scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Global daily 9 km remotely sensed soil moisture (2015–2025) with microwave radiative transfer-guided learning

Sijia Feng[1,14], Aoyang Li[2,3,4,5,14], Rui Zhou[2,3,4,5], Klaus Butterbach-Bahl[1,6], Kaiyu Guan[2,3,4,5], Zhenong Jin[7], Majken C. Looms[8], Sherrie Wang[9,10], Christian Igel[11,12], Claire Treat[1], Jørgen Eivind Olesen[13] & Sheng Wang[1,2] ✉

Accurate estimation of surface soil moisture (SM) in terrestrial ecosystems is essential for understanding hydroclimate dynamics. The L-band Soil Moisture Active Passive (SMAP) mission provides 9-km global daily surface SM by using a microwave radiative transfer model (RTM)-based algorithm. However, the accuracy of SMAP SM is limited in regions with dense vegetation cover and complex surface conditions, due to the empirical parameterization and oversimplified radiative transfer processes. To overcome the limitations, we developed a Process-Guided Machine Learning (PGML) framework to integrate RTM theories and deep learning to predict global daily surface 9-km SM from April 2015 to June 2025. Informed by domain knowledge, we developed the PGML model structure using RTM and hydrological theories, designed a Kling-Gupta efficiency-based cost function, pretrained it with RTM simulations, and fine-tuned it with *in-situ* measurements. The independent validation shows that PGML SM has strong agreement with *in-situ* measurements ($R = 0.868$ and unbiased RMSE $= 0.054$ m^3/m^3). This study highlights the potential of PGML to enhance the accuracy of satellite SM, thereby supporting improved water resources and ecosystem management.

Background and Summary

Surface soil moisture (SM) is a critical hydroclimatic variable that influences the global water, carbon, and energy cycles[1,2]. It plays a crucial role in regulating evapotranspiration, modulating plant water stress, and governing interactions between the land and the atmosphere[3]. Furthermore, SM is a key indicator for assessing ecosystem productivity, forecasting weather and climate features and extremes, and supporting early warning systems for natural hazards[4–6]. Therefore, accurate spatial and temporal SM information is vital for these applications and

[1]Pioneer Center Land-CRAFT, Department of Agroecology, Aarhus University, Aarhus, 8000, Denmark. [2]Agroecosystem Sustainability Center, Institute for Sustainability, Energy, and Environment, University of Illinois Urbana-Champaign, Urbana, IL, 61801, USA. [3]Department of Natural Resources and Environmental Sciences, College of Agricultural, Consumers, and Environmental Sciences, University of Illinois Urbana-Champaign, Urbana, IL, 61801, USA. [4]National Center for Supercomputing Applications, University of Illinois Urbana-Champaign, Urbana, IL, 61801, USA. [5]Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL, 61801, USA. [6]Karlsruhe Institute of Technology, Institute for Meteorology and Climate Research, Atmospheric Environmental Research (IMK-IFU), Kreuzeckbahnstrasse 19, Garmisch-Partenkirchen, 82467, Germany. [7]Department of Bioproducts and Biosystems Engineering, University of Minnesota, Saint Paul, MN, 55108, USA. [8]Department of Geosciences and Natural Resource Management (IGN), University of Copenhagen, Copenhagen, 1165, Denmark. [9]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA. [10]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA. [11]Department of Computer Science, University of Copenhagen, Copenhagen, 2100, Denmark. [12]Pioneer Centre for Artificial Intelligence, Copenhagen, 1350, Denmark. [13]Department of Agroecology, Aarhus University, Blichers Allé 20, Tjele, 8830, Denmark. [14]These authors contributed equally: Sijia Feng, Aoyang Li. ✉e-mail: shengwang12@gmail.com

for advancing our understanding of ecohydrological processes. Despite its importance, capturing the global spatiotemporal variability of SM through frequent observations remains a major challenge.

Satellite remote sensing provides continuous, cost-effective, and global observations, and has been widely used in SM estimation. Microwave remote sensing is a powerful and reliable approach due to its ability to penetrate clouds and operate under all weather conditions. Among microwave techniques, passive microwave sensing at L-band (1.41 GHz) is especially effective for retrieving SM from the top ~5 cm of soil[3]. Currently, the Soil Moisture Active Passive (SMAP)[7] mission from the National Aeronautics and Space Administration (NASA) and the Soil Moisture and Ocean Salinity (SMOS)[8] mission from the European Space Agency (ESA) are specifically designed to provide global daily surface SM products at L-band.

Current SM products are typically derived from microwave radiative transfer models (RTMs), enabling consistent and physically based SM retrievals across diverse landscapes. The strength of RTMs lies in their solid physical foundation: they explicitly simulate microwave scattering and absorption by vegetation and soil, allowing them adapt to varying land cover conditions[9,10]. Building on these strengths, several enhanced RTM-based algorithms have been developed for L-band passive microwave observations, including the multi-temporal dual-channel algorithm that assumes vegetation remains unchanged within around a week[11], the multi-channel collaborative algorithm that assumes vegetation optical depth varies with polarization while SM is independent of polarization[12], and the mono-angle L-band microwave emission of the biosphere model where vegetation parameters are initialized from optical observations and SM is retrieved independently[13]. Despite these advancements, RTM-based SM products still exhibit significant uncertainties, particularly in areas with dense vegetation, actively growing crops and complex land surface with strong spatial heterogeneity[14,15]. These uncertainties are primarily attributed to a combination of fundamental physical constraints and inherent model limitations. Specifically, the physical constraints include the limited penetration capability of microwave signals through vegetation canopies and confounding signal scattering introduced by terrain, land management and rainfall events[16]. The model limitations come from the empirically derived vegetation parameters[17–19] and surface roughness representations[16,20–22], as well as oversimplified microwave radiative transfer processes (e.g., neglecting multiple scattering between the surface and vegetation and assuming the canopy is a homogeneous medium)[10].

To overcome the limitations of RTM-based algorithms, data-driven machine learning (ML) approaches have increasingly been adopted to leverage microwave observations and hydroclimate datasets for SM estimations[23–25]. A range of ML models have been applied, including random forest[24], neural networks[23,25], and long-short-term memory[26] (LSTM). For example, Lei *et al.*[24] used random forest regression to estimate global daily SM from SMAP observations with an unbiased root mean square error (ubRMSE) of 0.05 m^3/m^3. Similarly, Gao *et al.*[23] developed a deep neural network to improve SM estimates by incorporating SMAP observations alongside multi-source land surface features (surface temperature, soil texture, etc.), achieving an R of 0.697 and a ubRMSE of 0.057 m^3/m^3. While these ML methods show great promise for improving the accuracy of global SM estimates, they still suffer from the lack of embedded physical knowledge and limited generalizability in data-scarce regions[24,27]. These shortcomings may reduce model robustness and transferability when applied across diverse environmental conditions.

Process-guided machine learning (PGML)—also known as knowledge-guided or physics-informed machine learning—offers a promising pathway to bridge the gap between process-based physical models and data-driven approaches[28,29]. PGML integrates ML with domain knowledge through various strategies, such as designing model architecture based on domain knowledge, developing physics-guided loss functions, and training models using simulation data from physical models[30,31]. These strategies enhance the flexibility and modularity of PGML, enabling it to effectively assimilate multi-source data, including satellite observations, RTM simulations, hydroclimatic data, and field measurements[32]. PGML has been successfully applied to simulate key ecosystem states and dynamics from biogeochemical processes, such as soil organic carbon changes[29] and plant biomass in croplands[33]. In the context of SM estimation, PGML frameworks have been used to fuse multi-source hydroclimate variables with *in-situ* measurements, and to incorporate physical constraints into model loss functions[34–37]. Wang *et al.*[34] incorporated a physical constraint into a LSTM model by applying the water transport mechanism in the unsaturated zone. Based on the Richardson-Richards equation, Zhang *et al.*[37] developed a physical constraint loss function for the convolutional neural networks and LSTM model to predict multi-layer SM in eastern China. While these approaches have improved the accuracy of SM estimates, many of them treat domain knowledge as an external constraint, rather than fully embedding physical mechanisms within the model structure, thereby limiting their ability to capture complex physical relationships. Accordingly, further development of PGML approaches that tightly couple radiative transfer principles, *in-situ* SM measurements, and multi-source satellite and model data is essential for enhancing the accuracy and generalizability of global SM estimation.

To produce accurate global surface SM estimates, this study developed a microwave radiative transfer process-guided machine learning framework that integrates microwave RTM with deep learning. Specifically, guided by microwave radiative transfer processes, we developed an ML model for SM estimation using remote sensing observations, hydroclimate features and *in-situ* SM measurements. The RTM physics and ML model are deeply embedded in mechanisms cooperating and external constraints, including simulating pre-trained datasets, informing the selection of input features, and guiding the design of model structure and loss functions. Based on this framework, a global daily surface SM dataset at 9 km spatial resolution from April 2015 to June 2025 was produced.

Methods

Datasets and pre-processing. We developed the PGML model using multi-source remote sensing observations and hydroclimate datasets (Table 1). All the datasets used in PGML were projected to the Equal-Area Scalable Earth Grid 2.0 (EPSG: 6933) with 9-km spatial resolution, consistent with the spatial reference of SMAP observations.

Variable	Source	Spatiotemporal resolution	Unit	Range
Brightness Temperature (Tb)	SPL3SMP_E[38]	9 km, daily	K	0–330
Normalized Difference Vegetation Index (NDVI)	MOD13C1[42]	0.05°, 16-day	/	0–1
Clay fraction (Cf)	Soilgrid250[47]	250 m	/	0–0.8
Soil bulk density (Bd)	Soilgrid250[47]	250 m	g/cm^3	0–1.8
Land cover types (Lc)	MCD12C1[46]	0.05°, yearly	/	/
Precipitation (Prep)	ERA5-Land[45]	0.1°, daily	mm	0–250
Evaporation (Evap)	ERA5-Land[45]	0.1°, daily	mm	0–15
Land Surface Temperature (LST)	ERA5-Land[45]	0.1°, daily	K	200–330
Köppen-Geiger climate zone (Climate)	Beck. *et al.*[48]	1 km	/	/
In-situ Soil Moisture (*In-situ* SM)	ISMN[49], Fluxnet[50–54] and published papers[55–65]	Daily	m^3/m^3	0-1

Table 1. Details of data used in microwave-based radiative transfer model (RTM) and RTM process-guided machine learning (PGML). Note: The ranges represent valid land-surface values.

Passive microwave observations. The L-band SMAP mission provides vertically (V) and horizontally (H) polarized surface brightness temperature (Tb) at 40°, with overpasses at 6 AM (descending) and overpasses at 6 PM (ascending) since April 2015[7]. We collected the Tb observations at both V- and H-polarization from SMAP Level 3 Enhanced Soil Moisture version 6[38] (SPL3SMP_E, https://nsidc.org/data/spl3smp_e/versions/6) from April 2015 to June 2025. Due to the thermal consistency between the vegetation canopy and bare soil surface in the early morning[39], only Tb observations at 6 AM (local time) were collected to estimate SM.

Vegetation water content. We estimated the Vegetation Water Content (VWC) using an empirical equation of NDVI with land cover-based parameters[40].

$$VWC = (1.9134 \times NDVI^2 - 0.3215 \times NDVI) + sf \times (NDVI_{\max} - 0.1)/(1 - 0.1) \quad (1)$$

Here, the first part with NDVI describes the canopy water content, the second part with *sf* estimates the vegetation stem water content. $NDVI_{\max}$ is the maximum during the whole study period, *sf* is the static parameter for vegetation stem fractions collected from the land-cover-based lookup table in the SMAP DCA document[41]. Note that the daily NDVI is used instead of $NDVI_{\max}$ for croplands and grasslands[40].

As the dynamic NDVI can capture the surface vegetation variation better than climatological NDVI and further improve SM retrieval accuracy[19], the daily NDVI composited by 0.05° 16-day NDVI from MOD13C1[42] (https://www.earthdata.nasa.gov/data/catalog/lpcloud-mod13c1-061) was used for estimating VWC. The pre-processing steps include quality control (remove poor data with quality flag > 1), reprojection, resampling to 9-km spatial resolution, linear interpolation, Savitzky-Golay filtering[43] with a smoothing window size of 9 and a 6th-order polynomial[44], and removing the instances where more than 2 sets of 16-day NDVI composites are missing consecutively.

Surface temperature, precipitation and evaporation. The surface temperature, precipitation and evaporation data were acquired from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5-Land hourly dataset[45] (https://cds.climate.copernicus.eu/doi/10.24381/cds.e2161bac). These variables provide global spatial-temporal continuous coverage, with hourly data at a spatial resolution of 0.1° (~10 km). Daily surface temperature was collected from the average hourly surface temperature of each day. Daily precipitation and evaporation were acquired from the total precipitation and evaporation at 12 PM each day. Then, these variables were reprojected and resampled to the 9-km EASE-Grid 2.0.

Ancillary data. Ancillary data for SM estimation include land cover types, soil texture, and climate types. The land cover classification map based on the International Geosphere-Biosphere Program (IGBP) was extracted from the MCD12C1[46] (https://www.earthdata.nasa.gov/data/catalog/lpcloud-mcd12c1-061, Figure S1a and Table S1). The soil bulk density and clay fraction were collected from the SoilGrid250[47]. The Köppen-Geiger climate classification map was produced by Beck *et al.*[48], with five major climate zones including tropical, arid, temperate, cold, and polar regions (Figure S1b). All the ancillary datasets were processed into the EASE-Grid 2.0 projection with 9-km spatial resolution.

***In-situ* soil moisture measurements.** We collected global *in-situ* SM measurements from the International Soil Moisture Network[49] (ISMN, https://ismn.earth/en/), AmeriFlux[50] (https://ameriflux.lbl.gov/resources/logos-acknowledgments/), Integrated Carbon Observation System[51–53] (ICOS, https://www.icos-cp.eu/), JapanFlux[54] (https://ads.nipr.ac.jp/japan-flux2024/) and previous related studies[55–65] during April 2015 to June 2025. The following four processing steps were taken to obtain reliable daily SM measurements. (*i*) To consider the depth of passive microwave penetration through the land surface, only the top 5 cm measurements were used in this study. (*ii*) All sites were filtered through strict quality control approaches, including removing missing, erroneous and anomalous values based on the quality flag (quality flag≠G). (*iii*) To ensure a consistent temporal resolution across multi-source datasets and maximize data availability, daily SM was extracted directly or calculated by averaging hourly SM measurements (at least 18 valid measurements per day). (*iv*) To reduce the

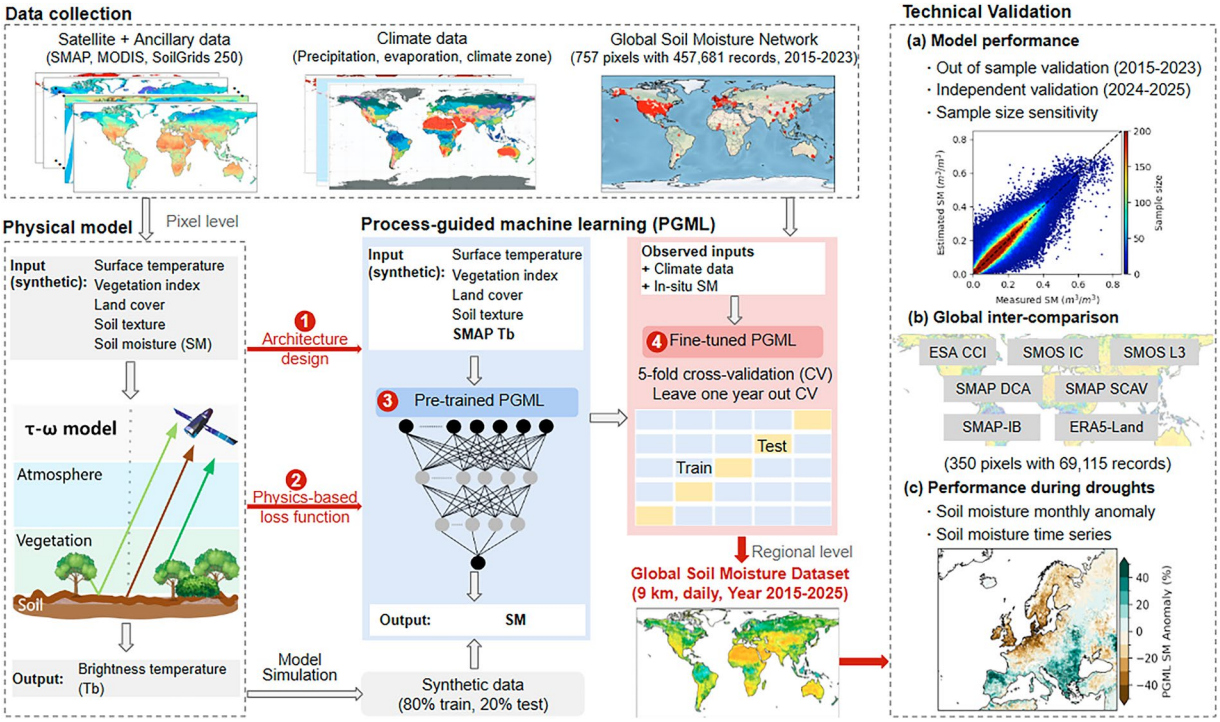

Fig. 1 Overview of the framework used for PGML development and assessment.

uncertainty from the spatial mismatch, *in-situ* measurements of sites located within one pixel were averaged arithmetically, and pixels with at least 30 valid records were retained. It should be noted that daily SM aggregation was adopted to ensure temporal consistency across predictors while maintaining compatibility with satellite observations acquired at different overpass times. Finally, a total of 757 pixels from 1,197 sites were included in this study. The *in-situ* data were used for (*i*) PGML five-fold cross-validation (2015–2023, 757 pixels with 457,681 records), (*ii*) temporally independent validation using data at the same pixels from different years (2024–2025, 469 pixels with 58,923 records), and (*iii*) SM products intercomparison (2015-2023, 350 sites with 69,115 records).

Process-guided machine learning to estimate soil moisture. This PGML employs the multi-layer perceptron (MLP) structure to integrate relationships between various hydroclimate features and radiative transfer processes with radiative transfer physics from RTM[10] for guidance (Fig. 1). Specifically, developing PGML for SM estimation can be divided into 4 steps. (*i*) Developing the architecture for an MLP based on the causal relations derived from RTM. (*ii*) Designing the physics-guided cost function for both pre-training and fine-tuning processes. (*iii*) Pre-training PGML using synthetic data generated from RTM. (*iv*) fine-tuning PGML using *in-situ* measurements, remote sensing data, and climate data of collected pixels from April 2015 to December 2023. After these steps, we applied fine-tuned PGML to estimate global daily SM at 9-km spatial resolution from April 2015 to June 2025.

Structure of PGML. The PGML was developed based on the MLP and incorporated physical knowledge through feature selection and model training strategy. We applied two steps to train PGML (pre-training and fine-tuning), resulting in a model structure with flexible input dimensionality. For pre-training, PGML consists of an MLP with an input of 22 features. Among these features, the Lc (Land cover types) is categorical and was encoded using one-hot encoding into 16 columns, resulting in a total of 22 input features. Its network includes 4 fully connected (linear) layers of sizes 256, 128, 64, and 32, each followed by a ReLU nonlinearity activation, and a final linear output layer to estimate surface SM. The 32-dimensional representation obtained from this stage serves as a latent embedding learned from RTM-informed variables. For fine-tuning, 3 additional climatic variables (precipitation, evaporation and climate zones) were incorporated. The climatic zone follows the first letter of the Köppen-Geiger classification (5 classes, Fig. S1b) and was encoded using one-hot encoding. These features (precipitation, evaporation and 5 climatic zones) were concatenated with the 32-dimensional embedding produced during pre-training. The fine-tuning head processes this $32+7$ dimensions input with fully connected linear layers of sizes 256, 128, 64, 32, 16, and 8, each followed by a ReLU activation, and the final linear layer followed by sigmoid activation. The sigmoid layer serves as a physical constraint to ensure that the predicted SM values remain within the physically valid range of 0–1 m^3/m^3. This design preserves the RTM-based knowledge from pre-training while allowing the model to adapt to the expanded feature space introduced in fine-tuning.

The physical law-based cost function. To further enhance physical consistency, the Kling-Gupta efficiency (KGE)[66] was used to consider the correlation coefficient, bias and standard deviation (STD) between

simulations and measurements. The KGE is widely used to assess models' ability to capture the temporal dynamics of the target variable and quantify errors[67,68]. We minimized the squared distance term in the original KGE formulation. The KGE-style cost function is defined below:

$$L_{KGE} = (R - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2 + \lambda \sum_i p(sim_i) \tag{2}$$

$$\alpha = \frac{STD_{sim}}{STD_{meas}} \tag{3}$$

$$\beta = \frac{\overline{sim}}{\overline{meas}} \tag{4}$$

$$p(sim_i) = \begin{cases} (0 - sim_i)^2 & sim_i < 0 \\ 0 & 0 \le sim_i \le 1 \\ (sim_i - 0)^2 & sim_i > 1 \end{cases} \tag{5}$$

Here *sim* is the simulation, *meas* is the measurements, R is the Pearson correlation coefficient, α is the term representing the variability of prediction errors, β is the bias, *STD* is the standard deviation of data, and $\overline{sim}$ is the average of the simulations, and $\overline{meas}$ is the averaged measurements. This cost function is equivalent to minimizing $(1 - KGE)^2$, while additionally incorporating a soft physical constraint through the penalty term $\lambda \sum_i p(sim_i)$. In L_{KGE}, the bias and variability terms (β, α) act as soft constraints to penalize the deviations from the target values (equal to 1) but are not enforced exactly. The additional boundary penalty $p(sim_i)$ discourages predictions outside the physically valid range (0 to 1), ensuring that SM estimates remain valid without introducing non-differentiable projections. Additionally, we deliberately avoided hard clipping or projection operations, as these can interrupt gradient flow. All statistics (R, means, and STD) are computed on the samples used in each training step. To avoid numerical instability when the variance is close to zero, small numerical epsilons are added to the denominators.

Pre-training PGML. For pre-training PGML, we applied the τ-ω model[10] to generate synthetic data. This model is applied in the satellite SM retrieval algorithms[41,69] to produce satellite SM products. Based on the radiative transfer processes, the τ-ω model simulates the upward radiation from the land surface through vegetation layers briefly and effectively[10]. We conducted 500,000 records from the τ-ω model by sampling inputs within predefined valid ranges (Table 1) randomly and uniformly, ensuring broad coverage of potential variable combinations. While for vegetation water content, the input ranges were further adjusted by land-cover type to better describe realistic surface conditions[40]. Specifically, the synthetic data contains Tb simulations with inputs of land cover type, surface temperature, NDVI, soil clay fraction, soil bulk density, and SM. Further, we pre-trained PGML using synthetic data to learn radiative transfer processes. The synthetic dataset was randomly split into two parts, 80% of the whole dataset for training and 20% of it for testing.

Fine-tuning PGML. For fine-tuning PGML, we used global *in-situ* SM measurements with RTM inputs and climate data from April 2015 to December 2023, with a total of 457,681 pixel-level records. Then, we conducted a 5-fold cross-validation for the fine-tuning process. The input layer of the pre-trained PGML was extended to incorporate additional climate variables (precipitation, evaporation, and climate zone). After training, the 5 models' outputs were collected to obtain the ensemble averages and STD. In addition, we fine-tuned the PGML using leave-one-year-out cross-validation to assess the temporal generalizability of the model, as reported in Table S5 and Fig. S2.

Evaluation scheme. To obtain reliable evaluation results, we assessed PGML SM using both *in-situ* SM measurements and other SM products. The comparison against 7 global SM products, including 5 L-band SM products, 1 active-passive combined SM product and 1 model-driven SM product. Detailed information on these SM products is summarized in Table 2.

The European Space Agency Climate Change Initiative SM[70] (ESA CCI SM, https://data.ceda.ac.uk/neodc/esacci/soil_moisture/data/daily_files) is generated by combining multi-sensors from both passive and active microwave missions. The SMOS Level 3 SM[69] (SMOS L3 SM, https://www.catds.fr/Products/Products-over-Land) and SMOS INRA-CESBIO SM[71] (SMOS-IC SM, https://ib.remote-sensing.inrae.fr/index.php/smos-ic-v2-product-documentation/) are produced using the multi-orbit algorithm applied to SMOS observations. Based on RTM, the SMAP team produces the SM products[38] (https://nsidc.org/data/spl3smp_e/versions/6) through the Dual Channel Algorithm (DCA) and Single Channel Algorithm at V-polarization (SCAV). The SMAP-INRAE-BORDEAUX SM[13] (SMAP-IB SM, https://ib.remote-sensing.inrae.fr/index.php/smap-ib-product-documentation/) is retrieved through the L-band microwave emission of the biosphere model. ERA5-Land SM[45] (https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=overview) is a reanalysis dataset produced by replaying the land component of the ECMWF ERA5 dataset.

To assess SM products, we implemented strict quality control for all SM products, including excluding permanent water, snow- and ice-covered regions, urban, and radio frequency interferences (RFI)-contaminated

SM product	Source	Temporal resolution	Spatial resolution	Available period (up to October 2025)
ESA CCI	ESA CCI V09.1[70]	Daily	0.25 °	April 2015–December 2023
SMOS-IC	SMOS-IC V2[71]	Daily	25 km	April 2015–August 2024
SMOS L3	SMOS L3 V700[69]	Daily	25 km	April 2015–December 2024
SMAP DCA	SPL3SMP V6[38]	Daily	9 km	April 2015–October 2025
SMAP SCAV	SPL3SMP V6[38]	Daily	9 km	April 2015–October 2025
SMAP-IB	SMAP-IB V2[13]	Daily	36 km	April 2015–August 2024
ERA5-Land	ERA5-Land[45]	Daily	0.1 °	April 2015–October 2025

Table 2. Soil moisture products for inter-comparison.

Bit	Information	Value and interpretation
0	Static water / urban area	1: Covered by water, urban or ice/snow 0: Otherwise
1	Frozen ground	1: Surface temperature <273.15 K 0: Otherwise
2	Dense vegetation	1: Vegetation water content >5 kg/m^2 0: Otherwise
3	Valid range	1: Soil moisture <0 or $>$ porosity 0: Otherwise
4–7	Undefined	0 (not used)

Table 3. Quality flag (8 bit) for PGML SM.

areas. For inter-comparison, all SM products were kept at original spatial resolutions, and their SM values were extracted at the 9-km pixel of *in-situ* average measurements to maintain consistent comparison across datasets. The metrics for evaluation included R ($p < 0.05$), bias, RMSE, and ubRMSE.

Data Records

The PGML SM[72] dataset can be accessed from Zenodo at https://doi.org/10.5281/zenodo.15826989.

The PGML SM[72] dataset contains global daily surface SM with a spatial resolution of 9 km, in the unit of m^3/m^3, from April 2015 to June 2025. This dataset is stored in Network Common Data Form (NetCDF) format with one file per day, defined by two dimensions (latitude and longitude), surface SM, ensemble STD, and a quality-control flag.

The file name is designed as "PGML_SM_yyyymmdd_Vx.nc", where "yyyy" stands for year, "mm" stands for month, "dd" stands for day, and 'Vx' stands for version ID. For example, "PGML_SM_20170101_V2.nc" contains the global surface SM distribution on the first day of January 2017 in version 2. Naming convention: PGML_SM_yyyymmdd_Vx.nc, Variable: lat_center, lon_center, pgml_sm, pgml_sm_std and qc (8 bit). The missing value is filled with NAN. A quality-control flag (qc) is assigned for SM estimation. The detailed information of qc is listed in Table 3. The PGML SM product is provided on the 9-km ESAE-Grid 2.0 (EPSG: 6933).

Technical Validation

Model performance. The PGML shows high accuracy, data efficiency and generalization ability in estimating SM (Fig. 2). The out-of-sample validation and independent validation demonstrate strong agreement between PGML SM and *in-situ* SM measurements at the pixel level (Fig. 2a,b), as indicated by the high correlation ($R > 0.85$) and low error statistic (RMSE <0.06 m^3/m^3). In addition, as the sample fraction increases from 0.05% to 70%, the accuracy of PGML SM improves progressively and plateaus near 40%, with R rising by around 0.4 and RMSE falling by more than 0.05 m^3/m^3 for both the pre-trained and fine-tuned PGML (Fig. 2c). The fine-tuned PGML shows higher accuracy than the pre-trained PGML at small sample sizes (e.g., 0.05% and 0.1%). Furthermore, the accuracy of PGML SM was improved progressively, where the pre-training strategy contributed most, with R increasing by 0.022 and RMSE reducing by 0.003 m^3/m^3 (Fig. 2d). The performance of the PGML model trained with the leave-one-year-out cross-validation strategy can be found in Table S5 and Fig. S2.

Inter-comparison of SM accuracy by using *in-situ* SM measurements. To further evaluate the PGML SM (five-model ensemble average), we compared it with seven widely used SM products (Table 2) against *in-situ* SM measurements at the pixel level, with a total of 69,115 records from April 2015 to December 2023. PGML SM outperforms other SM products, with $R = 0.923$, bias $= -0.001$ m^3/m^3, and ubRMSE $= 0.040$ m^3/m^3 (Fig. 3a). Compared with alternative SM products, PGML SM achieves improvements in accuracy, with R increasing by around 0.3 and RMSE reducing by more than 0.05 m^3/m^3. Moreover, PGML exhibits minimal bias in both dry (<0.2 m^3/m^3) and wet (>0.4 m^3/m^3) SM conditions, and its estimates present a lower dispersion. Other SM products exhibit various limitations. Although ESA CCI SM achieves favorable overall metrics, it overestimates SM under dry conditions (<0.2 m^3/m^3) and underestimates SM under wet conditions (>0.4 m^3/m^3) (Fig. 3b). SMOS-IC and SMOS L3 SM show poor performance in dry SM with large underestimations (Fig. 3c-d), which may be due to the uncertainties in model parameters and interferences of multi-angle observations[71]. As for

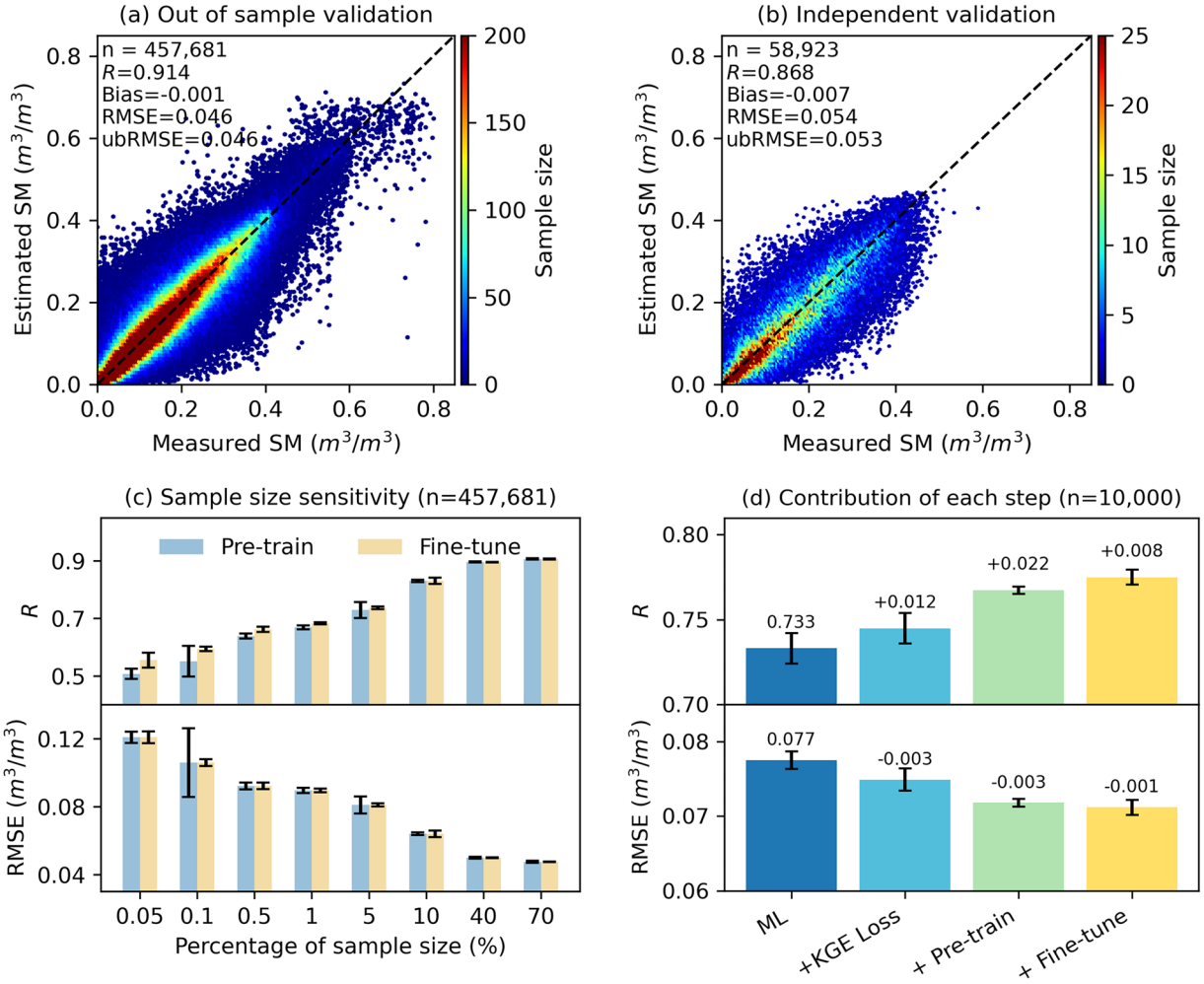

Fig. 2 Evaluation of PGML performance. (**a**) Out-of-sample validation against *in-situ* SM measurements from 2015 to 2023. (**b**) Independent validation against *in-situ* SM measurements from 2024 to 2025. (**c**) Sample size sensitivity of PGML in *R* and RMSE, with bars showing mean values and error bars indicating standard deviation. (**d**) Stepwise contribution of PGML, with bars showing mean values and error bars indicating standard deviation.

SMAP DCA SM and SCAV SM (Fig. 3e,f), the application of the Tb-based cost function and saturated SM range effectively constrains the values and enhances the accuracy[41]. SMAP-IB shows large uncertainty in the range of 0.1–0.3 m^3/m^3, where overestimate SM occasionally (Fig. 3g). The model-derived ERA5-Land SM tends to systematically overestimate SM under dry conditions, with the density contours shifting above the 1:1 line (Fig. 3h). Overall, the superior performance of PGML SM demonstrates model's strength in SM estimation, which can be attributed to its ability to learn both the mechanism of microwave radiative transfer processes and the intrinsic patterns described by *in-situ* SM measurements.

To further investigate the performance of PGML SM across different land covers, we selected the 5 most accurate SM products overall and reported their validation metrics (Fig. 4), and spatial distributions of pixel-level performance metrics were presented in Figs. S3–6.

Compared with other SM products, PGML SM demonstrates superior performance across validation metrics, with the most significant improvement observed in absolute bias (Fig. 4). PGML SM exhibits lower bias and better ability to capture SM dynamics and accurate values than others. For example, PGML substantially enhances the accuracy of SM estimates in North American forests (Fig. S4a), where dense vegetation limits microwave penetration. As for crop fields (CRL and CRM) across temperate regions of Europe and North America, PGML effectively corrects the systematic biases for most pixels. These improvements come from incorporating climate variables and *in-situ* SM measurements into the model framework, enabling the model to better capture temporal dynamics and management practices (e.g., irrigation and harvest). These improvements indicate that PGML can reduce systematic errors and capture robust SM dynamics across diverse land surface and climate regimes.

Inter-comparison of spatiotemporal characteristics. We conducted global comparisons of annual averaged SM and monthly SM dynamics across latitudes between PGML SM and the other seven SM products to assess their spatiotemporal characteristics.

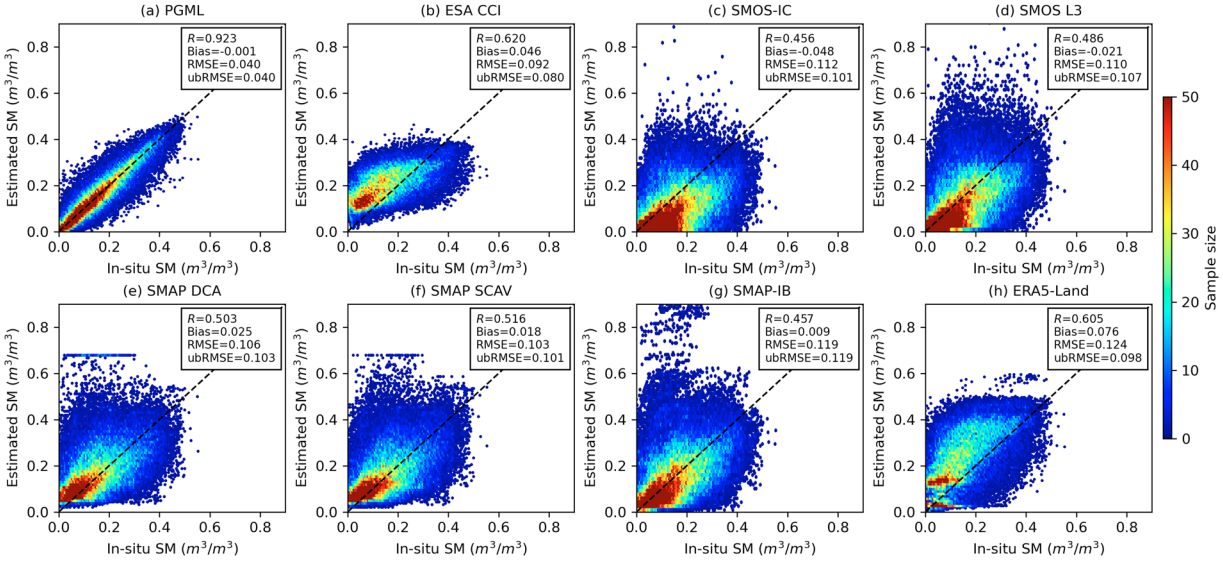

Fig. 3 Evaluation results of eight SM products against *in-situ* SM measurements with 69,115 records: (**a**) PGML, (**b**) ESA CCI, (**c**) SMOS-IC, (**d**) SMOS L3, (**e**) SMAP DCA, (**f**) SMAP SCAV, (**g**) SMAP-IB, and (**h**) ERA5-Land.

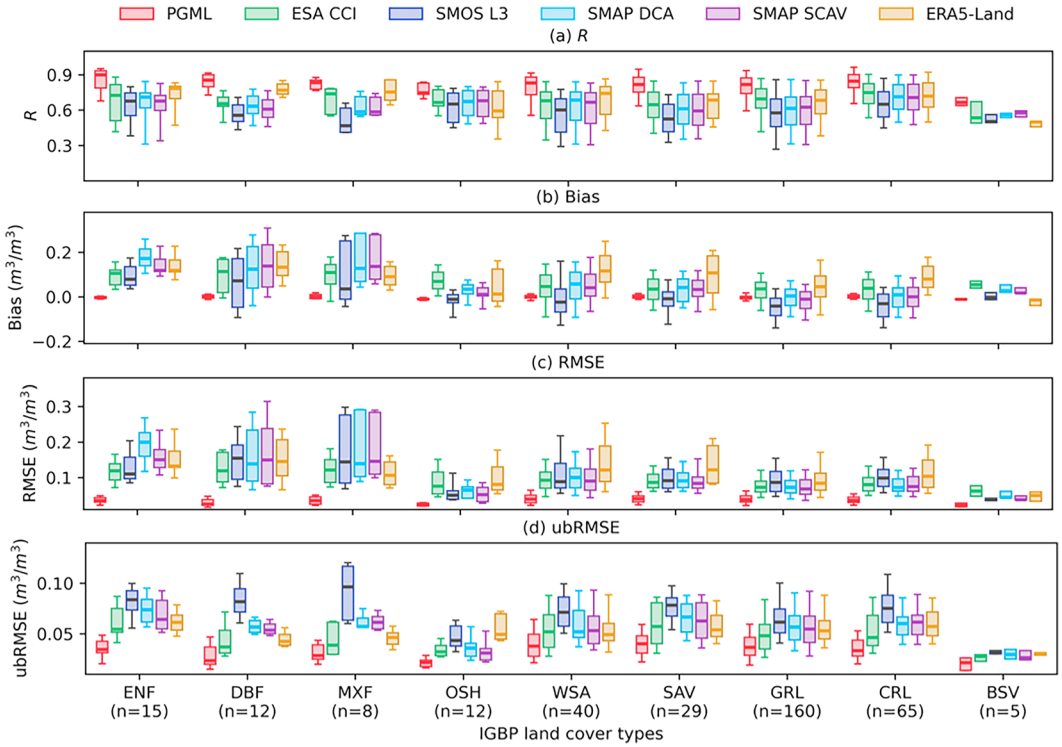

Fig. 4 Evaluation results of six SM products across different land cover types with 350 pixels: (**a**) *R*, (**b**) Bias, (**c**) RMSE and (**d**) ubRMSE. The horizontal line denotes the median. Only land cover types with more than 3 valid pixels were shown here.

PGML SM exhibits a similar spatial distribution to other SM products (Fig. 5). In mid-latitude regions where *in-situ* SM measurements are more abundant, PGML further demonstrates its strength by capturing finer spatial variability (Figs. S3–6). However, PGML tends to produce drier SM estimates in the tropics near the equator (Fig. 5a). This difference is mainly caused by the sparse availability and limited representativenessof *in-situ* SM measurements in this region. In regions where *in-situ* SM measurements are sparse, PGML relies heavily on the radiation transfer mechanism and training data from areas with similar hydroclimate features. As a result, PGML provides physically consistent estimations, although some regional discrepancies with other SM

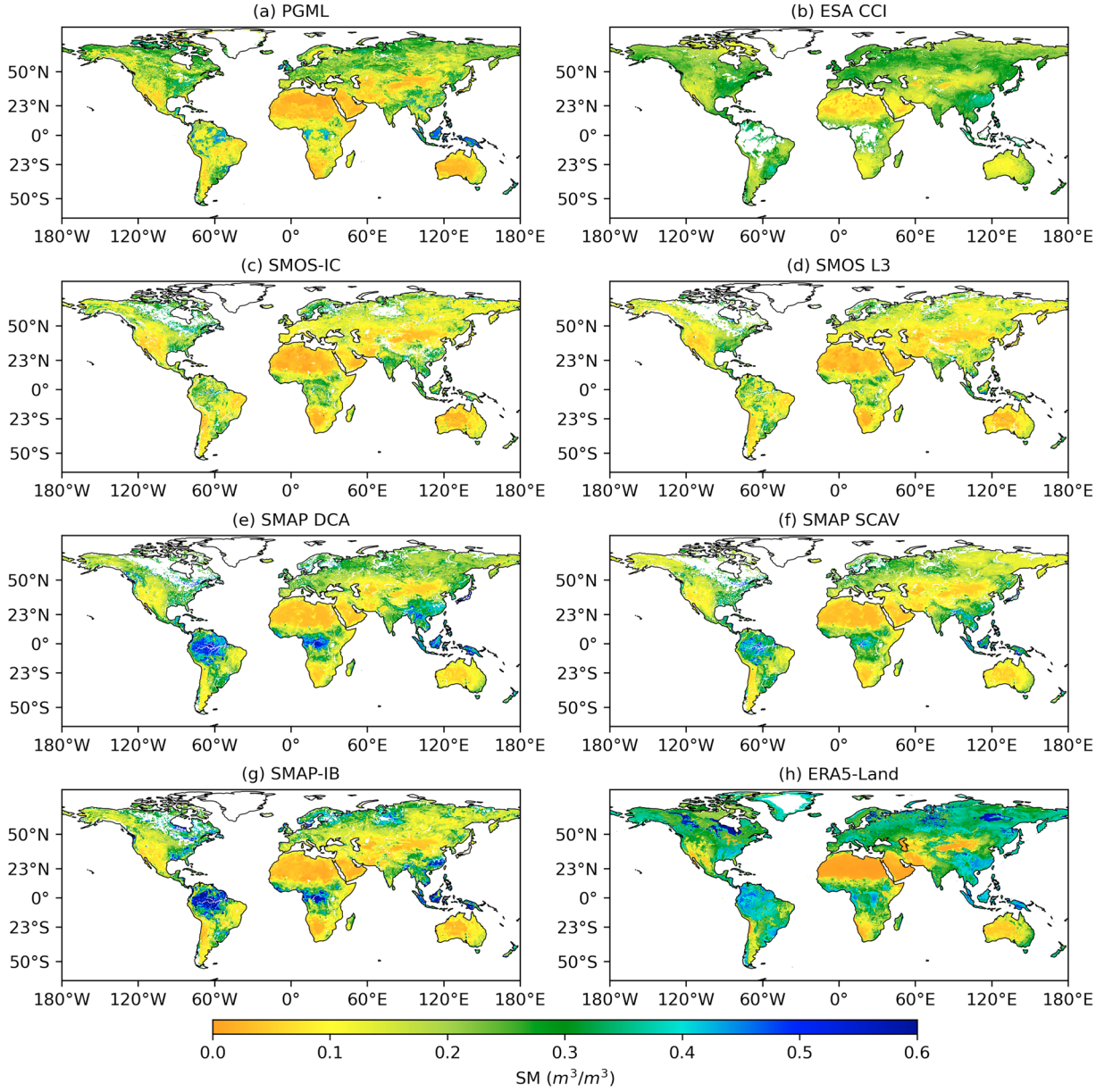

Fig. 5 Global patterns of annual averaged SM from April 2015 to December 2023: (**a**) PGML, (**b**) ESA CCI, (**c**) SMOS-IC, (**d**) SMOS L3, (**e**) SMAP DCA, (**f**) SMAP SCAV, (**g**) SMAP-IB, and (**h**) ERA5-Land.

products may remain. In contrast, conventional RTM-based SM products depend strongly on the assumed RTM structure and parameterization strategy. In dense forests, microwave signals are dominated by canopy emission rather than soil due to the limited penetration depth. This substantially reduces the sensitivity of microwave observations to SM, resulting in large uncertainty in SM estimations and often tends to exhibit a positive deviation[15,56]. Besides, PGML SM presents a higher STD (~0.05 m^3/m^3) over vegetation coverages than bare surface, reflecting the reduced sensitivity of remote sensing observations to SM under dense canopies. Importantly, regions with notable discrepancies lack sufficient ground-truth measurements for validation, which remains a challenge in assessing the accuracy of different SM products[73,74].

PGML SM exhibits similar spatiotemporal patterns to others and captures the meridional migration of the Intertropical Convergence Zone (Fig. 6). The primary difference among these SM products lies in the absolute magnitude of SM estimates. For SMAP-based SM (Fig. 6e–g), maximum values are observed near the equator, where persistently humid climate and abundant precipitation maintain high SM levels exceeding 0.4 m^3/m^3. PGML SM reaches approximately 0.25 m^3/m^3 at around 55°N during the growing season (April to October), consistent with the temporal dynamics presented in other SM products. Outside the growing season, the low temperature and presence of snow or ice can weaken the sensitivity of the passive microwave observations to SM variations. This uncertainty is further exacerbated by the limited availability of *in-situ* SM measurements for calibration in high latitudes during the cold season. Despite these challenges, PGML SM maintains a continuous

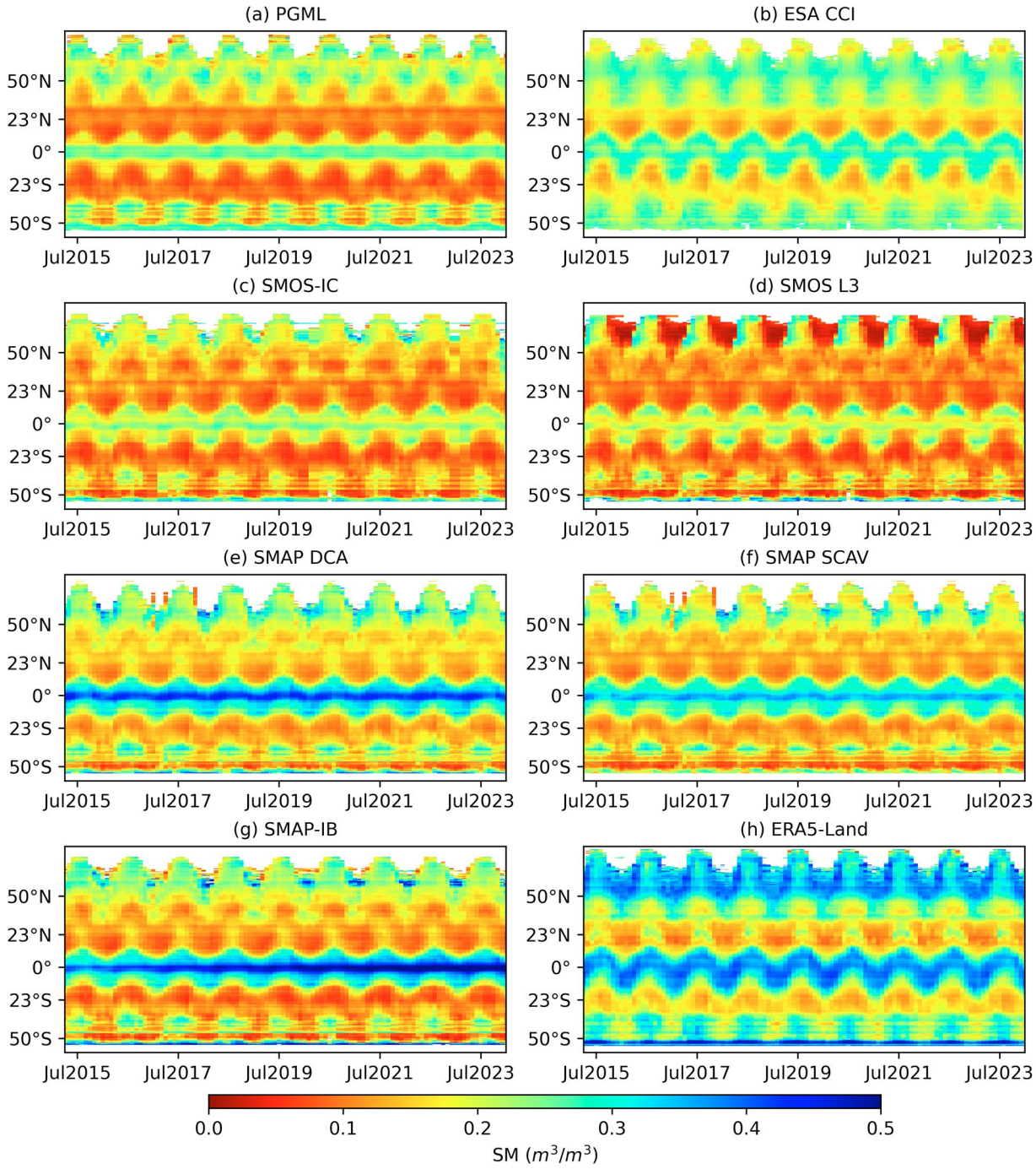

Fig. 6 Hovmöller diagrams[78] for latitudinal monthly SM from April 2015 to December 2023:(**a**) PGML, (**b**) ESA CCI, (**c**) SMOS-IC, (**d**) SMOS L3, (**e**) SMAP DCA, (**f**) SMAP SCAV, (**g**) SMAP-IB, and (**h**) ERA5-Land.

and coherent latitudinal pattern throughout the entire period, as also confirmed by evaluation results against *in-situ* SM measurements.

Evaluation of PGML SM during droughts. To investigate the capability of PGML SM during extreme droughts, the European drought[75,76] in 2018 was selected as a case study. Monthly SM anomalies in 2018 (Fig. 7a and Fig. S9) were calculated based on monthly PGML SM for the reference period from April 2015 to June 2025. ESA CCI SM was included for comparison due to its good performance (Figs. 3–4). Furthermore, we extracted three pixels with *in-situ* SM measurements located in drought-affected areas to demonstrate the temporal SM variations(Fig. 7d–f).

In July 2018, more than 30% of Europe experienced extreme drought, as indicated by negative SM anomalies (Fig. 7a). Particularly, regions such as northern Germany, Denmark, the United Kingdom and the Netherlands exhibited substantial SM deficits, driven by reduced precipitation and increased evapotranspiration associated with

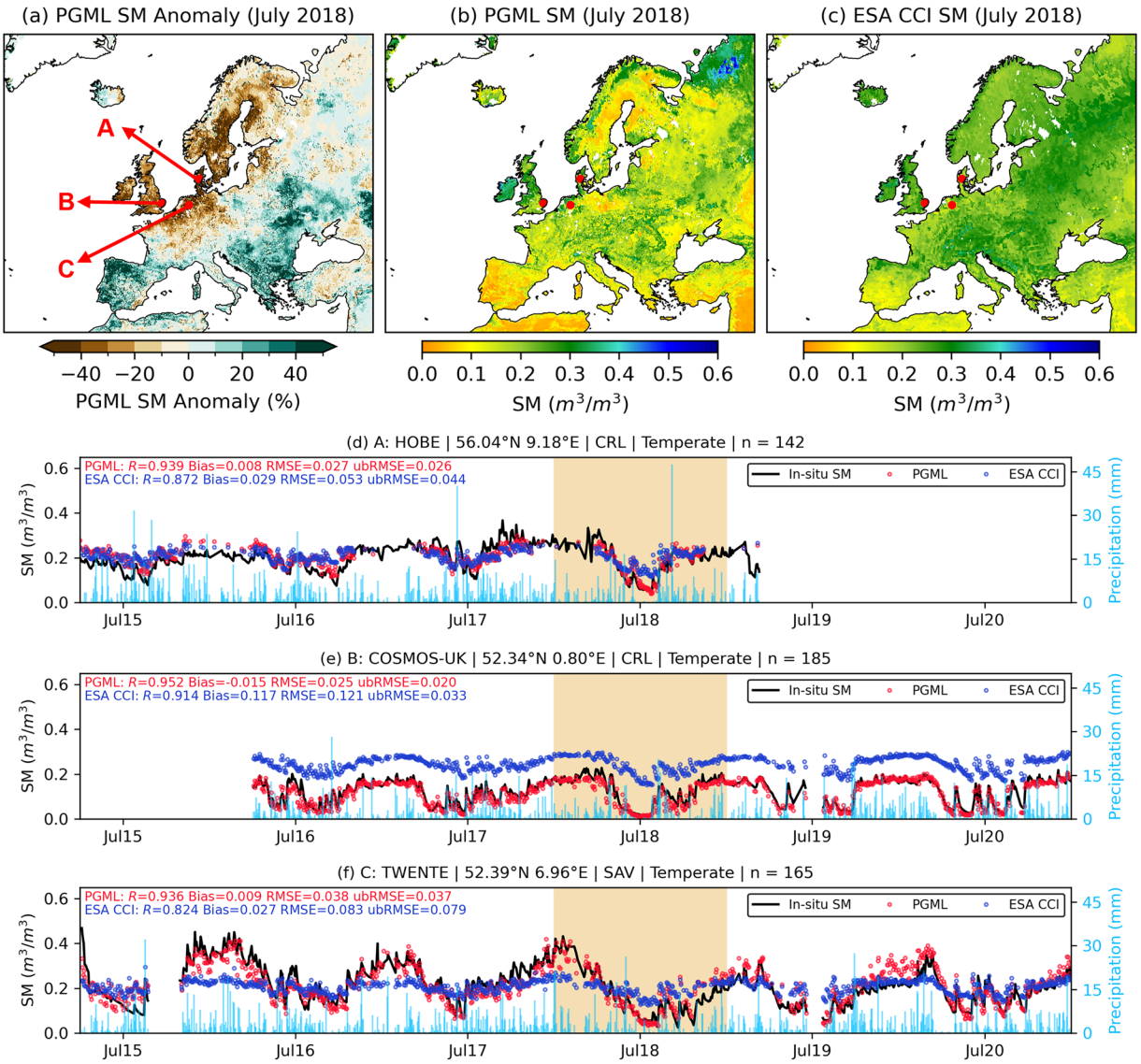

Fig. 7 The monthly PGML SM anomaly (**a**) and monthly averaged SM (**b,c**) for Europe in July 2018, and SM dynamics of selected pixels with precipitation (**d–f**). The validation metrics for 2018 (brown shadow) are reported in sub-figures (**d–f**).

high temperatures[77]. For example, in the northern Netherlands (Fig. 7f), SM declined to approximately 0.1 m^3/m^3 in July 2018 and remained about 20% below the multi-year July averaged SM. Both PGML SM and ESA CCI SM can capture the overall drought signal by decreasing, while PGML SM achieves higher R and lower systematic bias. Specifically, at the COSMOS-UK pixel (Fig. 7e), PGML and ESA CCI exhibit almost identical ability to capture temporal trends ($R = 0.952$ for PGML v.s. $R = 0.914$ for ESA CCI), whereas PGML shows a much smaller deviation (bias $= -0.015$ m^3/m^3) compared with ESA CCI (bias $= 0.117$ m^3/m^3). Overall, PGML outperforms ESA CCI in capturing SM dynamics during drought events, providing more accurate estimates and highlighting its potential for drought detection.

Usage Notes

In this study, we produced a global, daily, 9-km surface soil moisture (SM) dataset from April 2015 to June 2025 using a microwave radiative transfer-based physical-guided machine learning (PGML) method. This dataset provides a valuable complement to existing satellite SM products and enhances capability for applications in hydrology, climate research and agriculture from the regional to global scale.

In regions with limited *in-situ* SM networks, such as high latitudes and forests, the uncertainty of PGML SM is mainly driven by the lack of ground measurements to constrain the estimates. Although the PGML incorporates RTM-based knowledge, the scarcity of ground data limits local optimization and validation, making it difficult to improve accuracy and assess uncertainty. Users should be aware of potential spatial mismatches between site-scale and pixel-scale estimations. Although PGML demonstrates robust performance from site to pixel scales, fine-tuning with *in-situ* data might slightly influence the stability of the physical relationship

and mechanism, especially in heterogeneous regions. This may introduce uncertainty into applications such as dry-down rates, drought detection, and crop yield estimation. Therefore, we recommend using PGML SM together with other available satellite-based SM products in areas with complex surface conditions or sparse SM networks. Furthermore, since PGML uses remote sensing inputs, SM cannot be estimated when satellite observations are unavailable. These missing days are documented in the dataset release notes. Overall, this new PGML SM provides a valuable complement to existing SM products and enhances capability for large-scale hydroclimate analysis.

Data availability

The global soil moisture dataset[72] published in this study is available from Zenodo at https://doi.org/10.5281/zenodo.15826989. All external input datasets used in this research (e.g., SMAP brightness temperatures, ERA5-Land meteorological variables, MODIS NDVI) are publicly available from their original repositories, as cited in the manuscript.

Code availability

Data processing and analysis were conducted using Python version 3.13. The code is available on GitHub at https://github.com/SkyeFengg/PGML-SM.

Received: 10 July 2025; Accepted: 26 January 2026;
Published online: 12 February 2026

References

1. Oki, T. & Kanae, S. Global Hydrological Cycles and World Water Resources. *Science* **313**, 1068–1072 (2006).
2. Jung, M. *et al*. Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature* **467**, 951–954 (2010).
3. Calvet, J.-C. *et al*. Sensitivity of Passive Microwave Observations to Soil Moisture and Vegetation Water Content: L-Band to W-Band. *IEEE Trans. Geosci. Remote Sensing* **49**, 1190–1199 (2011).
4. Fu, Z. *et al*. Global critical soil moisture thresholds of plant water stress. *Nature Communications* **15**, 4826 (2024).
5. Liu, L. *et al*. Soil moisture dominates dryness stress on ecosystem production globally. *Nat Commun* **11**, 4892 (2020).
6. Saeedi, M., Sharafati, A., Brocca, L. & Tavakol, A. Estimating rainfall depth from satellite-based soil moisture data: A new algorithm by integrating SM2RAIN and the analytical net water flux models. *Journal of Hydrology* **610**, 127868 (2022).
7. Entekhabi, D. *et al*. The Soil Moisture Active Passive (SMAP) Mission. *Proc. IEEE* **98**, 704–716 (2010).
8. Kerr, Y. H. *et al*. The SMOS Mission: New Tool for Monitoring Key Elements ofthe Global Water Cycle. *Proc. IEEE* **98**, 666–687 (2010).
9. Ulaby, F. T. & Wilson, E. A. Microwave Attenuation Properties of Vegetation Canopies. *IEEE Transactions on Geoscience and Remote Sensing* **GE-23**, 746–753 (1985).
10. Mo, T., Choudhury, B. J., Schmugge, T. J., Wang, J. R. & Jackson, T. J. A model for microwave emission from vegetation-covered fields. *Journal of Geophysical Research: Oceans* **87**, 11229–11237 (1982).
11. Konings, A. G. *et al*. Vegetation optical depth and scattering albedo retrieval using time series of dual-polarized L-band radiometer observations. *Remote Sensing of Environment* **172**, 178–189 (2016).
12. Zhao, T. *et al*. Retrievals of soil moisture and vegetation optical depth using a multi-channel collaborative algorithm. *Remote Sensing of Environment* **257**, 112321 (2021).
13. Li, X. *et al*. A new SMAP soil moisture and vegetation optical depth product (SMAP-IB): Algorithm, assessment and inter-comparison. *Remote Sensing of Environment* **271**, 112921 (2022).
14. Colliander, A. *et al*. Comparison of high-resolution airborne soil moisture retrievals to SMAP soil moisture during the SMAP validation experiment 2016 (SMAPVEX16). *Remote Sensing of Environment* **227**, 137–150 (2019).
15. Colliander, A. *et al*. Validation of Soil Moisture Data Products from the NASA SMAP Mission. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* **15**, 364–392 (2022).
16. Walker, V. A. *et al*. From field observations to temporally dynamic soil surface roughness retrievals in the U.S. Corn Belt. *Remote Sensing of Environment* **287**, 113458 (2023).
17. Gao, L., Sadeghi, M., Ebtehaj, A. & Wigneron, J.-P. A temporal polarization ratio algorithm for calibration-free retrieval of soil moisture at L-band. *Remote Sensing of Environment* **249**, 112019 (2020).
18. Feng, S. *et al*. Improved estimation of vegetation water content and its impact on L-band soil moisture retrieval over cropland. *Journal of Hydrology* **617**, 129015 (2023).
19. Feng, S. *et al*. Can real-time NDVI observations better constrain SMAP soil moisture retrievals? *Remote Sensing of Environment* **318**, 114569 (2025).
20. Walker, V. A., Hornbuckle, B. K., Cosh, M. H. & Prueger, J. H. Seasonal Evaluation of SMAP Soil Moisture in the U.S. Corn Belt. *Remote Sensing* **11**, 2488 (2019).
21. Chaubell, M. J. *et al*. Improved SMAP Dual-Channel Algorithm for the Retrieval of Soil Moisture. *IEEE Trans. Geosci. Remote Sensing* **58**, 3894–3905 (2020).
22. Zeng, J., Chen, K.-S., Cui, C. & Bai, X. A Physically Based Soil Moisture Index From Passive Microwave Brightness Temperatures for Soil Moisture Variation Monitoring. *IEEE Trans. Geosci. Remote Sensing* **58**, 2782–2795 (2020).
23. Gao, L. *et al*. A deep neural network based SMAP soil moisture product. *Remote Sensing of Environment* **277**, 113059 (2022).
24. Lei, F. *et al*. Quasi-global machine learning-based soil moisture estimates at high spatio-temporal scales using CYGNSS and SMAP observations. *Remote Sensing of Environment* **276**, 113041 (2022).
25. Yao, P. *et al*. A global daily soil moisture dataset derived from Chinese FengYun Microwave Radiation Imager (MWRI)(2010–2019). *Sci Data* **10**, 133 (2023).
26. Ma, H. *et al*. Surface soil moisture from combined active and passive microwave observations: Integrating ASCAT and SMAP observations based on machine learning approaches. *Remote Sensing of Environment* **308**, 114197, https://doi.org/10.1016/j.rse.2024.114197 (2024).
27. Han, Q. *et al*. Global long term daily 1 km surface soil moisture dataset with physics informed machine learning. *Sci Data* **10**, 101 (2023).
28. Wang, S. *et al*. Airborne hyperspectral imaging of cover crops through radiative transfer process-guided machine learning. *Remote Sensing of Environment* **285**, 113386 (2023).
29. Liu, L. *et al*. Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems. *Nat Commun* **15**, 357 (2024).
30. Karniadakis, G. E. *et al*. Physics-informed machine learning. *Nat Rev Phys* **3**, 422–440 (2021).

31. Willard, J., Jia, X., Xu, S., Steinbach, M. & Kumar, V. Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. *ACM Comput. Surv.* **55**, 1–37 (2023).
32. Karpatne, A. *et al*. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Trans. Knowl. Data Eng.* **29**, 2318–2331 (2017).
33. Wang, S. *et al*. Airborne hyperspectral imaging of nitrogen deficiency on crop traits and yield of maize by machine learning and radiative transfer modeling. *International Journal of Applied Earth Observation and Geoinformation* **105**, 102617 (2021).
34. Wang, Y. *et al*. A Deep Learning Approach Based on Physical Constraints for Predicting Soil Moisture in Unsaturated Zones. *Water Resources Research* **59**, e2023WR035194 (2023).
35. Bagheri, A., Patrignani, A., Ghanbarian, B. & Pourkargar, D. B. A physics-informed machine learning approach to predict soil water content for agricultural decision-making. *2024 American Control Conference (ACC)* 2–7 (2024).
36. Li, Z. *et al*. Global multi-scale surface soil moisture retrieval coupling physical mechanisms and machine learning in the cloud environment. *Remote Sensing of Environment* **329**, 114928 (2025).
37. Zhang, T. *et al*. Multi-layer grid-scale soil moisture estimation using spatiotemporal deep learning methods with physical constraints. *Journal of Hydrology* **657**, 133086 (2025).
38. O'Neill, P. *et al*. SMAP Enhanced L3 Radiometer Global and Polar Grid Daily 9 km EASE-Grid Soil Moisture, Version 6. *NASA National Snow and Ice Data Center Distributed Active Archive Center* https://doi.org/10.5067/M20OXIZHY3RJ (2023).
39. Lv, S., Wen, J., Zeng, Y., Tian, H. & Su, Z. An improved two-layer algorithm for estimating effective soil temperature in microwave radiometry using *in situ* temperature and soil moisture measurements. *Remote Sensing of Environment* **152**, 356–363 (2014).
40. Chan, S., Bindlish, R., Hunt, R., Jackson, T. & Kimball, J. *Ancillary Data Report Vegetation Water Content*. Report No. JPL D-53061 (Jet Propulsion Laboratory California Institute of Technology, 2013).
41. O'Neill, P. *et al*. *Algorithm Theoretical Basis Document Level 2 & 3 Soil Moisture (Passive) Data Products*. Report No. JPL D-66480 (Jet Propulsion Laboratory California Institute of Technology, 2021).
42. Didan, K. MODIS/Terra Vegetation Indices 16-Day L3 Global 0.05Deg CMG V061. *NASA Land Processes Distributed Active Archive Center* https://doi.org/10.5067/MODIS/MOD13C1.061 (2021).
43. Savitzky, A. & Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **36**, 1627–1639 (1964).
44. Chen, J. *et al*. A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter. *Remote Sensing of Environment* **91**, 332–344 (2004).
45. Copernicus Climate Change Service. ERA5-Land hourly data from 1950 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*, https://doi.org/10.24381/CDS.E2161BAC (2019).
46. Friedl, M. & Sulla-Menashe, D. MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 0.05Deg CMG V061. *NASA Land Processes Distributed Active Archive Center*, https://doi.org/10.5067/MODIS/MCD12C1.061 (2022).
47. Hengl, T. *et al*. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE* **12**, e0169748 https://www.soilgrids.org/ (2017).
48. Beck, H. E. *et al*. High-resolution (1 km) Köppen-Geiger maps for 1901–2099 based on constrained CMIP6 projections. *figshare* https://doi.org/10.6084/m9.figshare.c.6395666.v1 (2023).
49. Dorigo, W. *et al*. The International Soil Moisture Network: serving Earth system science for over a decade. *Hydrol. Earth Syst. Sci.* **25**, 5749–5804 (2021).
50. Novick, K. A. *et al*. The AmeriFlux network: A coalition of the willing. *Agricultural and Forest Meteorology* **249**, 444–456 (2018).
51. Drought 2018 Team, ICOS Ecosystem Thematic Centre, ICOS Ecosystem Thematic Centre & Trotta, C. Drought-2018 ecosystem eddy covariance flux product for 52 stations in FLUXNET-Archive format. *ICOS Ecosystem Thematic Centre*, https://doi.org/10.18160/YVR0-4898 (2020).
52. ICOS RI *et al*. Ecosystem final quality (L2) product in ETC-Archive format - release 2025-1. *ICOS Ecosystem Thematic Centre*, https://doi.org/10.18160/S6HM-CP8Q (2025).
53. Warm Winter 2020 Team, ICOS Ecosystem Thematic Centre, ICOS Ecosystem Thematic Centre & Trotta, C. Warm Winter 2020 ecosystem eddy covariance flux product for 73 stations in FLUXNET-Archive format—release 2022-1. *ICOS Ecosystem Thematic Centre*, https://doi.org/10.18160/2G60-ZHAK (2022).
54. Ueyama, M. *et al*. The JapanFlux2024 dataset for eddy covariance observations covering Japan and East Asia from 1990 to 2023. *Arctic Data archive System* https://ads.nipr.ac.jp/japan-flux2024/ (2024).
55. Zhang, P. *et al*. A 10 year (2009–2019) surface soil moisture dataset produced based on *in situ* measurements collected from the Tibet-Obs. *4TU.ResearchData*, https://doi.org/10.4121/12763700.v7 (2020).
56. Cho, K. *et al*. Calibration of the SMAP Soil Moisture Retrieval Algorithm to Reduce Bias Over the Amazon Rainforest. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* **17**, 8724–8736 (2024).
57. Spennemann, P. C., Fernández-Long, M. E., Gattinoni, N. N., Cammalleri, C. & Naumann, G. Soil moisture evaluation over the Argentine Pampas using models, satellite estimations and *in-situ* measurements. *Journal of Hydrology: Regional Studies* **31**, 100723 (2020).
58. Yang, K. *et al*. Network of soil temperature and moisture on the Pali (2015-2021). *National Tibetan Plateau Data Center* https://doi.org/10.11888/Terre.tpdc.301088 (2022).
59. Domine, F., Sarrazin, D., Nadeau, D., Lackner, G. & Belke-Brea, M. Hydrometeorological, snow and soil data from a low-Arctic valley in the forest-tundra ecotone in Northern Quebec. *PANGAEA* https://doi.org/10.1594/PANGAEA.964743 (2024).
60. Singh, G., Panda, R. K. & Mohanty, B. P. Spatiotemporal Analysis of Soil Moisture and Optimal Sampling Design for Regional-Scale Soil Moisture Estimation in a Tropical Watershed of India. *Water Resources Research* https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2018WR024044#support-information-section (2019).
61. Cosh, M., Kelly, V. & Colliander, A. SMAPVEX19-22 Millbrook Temporary Soil Moisture Network, Version 1. *NASA National Snow and Ice Data Center Distributed Active Archive Center* https://doi.org/10.5067/NXNJWN9933UI (2020).
62. Cosh, M., Kraatz, S. & Colliander, A. SMAPVEX19-22 Massachusetts Temporary Soil Moisture Network, Version 1. *NASA National Snow and Ice Data Center Distributed Active Archive Center* https://doi.org/10.5067/3LXL78PSKVXQ (2020).
63. Wang, C. *et al*. Chinese Soil Moisture Observation Network and Time Series Data Set. *figshare* https://doi.org/10.6084/m9.figshare.21302955.v2 (2022).
64. Boike, J. *et al*. Measurements in soil and air at Samoylov Station (2002-2018). *PANGAEA* https://doi.org/10.1594/PANGAEA.891142 (2018).
65. Boike, J., Miesner, F., Bornemann, N., Cable, W. L. & Grünberg, I. Trail Valley Creek, NWT, Canada Soil Moisture and Temperature 2016. *PANGAEA* https://doi.org/10.1594/PANGAEA.962726 (2023).
66. Gupta, H. V. & Kling, H. On typical range, sensitivity, and normalization of Mean Squared Error and Nash-Sutcliffe Efficiency type metrics. *Water Resources Research* **47**, 2011WR010962 (2011).
67. Liao, D., Niu, J., Du, T. & Kang, S. Improving Subsurface Soil Moisture Estimation Using a 2-Dimensional Data Assimilation Framework Incorporated With a Dual State-Parameter Scheme. *Water Resources Research* **60**, e2023WR035771 (2024).
68. Fatima, E. *et al*. Improved representation of soil moisture processes through incorporation of cosmic-ray neutron count measurements in a large-scale hydrologic model. *Hydrol. Earth Syst. Sci.* **28**, 5419–5441 (2024).
69. Al Bitar, A. *et al*. The global SMOS Level 3 daily soil moisture and brightness temperature maps. *Earth Syst. Sci.* **9**, 293–315 (2017).

70. Dorigo, W. *et al*. ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions. *Remote Sensing of Environment* **203**, 185–215 (2017).
71. Wigneron, J.-P. *et al*. SMOS-IC data record of soil moisture and L-VOD: Historical development, applications and perspectives. *Remote Sensing of Environment* **254**, 112238 (2021).
72. Feng, S. *et al*. Global daily 9 km remotely sensed soil moisture (2015–2025) with microwave radiative transfer-guided learning. *Zenodo* https://doi.org/10.5281/ZENODO.15826988 (2025).
73. Colliander, A. *et al*. Seasonal Dependence of SMAP Radiometer-Based Soil Moisture Performance as Observed Over Core Validation Sites. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium* 5320–5323 (2019).
74. Ambadan, J. T. *et al*. Evaluation of SMAP Soil Moisture Retrieval Accuracy Over a Boreal Forest Region. *IEEE Transactions on Geoscience and Remote Sensing* **60**, 1–11 (2022).
75. Bakke, S. J., Ionita, M. & Tallaksen, L. M. The 2018 northern European hydrological drought and its drivers in a historical perspective. *Hydrol. Earth Syst. Sci.* **24**, 5621–5653 (2020).
76. Rakovec, O. *et al*. The 2018–2020 Multi-Year Drought Sets a New Benchmark in Europe. *Earth's Future* **10**, e2021EF002394 (2022).
77. Peters, W., Bastos, A., Ciais, P. & Vermeulen, A. A historical, geographical and ecological perspective on the 2018 European summer drought. *Phil. Trans. R. Soc. B* **375**, 20190505 (2020).
78. Hovmöller, E. The Trough-and-Ridge diagram. *Tellus* **1**, 62–66 (1949).

Acknowledgements

This work was supported by the Danish Data Science Academy, which is funded by the Novo Nordisk Foundation (NNF21SA0069429) and VILLUM FONDEN (40516). This work was also supported by the VILLUM Young Investigator 2024 project (00072051), the Novo Nordisk Starting Grant (NNF23OC0087612), the SCALE project (AgriFoodTure, Innovation Fund Denmark), NASA ECOSTRESS Science and Applications Program (80NSSC23K0308), NASA Early Career Investigator Program in Earth Science (80NSSC24K1057), and the Global Wetland Center (NNF23OC0081089, Novo Nordisk Foundation). Additional funding was provided by the Pioneer Center for Research in Sustainable Agricultural Futures (Land-CRAFT), DNRF grant number P2, and Aarhus University. The author acknowledge NASA, ESA, INRAE and ECMWF for providing global SM products and land surface features, and thank ISMN, AmeriFlux, JapanFlux, ICOS and previous studies for providing available *in-situ* SM measurements.

Author contributions

S.F. and S.W. conceived this research and drafted the manuscript; S.F. and A.L. performed the experiments and data processing; S.W. provided supervision and guidance throughout the experiment and analysis; and all the authors reviewed and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-06721-6.

Correspondence and requests for materials should be addressed to S.W.