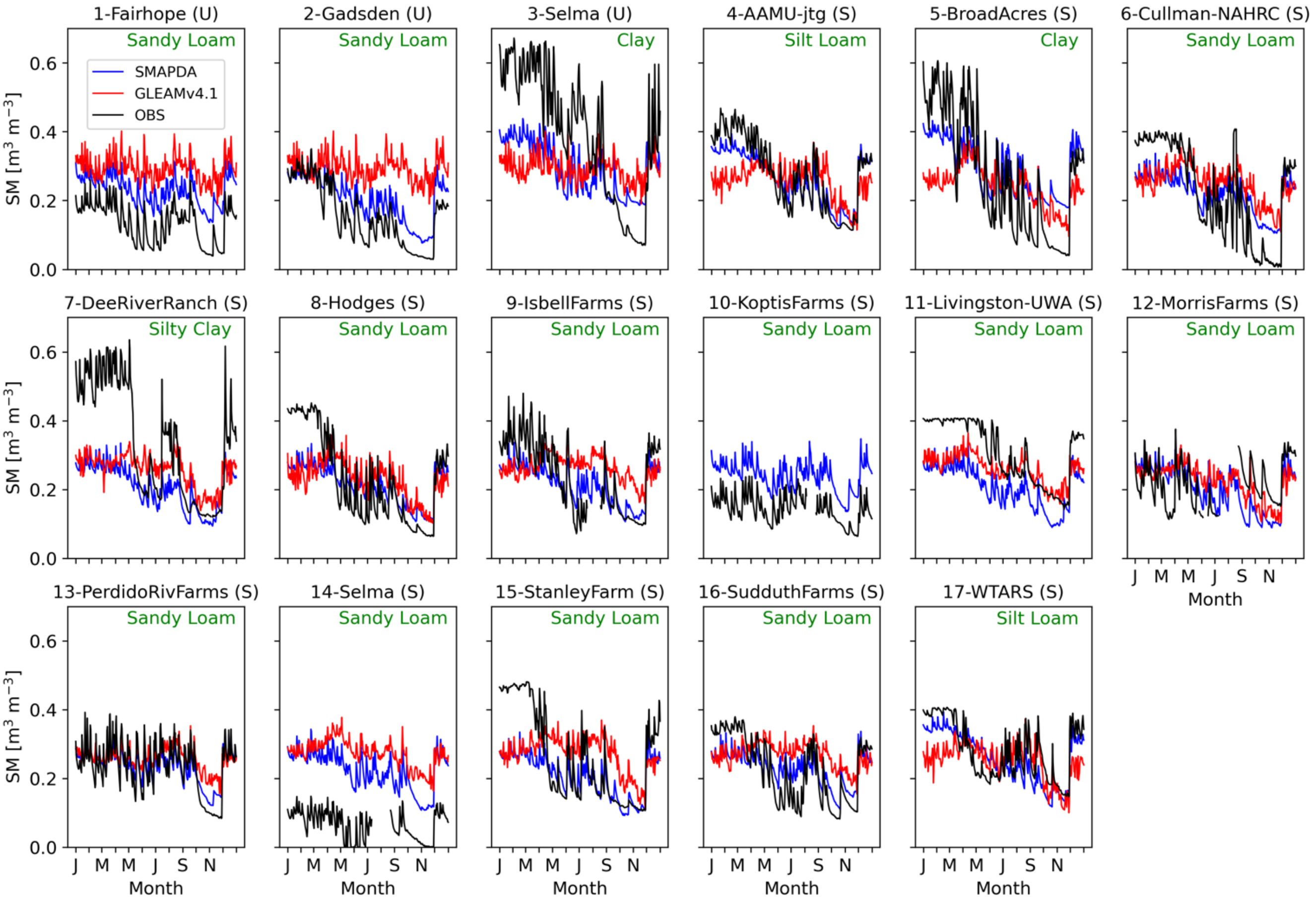

Figure 13 SM daily timeseries comparison at each in-situ observation location. The numbers given in the title above each panel correspond to the locations as marked by the numbered white circles in Figure 11. The site names and the corresponding observational networks (as indicated by either U (USCRN) or S (SCAN) in the parenthesis) are readable from the titles. Soil 500 texture type is indicated by green texts in the top right corner of each panel.

5. Summary and discussion

To facilitate an improved representation of local-to-regional scale SM distribution, we generated a high-resolution SM dataset at a 1-km grid spacing by assimilating the 9-km SMAP SM data into the Noah-MP land surface model. The dataset has 505 a spatial coverage over the east CONUS and has frequency of 6 hours for the entire 2016. The SMAP SM data assimilation is accomplished under the framework of NASA's Land Information System using the EnKF algorithm. In the DA simulation, 12 ensemble members were initialized by perturbing the selected variables in meteorological forcing data (NLDAS-2 and Stage IV). The subset of daily SMAP SM overpasses is assimilated hourly when applicable. The generated SM estimate is comprehensively assessed by using the in-situ SM observations collected in the networks of USCRN, SCAN, OKMet, and

ARM SGP and compared with the performance of other existing SM datasets such as the morning overpass of SMAP (SPL3SMP_E) data, ERA5-Land, GLASS SM, and GLEAMv4.1.

Overall, the evaluation result suggests the resulting soil moisture estimate, which we refer to as SMAPDA, exhibits the top performance among the examined datasets. While the SMAPDA and GLASS SM are considered the top two SM estimates based on the skill metrics computed against USCRN and SCAN observations (e.g., CCs are ~0.8 and ~0.7 and RMSEs are ~0.08 and ~0.09 m^3 m^{-3}, respectively), SMAPDA surpasses GLASS SM when validated against OKMet data (independent observations for both SMAPDA and GLASS SM). Being a fully data-driven ML product, the GLASS SM achieves a better one-to-one alignment with the observations than SMAPDA when evaluated by the in-situ data that used in its training process (USCRN and SCAN). However, the relative accuracy of GLASS SM and SMAPDA reverses when compared with the independent observations from OKMet, which implies the inclusion of physical constraints could be vital for a more consistent performance in SM estimate using the ML approach. From the analysis in anomalous errors, we show similar intrinsic errors among the selected SM datasets in some cases, which is most likely driven by overlapping data sources. Referenced by the OKMet observations, an investigation on the realism of estimated SM spatial heterogeneity indicates all SM estimates, including the SMAPDA and GLASS SM, persistently underestimate the observed variances (~ 25% less) across the sites over the study period. While the ERA5-Land estimate shows larger and more comparable variances as observed, it does not accurately represent those SM values individually.

In addition to SM data, we showed that SMAPDA data reasonably represent ST and even surface heat fluxes when compared against the observations measured in ARM SGP sites. This suggests the suite of SMAPDA dataset is useful in characterizing land-atmosphere interactions. Moreover, it is also analyzed with respect to the response to a drought that occurred over the southeastern U.S. during the fall of 2016. As one of the key components contributing to the drought, the reduction in SM is usually accompanied by increased evaporation in the water-limited scenario, which may potentially amplify and increase wildfire activity and stresses on agricultural production until new precipitation. We explored the relationships between SM and ET with a focus on Alabama quantitatively, utilizing concurrent GLEAMv4.1 data as the reference. Results indicate both datasets showed declined SM in summer and fall, with SMAPDA consistently displaying drier soil conditions compared to GLEAMv4.1. ET trends from both datasets were relatively close til June but diverged in summer, with SMAPDA estimating higher ET, exacerbating the drought conditions. Data also highlighted that model discrepancies, particularly in clay-rich soils, suggest the need for refined treatments of hydraulic process in models for accurate SM estimates.

A few uncertainties in our analysis are worth noting. For example, the evaluation result is most likely dependent on the data resolutions. Coarse resolution SM estimates such as ERA5-Land, GLEAMv4.1, and SMAP AM suffer from insufficient representativeness of subgrid SM variability, underscoring the necessity of high resolution to better characterize highly heterogeneous SM distributions. In addition, unresolved natural and anthropogenic processes such as surface and subsurface lateral flow (e.g, Yang et al. 2021), root water uptake and redistribution (e.g., Zeng 2001), dynamic groundwater water table and capillary rise (e.g., Miguez-Macho and Fan 2012), and irrigation (e.g., Yang et al. 2020) can potentially shift the SM estimates under various conditions. While our SM dataset encompassing much of the eastern CONUS is restricted to a one-year period (2016), our results demonstrate a promising approach that can be applied to any local domain of interest with potentially longer analysis periods. This dataset could be used as lower boundary conditions to drive other meteorological model experiments that investigate the impact of land-atmosphere coupling on boundary layer properties and clouds. Lastly, there are many more ML algorithms, such as neural networks, random forests, and support vector machines, have been applied to enhance the spatial and temporal resolution of soil moisture datasets and further improve accuracy in data-sparse regions (e.g., O. and Orth 2021; Han et al. 2023; Lei et al. 2022; Zhang et al. 2023). However, such approaches lack the inherent physical constraints of a data-assimilation approach. Future studies may include more ML-based products in the assessment and discuss the impacts of physical constraint on estimated SM as suggested in this study.

Acknowledgments

This research was supported by the Atmospheric Science Research (ASR) program as part of the DOE Office of Biological and Environmental Research. Pacific Northwest National Laboratory is operated by DOE by the Battelle Memorial Institute under contract DE-AC05-76RL01830. The modelling computation was performed on PNNL's Research Computing clusters. The authors would like to thank for all the technical supports through the LIS Github (https://github.com/NASA-LIS/LISF/discussions) and particularly the responsible research scientist David Mocko.

Data availability

The dataset generated and analyzed during the current study (Tai et al. 2024) is available on Zenodo at https://doi.org/10.5281/zenodo.14370563. The software package of the NASA Land Information System (LIS) can be downloaded through https://github.com/NASA-LIS/LISF (Kumar et al. 2006; Peters-Lidard et al. 2007). The enhanced SMAP Level 3 soil moisture product (SPL3SMP_E; O'Neill et al., 2020) is accessible at https://nsidc.org/data/spl3smp_e/versions/6. The NLDAS-2 data (Xia et al. 2012) is archived at https://disc.gsfc.nasa.gov/datasets/NLDAS_FORA0125_H_2.0/summary?keywords=NLDAS2. The Stage IV QPE product (Lin and Mitchell (2005)) is accessible at https://data.ucar.edu/dataset/ncep-emc-4km-gridded-data-grib-stage-iv-data. The USCRN and SCAN data are acquired from the International Soil Moisture Network (ISMN; https://ismn.earth/en/, last access: 1 February 2024). The Oklahoma mesonet soil moisture observations (OKMet; McPherson et al. (2007)) were acquired from the ARM discovery website: https://adc.arm.gov//discovery/#/results/s::sgpokmsoilX1.c1 (ARM facility (1998)). All other SM datasets please refer to Table 3. The STAMP (Kyrouac et al., n.d., https://doi.org/10.5439/1238260) and ECOR (Gaustad, n.d., https://doi.org/10.5439/1097546) data were sourced from the Atmospheric Radiation Measurement (ARM) user facility.

Competing interests

The contact author has declared that none of the authors has any competing interests.

References

Ahmad, J. A., B. A. Forman, and S. V. Kumar, 2022: Soil moisture estimation in South Asia via assimilation of SMAP retrievals. *Hydrology and Earth System Sciences*, **26**, 2221–2243, https://doi.org/10.5194/hess-26-2221-2022.

Arsenault, K. R., and Coauthors, 2018: The Land surface Data Toolkit (LDT v7.2) – a data fusion environment for land data assimilation systems. *Geoscientific Model Development*, **11**, 3605–3621, https://doi.org/10.5194/gmd-11-3605-2018.

Ball, J. T., I. E. Woodrow, and J. A. Berry, 1987: A Model Predicting Stomatal Conductance and its Contribution to the Control of Photosynthesis under Different Environmental Conditions. *Progress in Photosynthesis Research: Volume 4 Proceedings of the VIIth International Congress on Photosynthesis Providence, Rhode Island, USA, August 10–15, 1986*, J. Biggins, Ed., Springer Netherlands, 221–224.

Betts, A. K., 2002: Surface diurnal cycle and boundary layer structure over Rondônia during the rainy season. *Journal of Geophysical Research*, **107**, https://doi.org/10.1029/2001JD000356.

Brutsaert, W., 1982: *Evaporation into the Atmosphere*. Springer Netherlands,.

Chakraborty, A., M. Saharia, S. Chakma, D. Kumar Pandey, K. Niranjan Kumar, P. K. Thakur, S. Kumar, and A. Getirana, 2024: Improved soil moisture estimation and detection of irrigation signal by incorporating SMAP soil moisture into the Indian Land Data Assimilation System (ILDAS). *Journal of Hydrology*, **638**, 131581, https://doi.org/10.1016/j.jhydrol.2024.131581.

Chao, L., K. Zhang, S. Wang, Z. Gu, J. Xu, and H. Bao, 2022: Assimilation of surface soil moisture jointly retrieved by multiple microwave satellites into the WRF-Hydro model in ungauged regions: Towards a robust flood simulation and forecasting. *Environmental Modelling & Software*, **154**, 105421, https://doi.org/10.1016/j.envsoft.2022.105421.

Colliander, A., and Coauthors, 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.

Colliander, A., and Coauthors, 2022: Validation of Soil Moisture Data Products From the NASA SMAP Mission. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **15**, 364–392, https://doi.org/10.1109/JSTARS.2021.3124743.

Crow, W. T., and E. F. Wood, 2003: The assimilation of remotely sensed soil brightness temperature imagery into a land surface model using Ensemble Kalman filtering: a case study based on ESTAR measurements during SGP97. *Advances in Water Resources*, **26**, 137–149, https://doi.org/10.1016/S0309-1708(02)00088-X.

Diamond, H. J., and Coauthors, 2013: U.S. Climate Reference Network after One Decade of Operations: Status and Assessment. https://doi.org/10.1175/BAMS-D-12-00170.1.

Dirmeyer, P. A., and Coauthors, 2016: Confronting Weather and Climate Models with Observational Data from Soil Moisture Networks over the United States. https://doi.org/10.1175/JHM-D-15-0196.1.

Dorigo, W., and Coauthors, 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.

——, and Coauthors, 2021: The International Soil Moisture Network: serving Earth system science for over a decade. *Hydrology and Earth System Sciences*, **25**, 5749–5804, https://doi.org/10.5194/hess-25-5749-2021.

Ek, M. B., and A. a. M. Holtslag, 2004: Influence of Soil Moisture on Boundary Layer Cloud Development.

——, K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *Journal of Geophysical Research: Atmospheres*, **108**, https://doi.org/10.1029/2002JD003296.

Entekhabi, D., and Coauthors, 2010: The Soil Moisture Active Passive (SMAP) Mission. *Proceedings of the IEEE*, **98**, 704–716, https://doi.org/10.1109/JPROC.2010.2043918.

Farr, T. G., and Coauthors, 2007: The Shuttle Radar Topography Mission. *Reviews of Geophysics*, **45**, https://doi.org/10.1029/2005RG000183.

Fast, J. D., and Coauthors, 2018: Overview of the HI-SCALE Field Campaign: A New Perspective on Shallow Convective Clouds. *Bull. Amer. Meteor. Soc.*, https://doi.org/10.1175/BAMS-D-18-0030.1.

Friedl, M. A., and Coauthors, 2002: Global land cover mapping from MODIS: algorithms and early results. *Remote Sensing of Environment*, **83**, 287–302, https://doi.org/10.1016/S0034-4257(02)00078-0.

Gaustad, Krista. "Quality Controlled Eddy Correlation Flux Measurement (30QCECOR)." Atmospheric Radiation Measurement (ARM) user facility, n.d. Accessed August 21, 2024. https://doi.org/10.5439/1097546.

Gutman, G., and A. Ignatov, 1998: The derivation of the green vegetation fraction from NOAA/AVHRR data for use in numerical weather prediction models. *International Journal of Remote Sensing*, **19**, 1533–1543, https://doi.org/10.1080/014311698215333.

Han, Q., Y. Zeng, L. Zhang, C. Wang, E. Prikaziuk, Z. Niu, and B. Su, 2023: Global long term daily 1 km surface soil moisture dataset with physics informed machine learning. *Sci Data*, **10**, 101, https://doi.org/10.1038/s41597-023-02011-7.

Hawdon, A., D. McJannet, and J. Wallace, 2014: Calibration and correction procedures for cosmic-ray neutron soil moisture probes located across Australia. *Water Resources Research*, **50**, 5029–5043, https://doi.org/10.1002/2013WR015138.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hsu, H., and P. A. Dirmeyer, 2023: Soil moisture-evaporation coupling shifts into new gears under increasing CO2. *Nat Commun*, **14**, 1162, https://doi.org/10.1038/s41467-023-36794-5.

Jordan, R., 1991: A One-dimensional temperature model for a snow cover : technical documentation for SNTHERM.89. https://www.semanticscholar.org/paper/A-One-dimensional-temperature-model-for-a-snow-%3A-Jordan/ae2d518793624a2d5b9d5395a5dfdf2055c2b970 (Accessed October 26, 2024).

Katul, G. G., R. Oren, S. Manzoni, C. Higgins, and M. B. Parlange, 2012: Evapotranspiration: A process driving mass transport and energy exchange in the soil-plant-atmosphere-climate system. *Reviews of Geophysics*, **50**, https://doi.org/10.1029/2011RG000366.

Kerr, Y. H., P. Waldteufel, J.-P. Wigneron, J. Martinuzzi, J. Font, and M. Berger, 2001: Soil moisture retrieval from space: the Soil Moisture and Ocean Salinity (SMOS) mission. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 1729–1735, https://doi.org/10.1109/36.942551.

Koster, R. D., and Coauthors, 2004: Regions of Strong Coupling Between Soil Moisture and Precipitation. *Science*, **305**, 1138–1140, https://doi.org/10.1126/science.1100217.

Kumar, S. V., and Coauthors, 2006: Land information system: An interoperable framework for high resolution land surface modeling. *Environmental Modelling & Software*, **21**, 1402–1415, https://doi.org/10.1016/j.envsoft.2005.07.004.

Kumar, S. V., R. H. Reichle, C. D. Peters-Lidard, R. D. Koster, X. Zhan, W. T. Crow, J. B. Eylander, and P. R. Houser, 2008: A land surface data assimilation framework using the land information system: Description and applications. *Advances in Water Resources*, **31**, 1419–1432, https://doi.org/10.1016/j.advwatres.2008.01.013.

Kyrouac, Jenni, David Cook, Brian Ermold, Sujan Pal, Ryan Sullivan, and Evan Keeler. "Soil Temperature and Moisture Profiles (STAMP)." Atmospheric Radiation Measurement (ARM) user facility, n.d. Accessed August 23, 2024. https://doi.org/10.5439/1238260.

Lawrence, D. M., and Coauthors, 2019: The Community Land Model Version 5: Description of New Features, Benchmarking, and Impact of Forcing Uncertainty. *Journal of Advances in Modeling Earth Systems*, **11**, 4245–4287, https://doi.org/10.1029/2018MS001583.

Lei, F., V. Senyurek, M. Kurum, A. C. Gurbuz, D. Boyd, R. Moorhead, W. T. Crow, and O. Eroglu, 2022: Quasi-global machine learning-based soil moisture estimates at high spatio-temporal scales using CYGNSS and SMAP observations. *Remote Sensing of Environment*, **276**, 113041, https://doi.org/10.1016/j.rse.2022.113041.

Liang, S., and Coauthors, 2021: The Global Land Surface Satellite (GLASS) Product Suite. https://doi.org/10.1175/BAMS-D-18-0341.1.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges, 1994: A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research: Atmospheres*, **99**, 14415–14428, https://doi.org/10.1029/94JD00483.

Liu, P.-W., and Coauthors, 2021: Assessing Disaggregated SMAP Soil Moisture Products in the United States. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **14**, 2577–2592, https://doi.org/10.1109/JSTARS.2021.3056001.

Martens, B., and Coauthors, 2017: GLEAM v3: satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development*, **10**, 1903–1925, https://doi.org/10.5194/gmd-10-1903-2017.

McPherson, R. A., and Coauthors, 2007: Statewide Monitoring of the Mesoscale Environment: A Technical Update on the Oklahoma Mesonet. *J. Atmos. Oceanic Technol.*, **24**, 301–321, https://doi.org/10.1175/JTECH1976.1.

Miguez-Macho, G., and Y. Fan, 2012: The role of groundwater in the Amazon water cycle: 1. Influence on seasonal streamflow, flooding and wetlands. *Journal of Geophysical Research: Atmospheres*, **117**, https://doi.org/10.1029/2012JD017539.

Muñoz-Sabater, J., and Coauthors, 2021: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, **13**, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021.

Niu, G.-Y., Z.-L. Yang, R. E. Dickinson, L. E. Gulden, and H. Su, 2007: Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data. *Journal of Geophysical Research: Atmospheres*, **112**, https://doi.org/10.1029/2006JD007522.

——, and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research: Atmospheres*, **116**, https://doi.org/10.1029/2010JD015139.

Njoku, E. G., T. J. Jackson, V. Lakshmi, T. K. Chan, and S. V. Nghiem, 2003: Soil moisture retrieval from AMSR-E. *IEEE Transactions on Geoscience and Remote Sensing*, **41**, 215–229, https://doi.org/10.1109/TGRS.2002.808243.

O., S., and R. Orth, 2021: Global soil moisture data derived through machine learning trained with in-situ measurements. *Sci Data*, **8**, 170, https://doi.org/10.1038/s41597-021-00964-1.

Park Williams, A., B. I. Cook, J. E. Smerdon, D. A. Bishop, R. Seager, and J. S. Mankin, 2017: The 2016 Southeastern U.S. Drought: An Extreme Departure From Centennial Wetting and Cooling. *Journal of Geophysical Research: Atmospheres*, **122**, 10,888-10,905, https://doi.org/10.1002/2017JD027523.

Peters-Lidard, C. D., and Coauthors, 2007: High-performance Earth system modeling with NASA/GSFC's Land Information System. *Innovations Syst Softw Eng*, **3**, 157–165, https://doi.org/10.1007/s11334-007-0028-x.

Poggio, L., L. M. de Sousa, N. H. Batjes, G. B. M. Heuvelink, B. Kempen, E. Ribeiro, and D. Rossiter, 2021: SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL*, **7**, 217–240, https://doi.org/10.5194/soil-7-217-2021.

Rasheed, M. W., and Coauthors, 2022: Soil Moisture Measuring Techniques and Factors Affecting the Moisture Dynamics: A Comprehensive Review. *Sustainability*, **14**, 11538, https://doi.org/10.3390/su141811538.

Reichle, R. H., and R. D. Koster, 2004: Bias reduction in short records of satellite soil moisture. *Geophysical Research Letters*, **31**, https://doi.org/10.1029/2004GL020938.

Reichle, R. H., D. B. McLaughlin, and D. Entekhabi, 2002: Hydrologic Data Assimilation with the Ensemble Kalman Filter.

Robinson, D. A., and G. Kukla, 1985: Maximum Surface Albedo of Seasonally Snow-Covered Lands in the Northern Hemisphere.

Robock, A., K. Y. Vinnikov, G. Srinivasan, J. K. Entin, S. E. Hollinger, N. A. Speranskaya, S. Liu, and A. Namkhai, 2000: The Global Soil Moisture Data Bank.

Rouf, T., M. Girotto, P. Houser, and V. Maggioni, 2021: Assimilating satellite-based soil moisture observations in a land surface model: The effect of spatial resolution. *Journal of Hydrology X*, **13**, 100105, https://doi.org/10.1016/j.hydroa.2021.100105.

Sakaguchi, K., and Coauthors, 2022: Determining Spatial Scales of Soil Moisture—Cloud Coupling Pathways Using Semi-Idealized Simulations. *Journal of Geophysical Research: Atmospheres*, **127**, e2021JD035282, https://doi.org/10.1029/2021JD035282.

Santanello, J. A., C. D. Peters-Lidard, S. V. Kumar, C. Alonge, and W.-K. Tao, 2009: A Modeling and Observational Framework for Diagnosing Local Land–Atmosphere Coupling on Diurnal Time Scales. *Journal of Hydrometeorology*, **10**, 577–599, https://doi.org/10.1175/2009JHM1066.1.

Schaefer, G. L., M. H. Cosh, and T. J. Jackson, 2007: The USDA Natural Resources Conservation Service Soil Climate Analysis Network (SCAN). https://doi.org/10.1175/2007JTECHA930.1.

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

Sisterson, D. L., R. A. Peppler, T. S. Cress, P. J. Lamb, and D. D. Turner, 2016: The ARM Southern Great Plains (SGP) Site. *Meteorological Monographs*, **57**, 6.1-6.14, https://doi.org/10.1175/AMSMONOGRAPHS-D-16-0004.1.

Tai, S.-L., Yang, Z., Gaudet, B., Sakaguchi, K., Berg, L., Kaul, C. M., Qian, Y., Liu, Y., & Fast, J. (2024). A 1 km soil moisture data over eastern continental U.S. generated through assimilating SMAP data into the Noah-MP land surface model [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14370563

Tao, C., Y. Zhang, S. Tang, Q. Tang, H.-Y. Ma, S. Xie, and M. Zhang, 2019: Regional Moisture Budget and Land-Atmosphere Coupling Over the U.S. Southern Great Plains Inferred From the ARM Long-Term Observations. *Journal of Geophysical Research: Atmospheres*, **124**, 10091–10108, https://doi.org/10.1029/2019JD030585.

Taylor, C. M., A. Gounou, F. Guichard, P. P. Harris, R. J. Ellis, F. Couvreux, and M. De Kauwe, 2011: Frequency of Sahelian storm initiation enhanced over mesoscale soil-moisture patterns. *Nature Geosci*, **4**, 430–433, https://doi.org/10.1038/ngeo1173.

Torres, R., and Coauthors, 2012: GMES Sentinel-1 mission. *Remote Sensing of Environment*, **120**, 9–24, https://doi.org/10.1016/j.rse.2011.05.028.

Verseghy, D. L., 1991: Class—A Canadian land surface scheme for GCMS. I. Soil model. *International Journal of Climatology*, **11**, 111–133, https://doi.org/10.1002/joc.3370110202.

Wagner, T. J., D. D. Turner, L. K. Berg, and S. K. Krueger, 2013: Ground-Based Remote Retrievals of Cumulus Entrainment Rates. *J. Atmos. Oceanic Technol.*, **30**, 1460–1471, https://doi.org/10.1175/JTECH-D-12-00187.1.

Wang, C., and Coauthors, 2023: Chinese Soil Moisture Observation Network and Time Series Data Set for High Resolution Satellite Applications. *Scientific Data*, **10**, 424, https://doi.org/10.1038/s41597-023-02234-8.

Xia, Y., and Coauthors, 2012: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research: Atmospheres*, **117**, https://doi.org/10.1029/2011JD016048.

Yang, Z., and Coauthors, 2020: Understanding irrigation impacts on low-level jets over the Great Plains. *Clim Dyn*, **55**, 925–943, https://doi.org/10.1007/s00382-020-05301-7.

——, and Coauthors, 2021: Impact of Lateral Flow on Surface Water and Energy Budgets Over the Southern Great Plains—A Modeling Study. *Journal of Geophysical Research: Atmospheres*, **126**, e2020JD033659, https://doi.org/10.1029/2020JD033659.

Yang, Z.-L., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins. *Journal of Geophysical Research: Atmospheres*, **116**, https://doi.org/10.1029/2010JD015140.

Yin, J., and X. Zhan, 2020: Scale Impact of Soil Moisture Observations to Noah-MP Land Surface Model Simulations. *Remote Sensing*, **12**, 1169, https://doi.org/10.3390/rs12071169.

Zeng, X., 2001: Global Vegetation Root Distribution for Land Modeling.

Zhang, Y., and Coauthors, 2023: Generation of global 1&thinsp;km daily soil moisture product from 2000 to 2020 using ensemble learning. *Earth System Science Data*, **15**, 2055–2079, https://doi.org/10.5194/essd-15-2055-2023.

Lin, Y. and Mitchell, K. E.: The NCEP stage II/IV hourly precipitation analyses: development and applications, in: 19th Conf. Hydrology, available at: https://ams.confex.com/ams/pdfpapers/83847.pdf (August 2023), 2005.

Atmospheric Radiation Measurement (ARM) user facility. 1998. Oklahoma Mesonet Soil Moisture (OKMSOIL). 1998-01-01 to 2020-10-22, Southern Great Plains (SGP) External Data (satellites and others) (X1). Compiled by S. Giangrande, D. Wang and L. Gregory. ARM Data Center. Data set accessed 2022-10-30.