Remote Sensing of Environment 338 (2026) 115378

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

A novel soil moisture retrieval method via combining radiative transfer model and machine learning

Yurun Chen [a,1], Cheng Tong [b,c,*,1], Josh Qixuan Sun [d], Yulin Shangguan [b], Xiaodong Deng [b], Mark Crowley [d], Hongquan Wang [e], Yang Ye [c], Haijun Bao [c], Ruqi Huang [a]

[a] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
[b] College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China
[c] School of Spatial Planning and Design, Hangzhou City University, Hangzhou 310015, China
[d] Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
[e] Agriculture and Agri-Food Canada Lethbridge Research and Development Centre, Lethbridge, AB, Canada

ARTICLE INFO

Edited by Jing M. Chen

Keywords:
Soil moisture (SM)
Brightness temperature (TB)
Kolmogorov–Arnold network (KAN)
Radiative transfer model (RTM)
SMAP

ABSTRACT

Soil moisture (SM) is a key variable in the global water cycle, and passive microwave remote sensing is widely used for large-scale SM estimation. Although radiative transfer models (RTM) and machine learning (ML) approaches have advanced SM retrieval, interpretable methods that provide a direct, inspectable mapping from satellite observations to SM remain scarce. In this study, we present an initial attempt to develop such an interpretable retrieval framework by fusing the physical principles of the RTM with the mathematical formulation capability of the Kolmogorov–Arnold Network (KAN). The framework proceeds in two stages: first, a physically consistent training dataset is generated by the RTM that links SM with brightness temperature (TB), surface temperature (ST), and vegetation optical depth (VOD), thereby preserving key physical dependencies; second, KAN is trained on this dataset to extract explicit mathematical expressions that relate TB, ST and VOD to SM for different land-cover types. These derived expressions are applied to SMAP observations to produce global daily SM estimates for 2015–2023. Validation against in situ soil moisture measurements from global networks, including the International Soil Moisture Network (ISMN) and the Qinghai Lake Basin dense network (QLB-NET), shows that the KAN retrieval achieves an average correlation coefficient $R = 0.64$ and unbiased RMSE (ubRMSE) $= 0.07\ m^3/m^3$, comparable to the SMAP Level-3 product ($R = 0.65$, ubRMSE $=0.06\ m^3/m^3$) in capturing broad spatial patterns and seasonal dynamics. Beyond enabling accurate large-scale estimates, the explicit formula extracted by KAN clarifies how TB, ST, and VOD influence SM and provides a compact, direct retrieval expression that requires neither KAN retraining nor the iterative solving typical of traditional RTM approaches, thereby improving reproducibility and operational efficiency. This interpretable fusion of physical modeling and data-driven representation offers a novel pathway for advancing quantitative remote sensing retrievals and invites further refinement and operational evaluation.

1. Introduction

Soil moisture (SM) is a pivotal state variable in terrestrial ecosystems, governing the coupled water–energy–carbon exchanges at the land–atmosphere interface (Humphrey et al., 2021; McColl et al., 2017). As a critical hydrological parameter, SM influences key processes in agriculture, hydrology, and climate regulation, shaping phenomena such as crop productivity, runoff generation, and land–atmosphere feedbacks (Brocca et al., 2018; Seneviratne et al., 2010). Continuous, large-scale SM observations are therefore indispensable for advancing earth system science and supporting applications such as climate change assessment, water resource management, and hazard monitoring.

SM can be measured directly through in-situ monitoring networks, which provide high temporal fidelity at the point scale but are

* Corresponding author at: College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China
E-mail addresses: cyr22@mails.tsinghua.edu.cn (Y. Chen), Chengton@zju.edu.cn (C. Tong), q84sun@uwaterloo.ca (J.Q. Sun), yulinsg@zju.edu.cn (Y. Shangguan), xddeng@zju.edu.cn (X. Deng), mark.crowley@uwaterloo.ca (M. Crowley), Hongquan.Wang@USherbrooke.ca (H. Wang), yeyang@hzcu.edu.cn (Y. Ye), baohaijun@hzcu.edu.cn (H. Bao), ruqihuang@sz.tsinghua.edu.cn (R. Huang).
[1] These authors contributed equally to this work (co-first authors).

https://doi.org/10.1016/j.rse.2026.115378
Received 5 September 2025; Received in revised form 11 February 2026; Accepted 14 March 2026
Available online 20 March 2026

inherently limited by sparse station density and uneven spatial coverage (Peng et al., 2017). To overcome these spatial constraints, passive microwave remote sensing has emerged as the dominant retrieval approach, leveraging the dielectric contrast between liquid water and dry soil constituents, a physical mechanism that enables robust large-scale SM retrievals. Compared with optical/thermal sensors, microwave observations penetrate vegetation, probe deeper soil layers, and are less affected by clouds and atmospheric conditions, enabling day-night and all-weather monitoring. Capitalizing on these strengths, a series of satellites equipped with microwave sensors, including the Soil Moisture Active Passive (SMAP), Soil Moisture and Ocean Salinity (SMOS), and Advanced Microwave Scanning Radiometer 2 (AMSR2), have been launched, delivering invaluable time-series SM products (Entekhabi et al., 2010; Imaoka et al., 2010; Kerr et al., 2010).

Passive microwave remote sensing estimates SM from brightness temperature (TB) observations, whose variations reflect changes in soil dielectric properties. However, TB is also affected by factors such as surface roughness, vegetation, and temperature (Wigneron et al., 2017). To separate these effects, physical radiative transfer models (RTMs) are widely used, simulating microwave propagation through the soil–vegetation–atmosphere system to enable SM retrieval (Karthikeyan et al., 2017). Built on RTMs, a range of algorithms have been developed for SM retrieval, including the single-channel algorithm (SCA), the dual-channel algorithm (DCA), and the L-band Microwave Emission of the Biosphere (L-MEB) model. In addition to physical radiative transfer models, machine learning (ML) has gained increasing attention in SM retrieval and reconstruction (Zhang et al., 2025; Zhao and Sima, 2024), as it directly establishes relationships between input variables such as TB with the output prediction of SM, thereby circumventing the complexities associated with radiative transfer modeling. For instance, a Neural Network (NN) algorithm was trained using SMAP TB as the primary input and GEOS-5 SM as the output to estimate global SM (Kolassa et al., 2018). Qu et al. (2019) employed a Random Forest (RF) method to develop a predictive model linking SMAP SM with AMSR-E/AMSR2 TB, reconstructing a time series of SM over the Qinghai–Tibet Plateau. Moreover, some studies directly use ground-based station data to establish a nonlinear relationship with microwave TB through ML techniques for SM retrieval. Yuan et al. (2020) developed a Generalized Regression Neural Network (GRNN) that used SMAP TB and in-situ station measurements as references to estimate regional SM.

Notably, the two dominant approaches to SM retrieval (RTMs and ML) each exhibit distinct advantages and limitations. ML-based methods provide relatively straightforward implementation by bypassing the complexities of radiative transfer, thereby simplifying the inversion process through direct data-driven mapping. However, these approaches demand extensive training datasets and often display limited transferability across regions, sensors, or land-cover conditions. Moreover, the inherent "black-box" nature of ML algorithms poses interpretability challenges, which may undermine scientific confidence in the retrieved estimates. In contrast, RTM-based methods are grounded in well-established physical principles, ensuring theoretical consistency and broad generalizability across diverse environments. However, unlike other remote sensing parameters such as land surface temperature (LST), which can be directly retrieved from satellite observations using well-established analytical formulations (Zhao and Sima, 2024), SM retrieval from TB within an RTM framework generally does not rely on a simple explicit closed-form expression. Instead, SM retrieval typically requires constructing and iteratively minimizing a cost function within a radiative transfer framework, which substantially increases computational demands and limits the practicality of direct application. Moreover, this iterative inversion process often suffers from ill-posedness, where multiple parameter combinations can yield similar TB values, further complicating the retrieval (Konings et al., 2015). Recent research has increasingly focused on hybrid methodologies that integrate ML with RTMs. For example, Mao et al. (2023) developed an integrative framework that combines deep learning with physical modeling and statistical optimization, enabling the simultaneous retrieval of SM and LST from passive microwave observations. This approach demonstrated improved accuracy by leveraging synergistic dependencies between SM and LST within a multi-task learning paradigm. Similarly, Lee et al. (2025) proposed a deep neural network (DNN) architecture that embeds physical constraints into the retrieval process, replacing traditional algorithms while addressing scale mismatch and improving parameterization in the context of the SMAP mission. Li et al. (2025) developed a cloud-based intelligent framework that integrates physical principles with machine learning for global surface SM retrieval, incorporating environment-specific model selection and cloud-based implementation to enhance both accuracy and operational efficiency. Notably, while the potential of RTM-ML integration has been demonstrated in these pioneering studies, a critical limitation has been identified: current hybrid approaches have yet to establish a fully interpretable, physically consistent, and universally applicable physical formulation for SM retrieval. This gap is most evident in the core challenge of integrating data-driven learning with the physical constraints of first-principles radiative transfer models.

Recently, Kolmogorov-Arnold Networks (KANs) (Bozorgasl and Chen, 2024) were proposed, inspired by the Kolmogorov-Arnold representation theorem, which states that any high-dimensional continuous function can be represented as a superposition of univariate functions. Building on this principle, KANs introduce a neural network architecture that aims to represent complex input–output relationships in a formula-like manner. Instead of using fixed activation functions, the network employs learnable spline-parameterized functions constructed from a rich set of basis functions. Through training, KANs explicitly encode nonlinear relationships between inputs and outputs into mathematical expressions, enabling direct inspection and interpretation of the learned mapping. This design not only enhances interpretability but also achieves higher parameter efficiency and formula-expressiveness compared to traditional multilayer perceptrons (MLPs). Leveraging this formula-extraction capability, Lee et al. (2024) used KANs to derive a symbolic expression for predicting global warming potential, while Granata et al. (2024) compared KANs and Transformers in streamflow forecasting for Central European rivers and highlighted the substantial interpretability advantages of KANs.

Inspired by KANs, this study represents the first attempt to integrate RTM with the KANs architecture, leveraging KAN's powerful formula restoration capability to directly model nonlinear complex physical processes as interpretable mathematical equations for SM estimation. The organizational structure of this research is outlined as follows: Section 2 details the experimental datasets employed, Section 3 elaborates on the methodological framework, Sections 4 and 5 present the results and corresponding discussions, and Section 6 concludes the study.

2. Data

2.1. SMAP TB data

The SMAP satellite, launched by NASA in 2015, is designed to provide global measurements of SM and freeze-thaw states. It was initially equipped with a synergistic system comprising an L-band radiometer and radar. However, following the radar malfunction, the radiometer continued to operate independently. Surface brightness temperature is collected by the SMAP radiometer during both descending and ascending orbits through four channels: H-polarization (TBH), V-polarization (TBV), T3, and T4. TBH and TBV are primarily utilized for SM retrieval and freeze-thaw state monitoring, whereas T3 and T4 are dedicated to detecting Radio Frequency Interference (RFI), thereby ensuring the accuracy and reliability of the radiometer's measurements. In this study, we specifically selected vertically polarized (V-pol) SMAP TB values that had been corrected for both RFI and water body contamination, as indicated by the quality flags (O'Neill et al., 2020).

The daily TB data with a spatial resolution of 36 km used in this study covers the period from 2015 to 2023.

2.2. SMAP SM data

This study utilizes SMAP Level-3 global SM products derived from SCA-V, which employs only the TBV within a radiative transfer framework to retrieve SM. As single-polarization observations cannot resolve two unknown parameters, VOD obtained from climatological Normalized Difference Vegetation Index (NDVI) datasets was incorporated as a constraint, following established methodology (Njoku and li, 1999). To maintain temporal consistency with the TB dataset, all analyzed SM products were exclusively taken from descending overpasses, which occur at approximately 6:00 a.m. local solar time.

2.3. In situ data

The retrieval accuracy was validated using in situ SM measurements obtained from the International Soil Moisture Network (ISMN). This global data harmonization platform aggregates standardized SM observations through international collaboration from ground-based networks globally (Dorigo et al., 2011; Ma et al., 2019). Surface layer measurements (0–5 cm or 5 cm depth) were prioritized to match the microwave signal penetration depth characteristic of L-band retrievals. Furthermore, in situ soil moisture observations at 5 cm depth from the Tianjun dense soil moisture network in the Qinghai Lake Basin (QLB-NET), which provides high-density measurements within SMAP 36 km grid cells over the northeastern Tibetan Plateau, were also used for validation (Chai et al., 2024; Liu et al., 2024).

Fig. 1 displays the global distribution of validation stations used in this study. To ensure the reliability of the in-situ data, a rigorous quality control process was conducted. This involved filtering out stations with missing records, removing outliers, and retaining only high-quality observations for validation. Each in-situ station is collocated with the corresponding SMAP 36 km grid cell based on its geographic coordinates, ensuring a one-to-one station–pixel pairing for the validation analysis. The stations encompass diverse surface types spanning grasslands, croplands, bare land, and shrublands, enabling systematic evaluation of retrieval performance across heterogeneous terrain conditions. Key attributes of the selected validation sites, including station names and dominant land cover types, are summarized in Table 1.

2.4. Other auxiliary data

In addition to the SMAP data mentioned above, this study also utilizes two important parameters involved in the SMAP SM retrieval process: Surface Temperature (ST) and VOD. The ST data used in the SMAP algorithm primarily come from the first and second layers of soil temperature in NASA's GEOS-5 land modeling system. The effective ST is calculated using the surface and deep soil temperatures from this system, applying an empirical formula with coefficients. VOD is derived from vegetation water content (VWC), which is calculated using the Normalized Difference Vegetation Index (NDVI) and an empirical

Table 1
Descriptions of in situ stations used in this study.

Network name	Country	Main land cover	Sensors	Station number
CTP_SMTMN	China	Grassland	EC-TM	54
Naqu	China	Grassland	5TM	9
Maqu	China	Grassland	EC-TM	21
NGARI	China	Barren	5TM	18
QLB-NET	China	Diverse	CS655	60
PBO_H2O	America	Grassland, cropland and forest	GPS	152
USCRN	America	Diverse	Stevens Hydraprobe II Sdi-12	115
SCAN	America	Diverse	Hydraprobe Sdi-12	221
TxSON	America	Cropland	CS655	40
RSMN	Romania	Cropland and forest	5TM	19
REMEDHUS	Spain	Grassland and shrub	Stevens Hydra Probe	20
HOBE	Denmark	Cropland and forest	Decagon 5TE	29
SMOSMANIA	France	Grassland and forest	ThetaProbe ML2X	21
OZNET	Australia	Grassland and forest	Stevens Hydra Probe	19

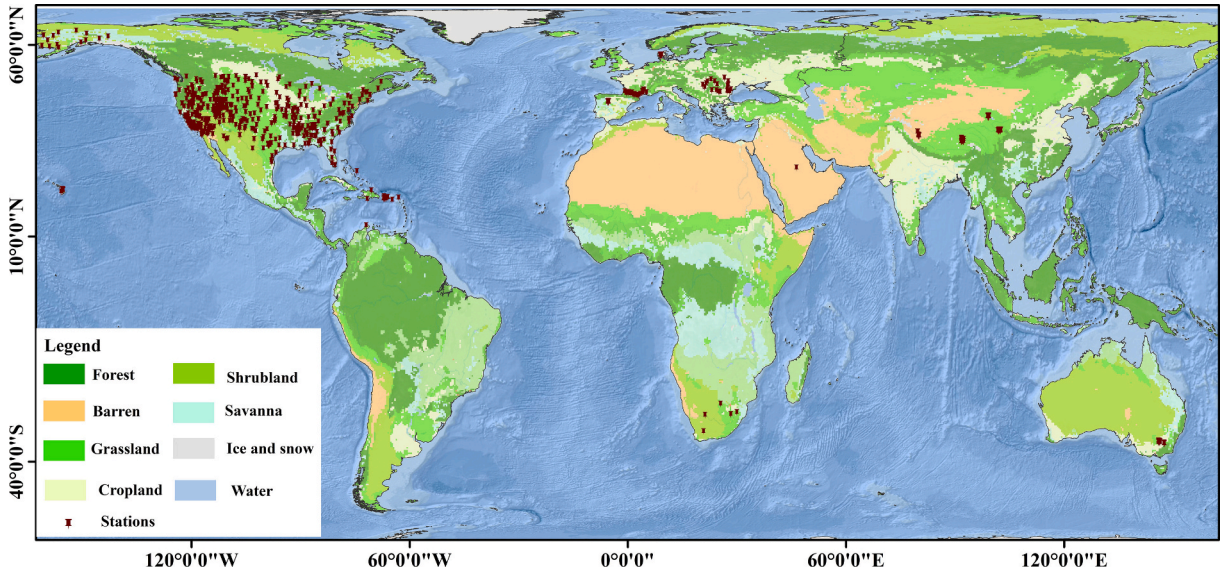

Fig. 1. Distribution of selected in situ stations and their corresponding land cover types. The background land-cover map is based on the IGBP classification and is aggregated for visualization into broader categories: forests (evergreen needleleaf forests, evergreen broadleaf forests, deciduous needleleaf forests, deciduous broadleaf forests, and mixed forests), shrublands (closed and open shrublands), savannas (woody and open savannas), grasslands, croplands (including cropland/natural vegetation mosaics), and barren land.

model. The relationship between VWC and NDVI is expressed through an equation that includes a "Stem Factor" related to land cover type. The VOD is then computed as the product of VWC and a constant parameter **b**, which varies by vegetation type and microwave frequency. (O'Neill et al., 2020).

3. Methods

To construct an explicit and physically consistent SM retrieval model, this study adopts a two-stage strategy that integrates an RTM with a KAN (see Fig. 2). The RTM is first used to simulate a physically consistent forward dataset linking SM, surface temperature, vegetation optical depth, and brightness temperature. Based on this dataset, KAN is then trained to learn an interpretable inverse mapping from satellite-observed variables to SM, which can be further expressed in an explicit mathematical form.

3.1. Rationale of SM retrieval

Soil moisture content has a direct impact on the dielectric properties of soil. The dielectric constant of dry soil is approximately 3, while that of pure water is around 80. The dielectric constant of SM falls between these two values. As a result, the relationship between SM and dielectric constant k' is described by the soil dielectric model (Topp model) (Topp et al., 1980) as follows:

$$k' = 3.03 + 9.3mv + 146mv^2 - 76.7mv^3 \quad (1)$$

where mv represents soil moisture, with units of m^3/m^3.

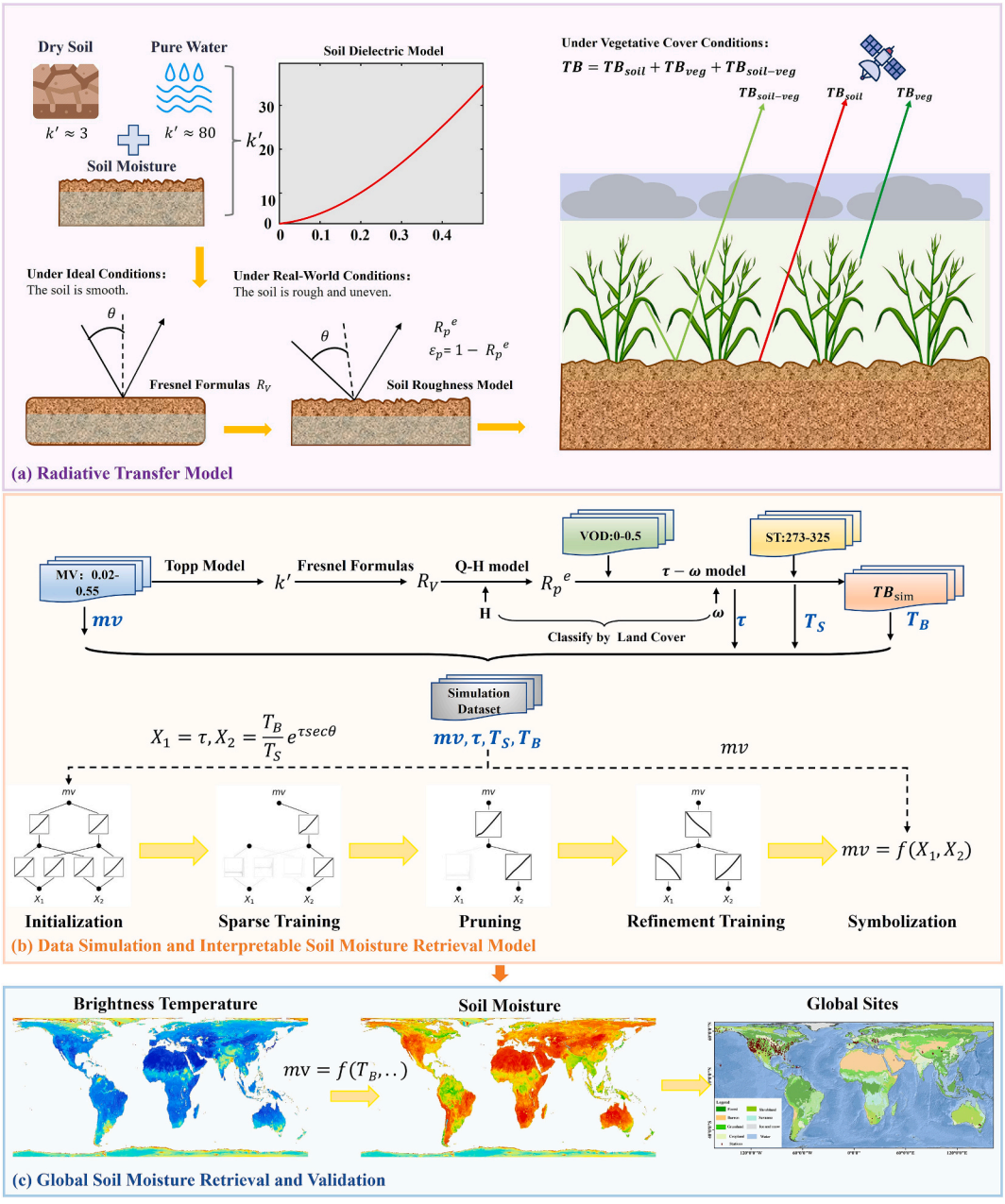

Fig. 2. Overview of the proposed soil moisture retrieval framework: (a) Forward radiative transfer modeling used to simulate physically consistent relationships among soil moisture, surface temperature, vegetation optical depth, and brightness temperature. (b) Generation of the training dataset and KAN-based inverse modeling, including sparse training, pruning, refinement, and symbolic formulation. (c) Application of the derived interpretable model to SMAP observations and validation against global in situ measurements.

Assuming a smooth soil surface, the surface reflectivity is primarily influenced by the soil dielectric constant and can be described using the Fresnel equation.

$$r_V = \left| \frac{k' cos\theta - \left(k' - sin^2\theta\right)^{0.5}}{k' cos\theta + \left(k' - sin^2\theta\right)^{0.5}} \right|^2 \quad (2)$$

In this equation, θ represents the satellite's incidence angle, and r_V denotes the soil reflectivity in V-polarization.

Under natural conditions, the soil surface is uneven, with varying degrees of roughness that significantly influence its actual reflectivity, deviating from the idealized assumption of a smooth surface. As a result, soil roughness models are commonly used to calculate the actual soil reflectivity, as shown below:

$$R_p{}^e = \left[(1-Q) \bullet r_p + Q \bullet r_q\right] \bullet \exp\left(-\mathrm{H} cos\theta^2\right) \quad (3)$$

Where $R_p{}^e$ represents the actual surface reflectance, r_p denotes the reflectance of smooth soil, and Q and H are roughness-related parameters, which are represented by empirical parameters according to different land cover types. p and q correspond to different polarizations.

In vegetation-covered soil layers, the signals acquired by satellites not only include radiation from the soil but also radiation contributed by the vegetation and the radiation resulting from the interaction between the vegetation and the soil layer, as follows. To describe the radiation transfer process between the soil and vegetation, the zero-order vegetation radiation transfer model (τ-ω model) (Mo et al., 1982) is commonly used, as follows:

$$\begin{aligned} TB = {} & T_S\left(1-R_p{}^e\right)\exp\left(-\tau sec\theta\right) + T_c(1-\omega)\left[1-\exp\left(-\tau sec\theta\right)\right] \\ & + T_c R_p{}^e (1-\omega)\left[1-\exp\left(-\tau sec\theta\right)\right]\exp\left(-\tau sec\theta\right) \end{aligned} \quad (4)$$

In this formula, T_S and T_c represent the effective temperatures of the soil and vegetation, respectively; ω is the single scattering albedo of the vegetation; and τ is the VOD. Here, TB denotes the upwelling brightness temperature at the top of the canopy, representing the combined contributions of the soil and vegetation layers in the τ-ω model.

3.2. Generation of training dataset based on RTM

In this study, a comprehensive data simulation was implemented to generate a dataset for KAN training based on the aforementioned radiation transfer process. The process begins with the modeling of SM variations, where the corresponding dielectric constant for each SM value is calculated using Eq. (1), forming the foundation for subsequent computations. Next, the ideal soil reflectance is determined under idealized conditions using the Fresnel equation, as described in Eq. (2). To account for surface roughness effects, the QH model in Eq. (3) is applied to derive the actual soil reflectance, ensuring the reflectance values better represent real-world conditions. The simulation process further accounts for the influence of vegetation and surface temperature, with variations in VOD and ST serving as inputs to the τ-ω model, as described in Eq. (4), which are crucial for accurately simulating TB values under diverse surface conditions. All parameter settings are shown in Table 2. It is important to note that certain parameters, such as the H parameter in the QH model and the ω parameter in the τ-ω model, are set according to surface type. In the SMAP retrieval algorithm, the Earth's land cover is classified into 17 types based on the International Geosphere-Biosphere Programme (IGBP) classification system (Friedl et al., 2002), as shown in Table 3, encompassing categories such as forests, shrublands, grasslands, and others. Therefore, according to our parameter setting range, we can obtain a large-scale dataset, comprising diverse combinations of mv, τ, T_S, and the corresponding TB (can be denoted as T_B) across various land cover types.

In practice, we generated 10 groups of RTM-simulated datasets, corresponding to the 10 land cover types considered in this study. Each group consists of approximately 75,000 samples, covering diverse combinations of mv, τ, T_S, and the corresponding T_B. To facilitate model development and evaluation, the samples in each group are randomly divided into 80% for training and 20% for testing, with the testing subsets also used for validation purposes.

Table 2
Parameter settings for simulated data.

Parameter	Range	Step	Model used
mv	0.02–0.55 m^3/m^3	0.01 m^3/m^3	Topp model
τ	0–0.5	0.01	τ-ω model
T_S	273–325 K	1 K	τ-ω model

Table 3
MODIS land-cover classes and the proposed reclassification scheme.

ID	MODIS land cover types	h	ω	New types
0	Water Bodies	0	0	–
1	Evergreen Needleleaf Forests	0.160	0.070	
2	Evergreen Broadleaf Forests	0.160	0.070	
3	Deciduous Needleleaf Forests	0.160	0.070	1
4	Deciduous Broadleaf Forests	0.160	0.070	
5	Mixed Forests	0.160	0.070	
6	Closed Shrublands	0.110	0.050	2
7	Open Shrublands	0.110	0.050	
8	Woody Savannas-	0.125	0.050	3
9	Savannas	0.156	0.080	4
10	Grasslands	0.156	0.050	5
11	Permanent Wetlands	0	0	6
12	Croplands Average	0.108	0.050	7
13	Urban and Built-up Lands	0.000	0.030	8
14	Crop-land/Natural Vegetation Mosaics	0.130	0.065	9
15	Snow and Ice	0	0	–
16	Barren	0.150	0	10

3.3. KAN training

After reclassification, our model treats h and ω as constants in each new land cover type, so we can focus on the three remaining inputs: TB (T_B), ST (T_S), and VOD (τ). After obtaining a large-scale dataset with known physical consistency, covering diverse environmental conditions and land surface types. We aim to establish a synergistic framework based on RTM and KAN for physically consistent and interpretable SM retrieval. To be specific, as shown in Fig. 3, the forward RTM model simulates a comprehensive, physically consistent dataset, and these

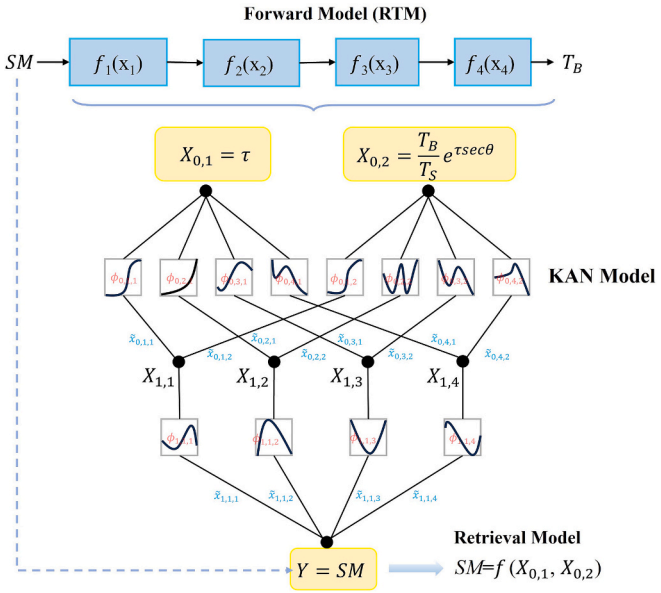

Fig. 3. The detailed structure of combining RTM and KAN.

RTM-simulated, satellite-like observables to learn an explicit mapping from the input variables to SM while remaining consistent with the underlying physical assumptions.

Since KANs are better suited for handling continuous functions, some discrete parameters such as surface roughness (h) and single scattering albedo (ω), which vary by land cover type, are not directly used as inputs. Instead, we reclassify the land cover types into 10 new categories based on their h and ω values. For each reclassified type, we treat h and ω as constants and generate a distinct sub-simulated dataset. This allows the model to focus on learning the continuous relationship between T_B, T_S, and τ as inputs and SM as the output. This reclassification process itself can be regarded as an important step that strengthens the physical consistency between RTM and KAN.

The KAN model is then trained on this synthetic dataset to approximate the inverse mapping: from T_B, T_S and τ to SM. Because the training samples originate entirely from RTM, the learned mapping is constrained to follow the physical laws embedded in the RTM equations. Moreover, KAN's architecture enables the learned function to be represented symbolically, allowing direct inspection of partial derivatives and the identification of key physical influences.

To conclude, RTM defines the structure and boundary of the physically plausible feature space, and KAN extracts interpretable, data-driven insights from within this space. As a result, the final symbolic expression produced by KAN inherits both the accuracy of data-driven learning and the reliability of physically informed constraints.

3.3.1. Initialization

A KAN architecture (Bozorgasl and Chen, 2024) consists of two components: edges and nodes. As shown in Fig. 4, KANs place learnable activation functions on edges, whereas MLPs use learnable weights on edges. The nodes in KANs, also referred to as neurons, simply sum incoming signals without applying non-linearities, while MLPs use fixed activation functions on nodes. There are no linear weight matrices in KANs, which increases their interpretability.

We use $[n_1, n_2, \ldots n_L]$ to represent the shape of an L-layers network, where n_i denotes the number of nodes in the i^{th} layer. The simulated dataset derived from the radiative transfer model is reorganized according to the findings of the variable analysis. We select τ and $\frac{T_B}{T_S}e^{\tau \sec\theta}$ as the network inputs, denoted as X_1 and X_2 for convenience. Subsequently, these two variables are extracted from the simulated dataset and used to train the constructed KAN with the corresponding simulated soil moisture values. The network, as illustrated in Fig. 4, is initialized with a layer structure of [2,2,1].

In the right part of Fig. 4, the black dots represent nodes, including inputs and outputs. The connecting lines represent activation functions between nodes, with curves illustrating their shapes. We denote the i^{th} node in the l^{th} layer as (l, i). For each activation function $\phi_{l,j,i}$ in a KAN layer $\mathbf{\Phi}_l$, l denotes the layer index, while i and j denote input and output node indices, respectively.

This structural design allows a KAN to serve as an interpretable function approximator, which is particularly suitable for learning physically constrained relationships in remote sensing applications.

3.3.2. Sparse training

Sparse training is introduced to suppress redundant or weak activation functions, so that the learned mapping focuses on the most physically meaningful relationships rather than fitting noise. We begin sparse training after loading the simulated data into the network M_0. For an L-layer KAN, the total training objective ℓ_{total} is defined as follows:

$$\ell_{total} = \ell_{pred} + \lambda_{L1}\sum_{l=0}^{L-1}|\Phi_l|_1 + \lambda_{entropy}\sum_{l=0}^{L-1} S(\Phi_l), \tag{5}$$

where λ_{L1} and $\lambda_{entropy}$ control L1 regularization and entropy regularization, respectively. These two regularizations encourage the learning of simpler and sparser models and allow the model to learn the importance of each node during training, distinguishing smooth activation functions that may be interpretable from less important activation functions. After the model converges, we obtain an updated network, denoted as M_0', which reflects the learned sparse patterns and relationships from the data.

3.3.3. Pruning

Next, we prune some edges and nodes of M_0' to reduce its complexity, yielding a more efficient and interpretable network, M_1. Specifically, we sparsify KANs at the node level. For each node, we define its incoming and outgoing scores as follows:

$$I_{l,i} = \max_k\left(|\phi_{l-1,i,k}|_1\right), O_{l,i} = \max_j\left(|\phi_{l+1,j,i}|_1\right). \tag{6}$$

We consider a node important if both incoming and outgoing scores exceed a threshold hyperparameter, set to $\theta = 10^{-2}$ by default. All unimportant nodes and their related activation functions are pruned. After pruning, the network retains only those relatively smooth activation functions and their associated nodes that are more interpretable. In the context of physical modeling, smoother activation functions tend to represent more stable and monotonic relationships, which are easier to interpret and more consistent with physical processes. The entire network structure becomes sparse, and the relationship between input and output becomes clearer.

3.3.4. Refinement training

After pruning the model, we proceed with an adjustment called refinement training, which is similar to sparse training as described in Eq. (5). In this phase, we reduce the regularization term and retain only l_{pred}. This modification allows the model to focus more on fitting the data

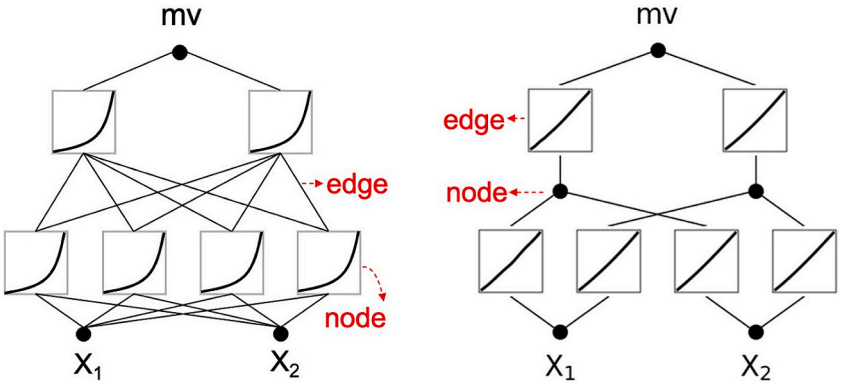

Fig. 4. Comparison between MLPs and KANs. MLPs: learnable weights on edges; fixed activation functions on nodes. KANs: learnable activation functions on edges; sum operation on nodes.

accurately. After this refinement training phase, we obtain M_1'.

3.3.5. Symbolification

The final step involves the symbolification of M_1' to convert the learned model into an explicit mathematical formula, denoted as M_{sym}. This step does not involve additional learning, but instead provides an explicit analytical approximation of the trained KAN model.

At this stage, we match each activation function with every candidate function in a basic function library consisting of simple and commonly used functional forms. For symbolification metrics, we use a weighted loss $\mathscr{L}$ that weighs model accuracy (measured as R^2) against model complexity $\mathscr{L}_{cplx}$ shown in Equation to reach a trade-off between approximation accuracy and functional simplicity. In this equation, $W_{cplx} \in [0,1]$ controls the relative importance of model complexity and approximation accuracy.

$$\mathscr{L} = W_{cplx}\mathscr{L}_{cplx} + (1 - W_{cplx})\mathscr{L}_{R^2} \tag{7}$$

We select the best-fitting basic function with minimum $\mathscr{L}$ to represent each activation function. The full lookup table for model complexity is provided in the appendix B.

This symbolic representation not only provides insights into the underlying relationships captured by the model, but also allows for easier interpretation and analysis.

3.4. Experiment setting and metrics

To evaluate the performance of the retrieval results, in situ SM measurements obtained from global Networks were used as reference values for validation, four error metrics are adopted: Root Mean Square Error (RMSE), Unbiased Root Mean Square Error (ubRMSE), Bias, and the Pearson correlation coefficient (R). The formulas for these metrics are as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\theta_{retrieval,i} - \theta_{reference,i}\right)^2} \tag{8}$$

$$ubRMSE = \sqrt{E\left\{\left(\theta_{retrieval} - E[\theta_{retrieval}]\right) - \left(\theta_{reference} - E[\theta_{reference}]\right)\right\}^2} \tag{9}$$

$$Bias = E[\theta_{retrieval} - \theta_{reference}] \tag{10}$$

$$R = \frac{E\left[\left(\theta_{retrieval} - E[\theta_{retrieval}]\right)\left(\theta_{reference} - E[\theta_{reference}]\right)\right]}{\sigma_{retrieval}\sigma_{reference}} \tag{11}$$

where $E[\cdot]$ denotes the mean processing, $\theta_{retrieval}$ refers to the retrieval SM values, and $\theta_{reference}$ refers to the reference SM values. σ represents the standard deviation of the retrieval and reference values.

4. Results

This section describes the results of training the KAN model outlined above on the modified classification dataset described in Section 3.2.

4.1. Experimental results on training dataset

After sparse training, the resulting KAN is shown in Fig. 5(a). During sparse training, certain nodes were identified as insignificant based on Eq. (7), resulting in these nodes and their associated activation functions becoming more transparent.

A clear outcome in the simulated datasets is that our KAN tends to retain only one node in the *sec*ond layer of the network. Fig. 5(b) shows that the pruned KAN simplifies to a structure of $[2, 1, 1]$. The result of refinement training is illustrated in Fig. 5(c). We observe that each activation function in the network fits the data better and becomes more like a common basic function. For symbolification, each activation function is processed separately. Table 4 shows the weighted losses $\mathscr{L}$ of these three activation functions in the second new land cover type. Results for all 10 new types are provided in the appendices.

For the symbolization of the three activation functions $\phi_{0,0,0}$, $\phi_{0,0,1}$, and $\phi_{1,0,0}$, we select e^x, x^2, and x functions. These selections were made after extensively screening all candidate functions in the library and comparing their weighted losses. The chosen functions consistently produced the lowest losses across the evaluations, and their mathematical forms closely resemble the refined activation functions learned by the KAN. Our model suggests the same basic functions for all 10 new land cover types. Thus, we can use the following simple expression to describe our symbolic retrieval model:

$$mv = \alpha(X_2 + \beta)^2 + \gamma e^{\delta X_1} + C. \tag{12}$$

Based on previous analysis, we designate τ and $\frac{T_B}{T_S}e^{\tau\sec\theta}$ as inputs to the network, referred to as X_1 and X_2 for convenience. Table 5 displays the detailed parameters after symbolization for our 10 new land cover types using simulated data. These parameters indicate that, across different cover types, the model parameters exhibit slight variations. However, the sign of each parameter remains consistent.

4.2. Comparison with SMAP product

To further quantify the reliability of the KAN SM retrieval relative to the SMAP SM product, a comprehensive comparative analysis was conducted between the KAN SM and the SMAP SM product. Fig. 6

Table 4
Weighted losses for activation functions across five basic mathematical functions, used in the symbolification process. Bold values indicate the lowest loss for each activation function.

	C	x^2	1/x	log(x)	e^x
$\phi_{0,0,0}$	−0.31	−0.79	−1.31	−0.91	**−1.66**
$\phi_{0,0,1}$	−0.58	**−1.68**	−1.15	−0.99	−1.36
$\phi_{1,0,0}$	**−0.64**	−0.05	−0.09	−0.10	−0.07

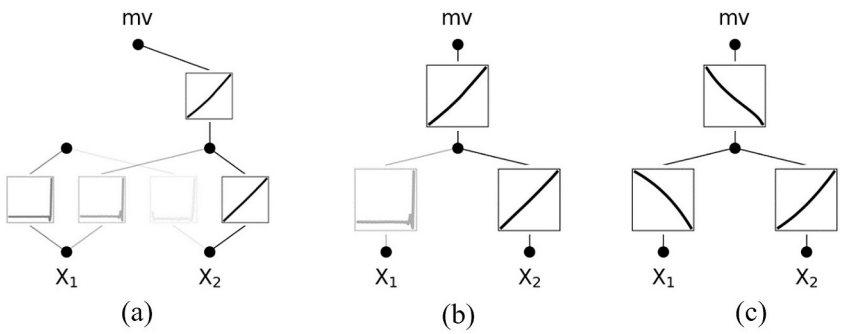

Fig. 5. Example of model training process on simulated dataset: (a) Network after sparse training, (b) Pruned KAN, and (c) Refined network.

Table 5
Detailed parameters of the symbolic model for 10 new land cover types.

Type	α	β	γ	δ	C
1	−0.57	0.36	0.83	2.27	0.19
2	−0.56	0.35	0.81	2.30	0.17
3	−0.57	0.35	0.82	2.30	0.17
4	−0.56	0.37	0.83	2.25	0.21
5	−0.58	0.34	0.83	2.30	0.17
6	−0.54	0.31	0.78	2.38	0.13
7	−0.56	0.35	0.81	2.30	0.17
8	−0.53	0.35	0.78	2.32	0.15
9	−0.56	0.36	0.82	2.27	0.19
10	−0.60	0.29	0.84	2.40	0.13

presents the global distribution of pixel-wise temporal evaluation metrics, including the correlation coefficient (R), ubRMSE, RMSE, and bias. At the global scale, the two datasets exhibit very strong agreement: R exceeds 0.9 for the vast majority of land pixels, yielding a global mean correlation of 0.98. Outside densely forested regions, the ubRMSE remains generally low, with most pixels showing differences smaller than 0.02 m^3/m^3 between KAN SM and SMAP SM. This indicates that the KAN retrieval not only reproduces the overall magnitude of SMAP SM, but also captures its temporal variability with high fidelity across most climate and land-cover regimes.

The spatial patterns in Fig. 6 also highlight where the agreement between the two products deteriorates. In regions with dense vegetation cover, particularly the Amazon rainforest and the Congo Basin, a limited number of pixels exhibit markedly elevated RMSE and substantial bias, reflecting systematic discrepancies between the two products. These high-error areas are consistent with the well-known difficulties of microwave SM retrieval under high-vegetation conditions (Ambadan et al., 2022; Park et al., 2024), and the associated error sources will be examined in more detail in the discussion section. Nevertheless, the overall spatial coherence of the R, ubRMSE, RMSE, and bias fields suggests that the KAN SM provides a temporally and spatially consistent approximation to SMAP SM at the global scale.

It is worth emphasizing that this level of agreement is achieved with fundamentally different retrieval strategies. The SMAP SM product is derived through an iterative inversion of the vegetation radiative transfer model, which repeatedly optimizes SM to match the observed TB for each pixel and time step (Tong et al., 2025). In contrast, the KAN approach directly applies an explicit functional relationship between TB and SM, learned from the physical-constraint training framework, to obtain SM in a single forward calculation. This avoids the computational cost and potential instability of iterative optimization, and highlights the strong potential and practical value of the KAN-based formulation for fast, large-scale SM retrieval while maintaining close consistency with an established benchmark product.

4.3. Validation results based on in situ measurements

The accuracy of KAN SM was further evaluated through comparison with in situ measurements from multiple ground networks, with the SMAP SM also included for reference. Each independent station was assessed by calculating the correlation coefficient (R), RMSE, ubRMSE, and bias. Because more than one station can fall within the same SMAP 36-km grid cell (and networks contain multiple stations), the validation statistics are summarized using boxplots (Fig. 7). Overall, both KAN SM and SMAP SM show comparable temporal consistency with in situ observations, indicating that they capture the timing and magnitude of SM variability reasonably well. The mean correlation coefficient across all stations was 0.65 for SMAP SM and 0.64 for KAN SM, suggesting very similar skill in reproducing temporal dynamics. Regarding errors, KAN SM exhibits a slightly higher ubRMSE (0.07 m^3/m^3) than SMAP SM (0.06 m^3/m^3). Despite this small difference, both datasets maintain robust performance in tracking SM fluctuations, which supports the reliability of the KAN-based approach while offering a more compact, computationally efficient, and interpretable formulation.

Specifically, the majority of stations yield R values above 0.6 for KAN SM, with notably strong correlations ($R > 0.8$) at the TxSon, NGARI, and CTP_SMTMN stations. In terms of error metrics, KAN SM generally shows slightly larger dispersion than SMAP SM, although the overall magnitude remains moderate. For instance, RMSE values for KAN SM were mostly below 0.15 m^3/m^3, but the mean RMSE (0.11 m^3/m^3) is higher than that of SMAP SM (0.08 m^3/m^3). Encouragingly, particularly low ubRMSE values (mean < 0.04 m^3/m^3) were observed at the TxSon and NGARI stations, indicating high fidelity in capturing anomalies after

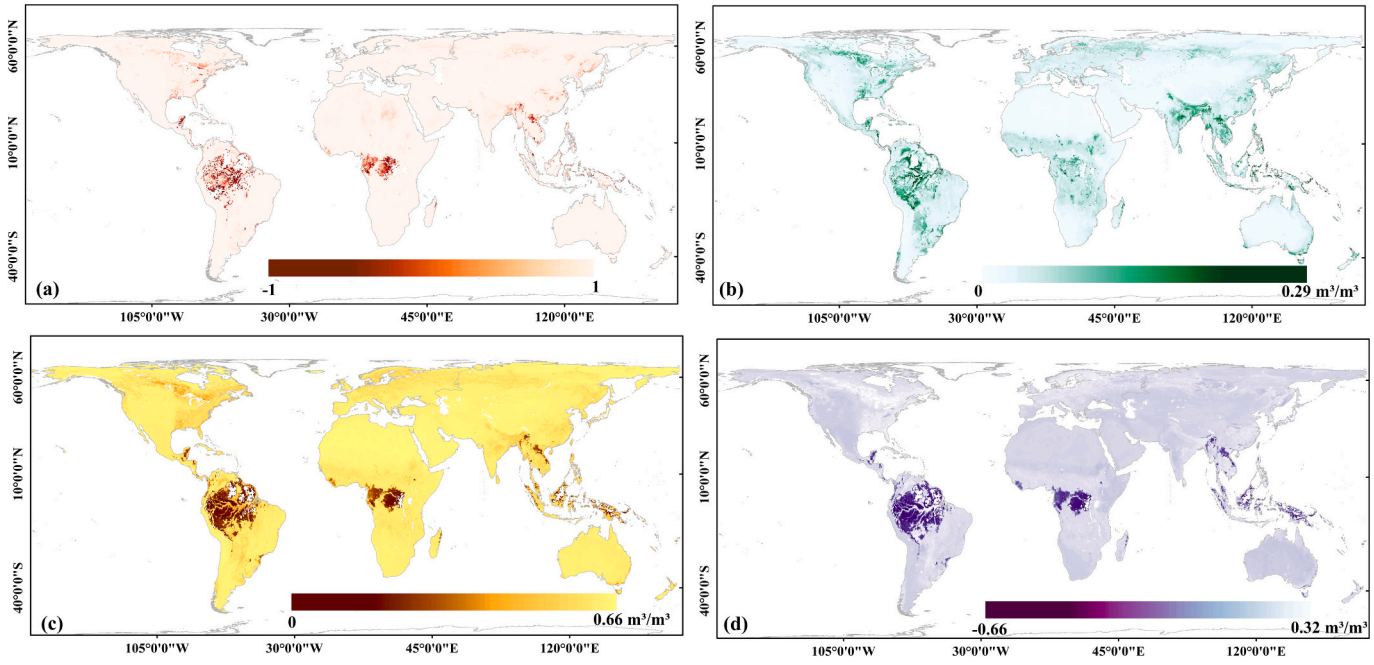

Fig. 6. Spatial distribution of temporal evaluation metrics between KAN SM and SMAP SM: (a) Pearson correlation coefficient (R), (b) ubRMSE, (c) RMSE, and (d) mean bias. All metrics are computed from the time series at each grid cell.

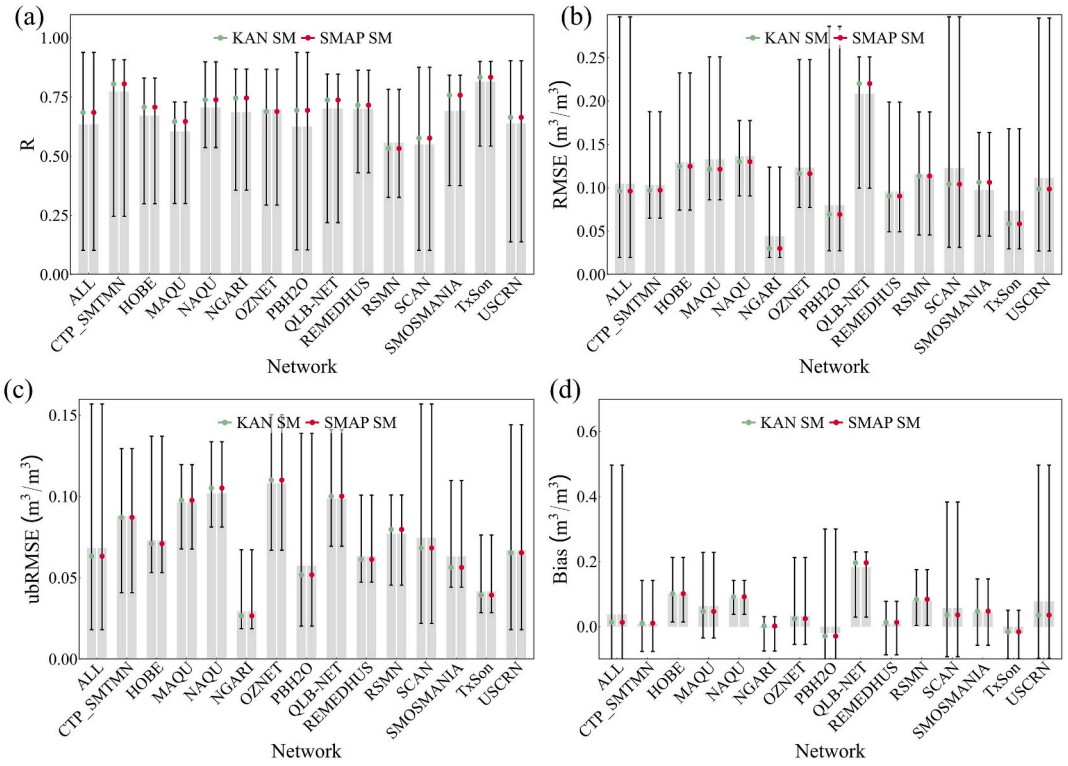

Fig. 7. Validation results of KAN SM and SMAP SM against in situ soil moisture measurements.

removing systematic bias. Both datasets exhibit predominantly positive biases, reflecting a tendency toward overestimation relative to in situ measurements, which may arise from retrieval uncertainty, representativeness/scale mismatch between point measurements and satellite footprints, and sub-pixel spatial heterogeneity. Although KAN SM does not show a consistent accuracy advantage over SMAP SM, its overall performance is still operationally satisfactory: the average ubRMSE (0.07 m^3/m^3) is below the commonly used application threshold of 0.08 m^3/m^3. More importantly, the KAN-based framework provides an explicit, formula-based TB–SM mapping (with auxiliary variables), improving transparency and facilitating rapid large-scale implementation, thereby supporting broader real-world applications.

5. Discussion

5.1. Analysis on model interpretability

Compared with traditional machine learning methods, a key advantage of the KAN model is its capacity to explicitly formulate mathematical expressions. As Section 4.1 shows, the interpretable SM retrieval process can be described as the following equation:

$$mv = \alpha\left(\frac{T_B}{T_S}e^{\tau\sec\theta} + \beta\right)^2 + \gamma e^{\delta\tau} + C. \tag{13}$$

In this study, we employed partial derivative analysis to systematically examine the sensitivity and influence patterns of T_B, T_S, and VOD (τ) on SM, thereby enhancing the interpretability of the model by explicitly revealing the functional roles of different input variables in shaping SM variability. As an example, we first analyze the partial derivative of SM with respect to T_B. The specific derivation result is as follows:

$$\frac{\partial mv}{\partial T_B} = \frac{2\alpha\left(\frac{T_B}{T_S}e^{\tau\sec\theta} + \beta\right)}{T_S}e^{\tau\sec\theta} \tag{14}$$

Since α is negative, this indicates that as T_B increases, mv decreases. This demonstrates an inverse relationship between the T_B signal and SM. Furthermore, as T_B increases, the magnitude of the partial derivative becomes larger, reflecting the more significant impact of T_B on SM.

A similar analysis of T_S is shown in Eq. (15):

$$\frac{\partial mv}{\partial T_S} = -\frac{2\alpha T_B e^{\tau\sec\theta}}{T_S^2}\left(\frac{T_B}{T_S}e^{\tau\sec\theta} + \beta\right). \tag{15}$$

Since the parameter α is negative, -2α is positive, indicating that the partial derivative of SM with respect to T_S is positive. In other words, SM and T_S are positively correlated. Although higher temperatures tend to enhance evapotranspiration and reduce SM, at the global scale, temperature increases are often accompanied by precipitation events, which strengthen the hydrological cycle and thereby exert an overall positive influence on SM (Seneviratne et al., 2010). From the perspective of the τ-ω model, T_S controls the effective emitting temperature of the soil layer that contributes to the observed brightness temperature, while SM primarily modulates the soil reflectivity and thus the emissivity term. In the RTM-generated training data, combinations of higher T_S and higher SM can produce similar brightness temperatures through this coupling of temperature and emissivity, so the learned explicit formula captures this co-variability as a positive $\frac{\partial mv}{\partial T_S}$. This behavior is therefore consistent with the combined effect of microwave emission physics and large-scale hydroclimatic co-variation, rather than reflecting only a local evaporative response.

Regarding τ, $\frac{\partial mv}{\partial t}$ is given by Eq. (16). The first term is negative due to α, while the latter term is positive due to the positive $\gamma\delta e^{\delta t}$. Therefore, τ has opposing effects on SM, but the positive effect from $\gamma\delta e^{\delta t}$ tends to dominate due to its exponential form, leading to an overall increase in mv.

$$\frac{\partial mv}{\partial \tau} = 2\alpha\sec\theta\left(\frac{T_B}{T_S}e^{\tau\sec\theta} + \beta\right)\frac{T_B}{T_S}e^{\tau\sec\theta} + \gamma\delta e^{\delta\tau}. \tag{16}$$

Previous research indicates that SM promotes vegetation growth (Engstrom et al., 2013; Li et al., 2022; Wang et al., 2020). This finding is

consistent with ecological understanding: sufficient SM supports vegetation growth, while increased vegetation cover can reduce direct soil evaporation and modify the near-surface microclimate, which may help retain moisture in the root zone and enhance SM persistence.

Additionally, we used the SMAP SM as an independent observational reference to examine whether the signs and spatial prevalence of the relationships implied by our theoretical derivations are supported by data. Specifically, we computed Spearman's rank correlation to characterize the monotonic associations between SM, and T_S, VOD and T_B across the SMAP record, and the results are summarized in Fig. 8. Regarding the relationship between SM and T_S, the majority (64%) of the SMAP data show a positive correlation between SM and T_S, which is broadly consistent with our theoretical analysis. For the relationship between SM and VOD, we find that more than 71% of the pixels display a positive correlation between SM and VOD. This supports the hypothesis in our model that higher vegetation density, represented by VOD, promotes soil moisture. When it comes to the relationship between SM and T_B, more than 86% of the pixels show a negative correlation between SM and T_B. Overall, these three relationships are highly consistent with our interpretability analysis, indicating that the mathematical formulation provides reliable and robust explanatory power at the global scale.

5.2. Error analysis

Based on the validation results against in situ measurements presented above, KAN SM exhibits notable discrepancies. Accordingly, this section examines the sources of these errors from two perspectives: (i) the performance of KAN SM across different land cover types, and (ii) the influence of model parameterization on retrieval accuracy.

5.2.1. Performance of KAN SM across different land cover types

To evaluate the performance of KAN SM across different land cover types, Fig. 9 presents a comparative analysis of KAN SM and SMAP SM over forests, shrublands, grasslands, croplands, and bare land. As shown in Fig. 9 (a), KAN SM exhibits a strong positive correlation with SMAP SM across all five land cover types, with correlation coefficients exceeding 0.98, and the correlations are statistically significant ($p < 0.001$), indicating robust agreement across land cover types. This indicates that KAN SM is capable of effectively capturing SM dynamics at the global scale, showing a high level of agreement with SMAP SM. However, notable differences in numerical accuracy are observed among the different land cover categories. Compared with SMAP SM, KAN SM exhibits larger errors in forested regions. As illustrated in Fig. 9 (b), the RMSE for forests reaches 0.16 m^3/m^3, substantially higher than that of the other four land cover types. Fig. 9 (d) further shows that KAN SM tends to underestimate SM in forests relative to SMAP SM, with a mean bias of −0.07 m^3/m^3, likely attributable to the relatively high SM values produced by SMAP SM retrievals in forested areas. In contrast, the errors across the remaining four land cover types are comparatively smaller, with croplands exhibiting a slightly higher ubRMSE of 0.03 m^3/m^3 relative to the other three categories. Overall, these results stratified by land-cover type indicate that KAN SM and the SMAP SM product exhibit consistently high temporal coherence, whereas the magnitude of retrieval errors varies markedly across land-cover classes. Forested regions emerge as the primary source of discrepancy, while non-forested classes generally exhibit low errors and limited bias. This pattern indicates that improving the representation of vegetation effects and related parameterizations is likely critical for further reducing uncertainties, particularly over dense-canopy areas, which we discuss in the next section.

5.2.2. Model and parameterization contributions to error

When evaluated against in situ measurements, KAN SM showed a higher ubRMSE compared to SMAP SM, indicating that directly retrieved SM still exhibits certain deviations from ground observations. One important factor contributing to this difference is the parameterization of the dielectric model. While SMAP retrieval relies on the RTM and solves SM through an iterative process that accounts for soil texture effects, our study employed the simplified Topp model to facilitate model construction, which inevitably introduced additional bias. Furthermore, the KAN model was trained using numerically simulated SM, VOD, and ST, which may contain interdependencies and stochastic variability, thereby introducing uncertainties when applied to SMAP TB.

Furthermore, across different land cover types, discrepancies were most pronounced in forested regions. KAN SM was substantially lower than SMAP SM, while SMAP SM in these areas was generally high (≈0.5–0.6 m^3/m^3), significantly exceeding the global average. This suggests that the differences cannot be attributed solely to the KAN model but are also closely linked to inherent overestimation in SMAP SM over forested areas. To further investigate the RTM error mechanism under dense vegetation, we simulated relationships among SM, ST, and TB under different VOD conditions (0.2, 0.4, 0.8) using the forward model (Fig. 10a). Results show that as VOD increases, the sensitivity of TB to SM decreases substantially, while vegetation emission contributes increasingly to TB, raising its overall value. Under these conditions, the cost function fails to ill-conditioned (Fig. 10c); the difference between simulated and observed TB does not approach zero even as SM increases, until SM reaches its texture-dependent upper limit. Consequently, RTM retrievals under high vegetation density tend to overestimate SM significantly. In contrast, under sparse vegetation (Fig. 10b), the cost function usually decreases then increases, converging near a minimum, enabling more accurate SM estimation. Overall, these results suggest that SMAP SM over forests may be prone to overestimation; therefore, although KAN SM is lower than SMAP SM in these regions, it may partially mitigate this systematic tendency.

5.3. Advantages, limitations, and future directions

In this study, we present a novel retrieval framework that integrates a radiative transfer model with machine learning to derive an SM retrieval algorithm in the form of an explicit mathematical formula. Conventional RTM-based approaches represent the microwave emission process through a chain of coupled physical sub-models (e.g., dielectric mixing, Fresnel reflectivity, surface roughness, and the τ-ω vegetation formulation). Although physically grounded, such schemes typically require many input parameters and are commonly implemented through iterative inversion, which increases computational cost and often obscures the final TB–SM relationship from direct inspection. The key advantage of the KAN model lies in its foundation on the Kolmogorov–Arnold representation theorem, which allows complex multivariate functions to be decomposed into a series of trainable univariate functions. These univariate functions are then combined through the network's activation functions into a single multivariate function, enabling a direct mapping from satellite observations to SM. This approach eliminates the need for iterative inversion, substantially improving computational efficiency, while the resulting symbolic formula retains a degree of physical interpretability, revealing meaningful relationships between SM and relevant variables. Although KAN shares a neural network architecture with MLPs, the KAN-based SM retrieval method proposed here outperforms MLPs in both accuracy and generalization. It also surpasses traditional white-box algorithms such as random forests, decision trees, and linear regression. Importantly, the symbolic formulas produced by KAN ensure stable and reproducible outputs, overcoming the inherent randomness often encountered in conventional machine learning models and effectively addressing the “black-box” limitation.

Like any machine learning approach, the proposed method still has certain limitations. First, the reliability of the KAN model largely depends on the training dataset, which is generated through a forward model and is designed to span a broad, physically plausible state space, whereas real-world variables are often interdependent and constrained. Moreover, the SMAP record is further shaped by observational sampling,

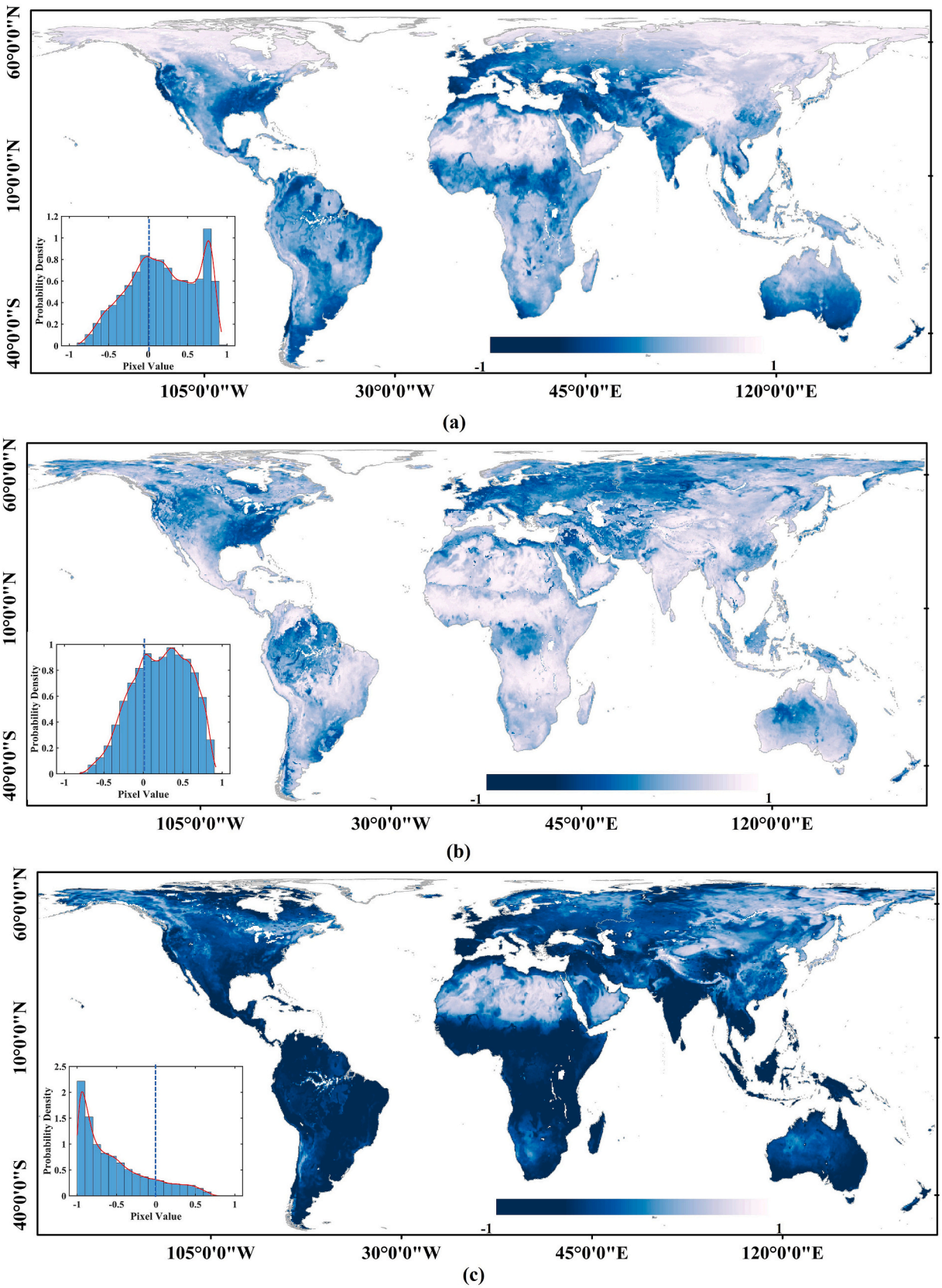

Fig. 8. Global-scale Spearman rank correlation coefficients between SMAP SM and various variables: (a) T_S, (b) VOD, and (c) T_B.

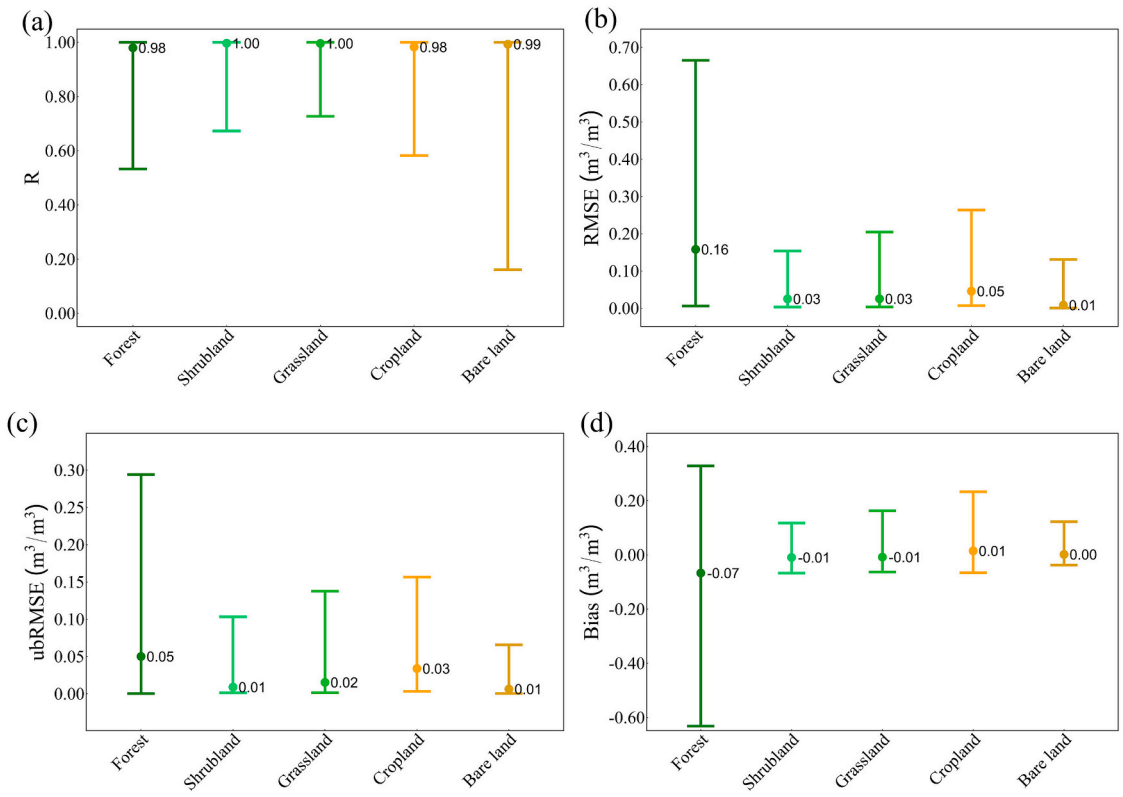

Fig. 9. Comparison of KAN SM and SMAP SM across different land cover types: (a) Correlation coefficient (R), (b) RMSE, (c) ubRMSE, and (d) Bias.

masks, and quality-control criteria, making it a constrained subset of this broader space. For instance, extreme conditions, such as very low SM coupled with unusually vigorous vegetation growth, may occur in the simulated scenarios, although such combinations rarely exist in natural environments due to the physiological and ecological limitations of vegetation under water stress. This mismatch introduces uncertainties that can affect retrieval accuracy. Second, the forward radiative transfer models used in this study, including the Topp dielectric model and the τ-ω vegetation model, have inherent limitations. For example, the Topp model does not account for soil texture effects on dielectric properties, and the τ-ω model is more suitable for low-stature vegetation. These factors may partly affect the training of the KAN model and the physical fidelity of the derived symbolic formulas. Furthermore, to simplify model construction, only single-polarization (V-polarization) TB data were used in this study, limiting the model's ability to fully exploit multi-channel microwave observations. H-polarization TB can serve as complementary input, while the dual-polarization Microwave Polarization Difference Index (MPDI) can partially reflect vegetation growth status and help correct the overestimation of SM by the τ-ω model under dense vegetation conditions (Park et al., 2024). These considerations indicate that, in future studies, integrating multi-channel TB data and incorporating additional physical constraints into the forward model may further enhance the accuracy and reliability of SM retrievals. Beyond SM, the proposed KAN-RTM framework also shows great potential for directly retrieving VOD. In the τ-ω model, VOD is the key parameter describing vegetation attenuation and emission. Therefore, the same strategy used here for SM—learning an explicit mapping among TB, ST, VOD, and other state variables under RTM constraints—can be extended to derive symbolic formulas for VOD as well. In a dual-channel configuration, KAN could be trained on simulated (or SMAP-like) TBH/TBV pairs to emulate a DCA-type joint retrieval, while still providing explicit expressions for both SM and VOD. Such extensions would enable the framework to jointly characterize soil and vegetation water status in an interpretable manner and represent a natural direction for future work.

Overall, our results demonstrate that integrating physically based modeling with an interpretable, formula-extracting neural architecture can provide a practical route to efficient and transparent quantitative retrievals. The framework is not inherently limited to SMAP and can be adapted to other microwave missions (e.g., SMOS, AMSR2, Fengyun-3) by adjusting sensor configuration and forward-model settings, thereby leveraging their multi-frequency and multi-polarization capabilities. With further refinements in forward-model parameterization and the incorporation of additional physical constraints and auxiliary datasets, the proposed approach may offer a more robust and scientifically interpretable alternative for remote sensing inversion of land-surface variables.

6. Conclusion

In this study, an interpretable soil moisture (SM) retrieval model was developed through the integration of the Kolmogorov-Arnold Network (KAN) architecture with the radiative transfer model (RTM). The effects of soil dielectric properties, surface roughness, and vegetation were considered to simulate SM and corresponding microwave brightness temperature (TB) using RTM, thereby generating diverse surface-type datasets. These datasets were subsequently utilized to train a KAN model, which leverages its powerful formula reconstruction capability to directly approximate complex nonlinear physical processes into interpretable mathematical expressions, thereby enabling the conversion of TB to SM.

To validate the feasibility of the proposed KAN-based model, it was applied to long-term SM retrieval using SMAP data. The accuracy of the retrieval results was evaluated through spatiotemporal analysis, comparisons with SMAP SM products, and validation with in-situ measurements. The results indicate that the retrieved SM effectively captures global SM dynamics and seasonal variations. Compared to SMAP official soil moisture products, the proposed method achieved a high correlation coefficient ($R = 0.98$) across various land cover types at the global scale.

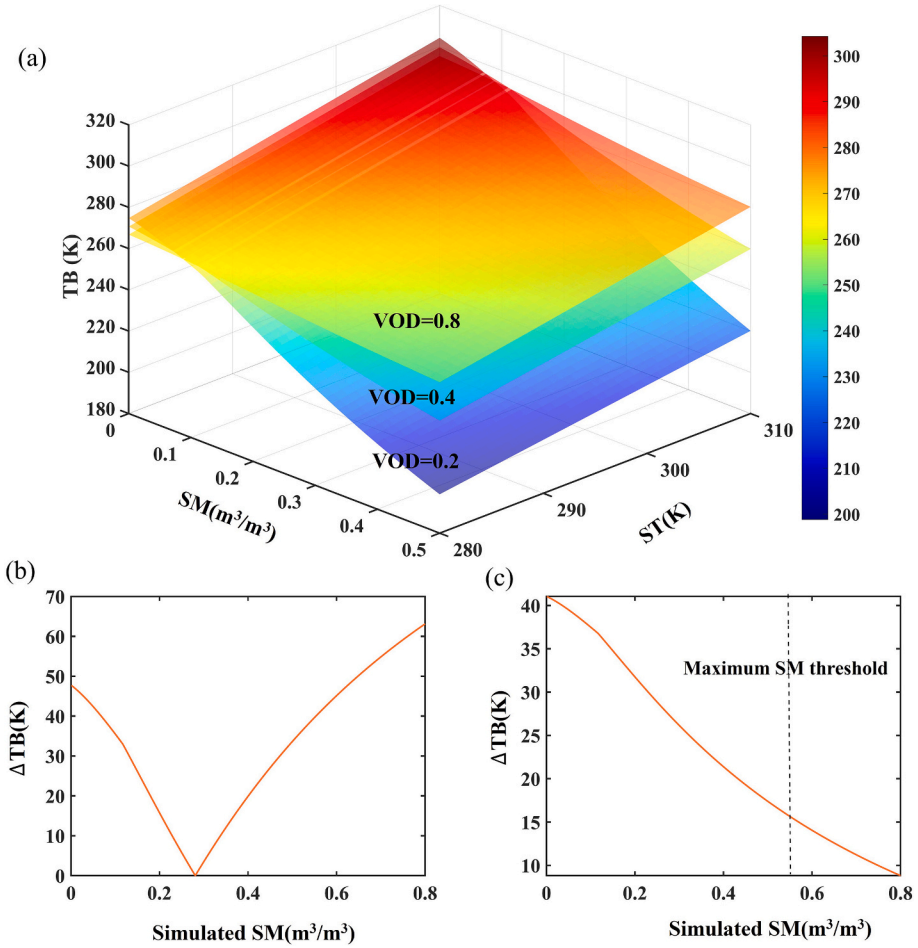

Fig. 10. Analysis of SM retrieval based on RTM under forested conditions: (a) Forward RTM-simulated relationships among SM, ST, and TB under different vegetation conditions. (b) Iterative function under normal conditions. (c) Iterative function under high vegetation conditions.

In addition, validation against ground-based measurements showed an average correlation of 0.64 and an ubRMSE of 0.07 m^3/m^3, further confirming the consistency of the retrieved SM with in situ observations. Finally, the interpretability analysis of the derived formulas through partial derivatives revealed that the proposed model effectively explains the physical relationships between SM and key influencing factors.

In summary, this study represents the first attempt to construct a globally interpretable SM retrieval formula using the KAN framework, demonstrating its feasibility and potential. Future research can further extend the application of KAN to the retrieval of other remote sensing parameters and facilitate the integration of multiple remote sensing models, thereby contributing to a deeper understanding of the physical mechanisms underlying satellite signals and surface parameters.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2026.115378.

CRediT authorship contribution statement

Yurun Chen: Writing – review & editing, Writing – original draft, Resources, Methodology. **Cheng Tong:** Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Formal analysis, Conceptualization. **Josh Qixuan Sun:** Writing – original draft, Methodology, Formal analysis. **Yulin Shangguan:** Visualization, Funding acquisition, Conceptualization. **Xiaodong Deng:** Validation, Supervision, Methodology. **Mark Crowley:** Writing – review & editing, Validation, Supervision. **Hongquan Wang:** Writing – original draft, Validation, Supervision. **Yang Ye:** Investigation, Funding acquisition, Conceptualization. **Haijun Bao:** Project administration, Methodology, Investigation. **Ruqi Huang:** Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (grant number 424B2005, grant number 42401420).

Appendix A

To further evaluate the generalization ability of the proposed KAN model, we compared its performance against several commonly used machine learning methods, including multilayer perceptron (MLP), decision trees (DT), linear regression (LR), and random forests (RF). For fairness, all models

were trained with the same input variables and dataset partitions as used in the KAN experiments. The results (Table A1) show that KAN consistently outperformed these baseline models in most land cover types, achieving higher correlation coefficients

Table A1
Performance comparison of KAN with other machine learning methods (MLP, DT, LR, RF) across 10 land cover types in the SMAP dataset.

Type	MLP	DT	LR	RF	KAN
1	−0.443	−0.118	−0.435	−0.118	0.010
2	0.975	0.975	0.946	0.976	0.985
3	0.883	0.279	0.743	0.280	0.983
4	0.987	0.971	0.957	0.972	0.991
5	0.993	0.989	0.958	0.991	0.993
6	0.997	0.998	0.997	0.998	0.995
7	0.967	0.980	0.953	0.981	0.981
8	1	1.000	1.000	1.000	1.000
9	0.968	0.949	0.936	0.951	0.979
10	0.972	0.984	0.985	0.985	0.985

MLP (Multilayer Perceptron): A feedforward neural network consisting of multiple layers of interconnected nodes. Each hidden layer applies nonlinear activation functions, enabling the model to approximate complex and nonlinear mappings between input features (e.g., TB, ST, VOD) and SM.

DT (Decision Tree): A hierarchical, rule-based model that splits the dataset into subsets based on feature thresholds. By following branches from root to leaf, the model generates predictions through a sequence of simple if–then rules, making it effective for capturing localized patterns in the data.

LR (Linear Regression): A statistical method that assumes a linear relationship between input variables and the target. It fits coefficients to minimize the difference between observed and predicted values. While computationally efficient, its ability to capture nonlinear dependencies is limited.

RF (Random Forest): An ensemble learning technique that constructs multiple decision trees using different subsets of the data and features. Predictions are obtained by aggregating outputs from all trees (majority vote or averaging), improving robustness and reducing overfitting compared to a single decision tree.

Daily SM was retrieved through the integration of SMAP TB, ST, and VOD data within the developed interpretable model at a spatial resolution of 36 km. The spatial distribution of retrieved SM across multiple temporal scales is illustrated in Fig. S1. Distinct spatial patterns were identified, consistent with global climatic and vegetation characteristics. To demonstrate the model's capability in capturing global SM patterns across diverse environmental conditions, we highlighted several representative regions in Fig. S1. Regions with humid climates and dense vegetation—such as the Congo basin, Eastern North America, and Southeast Asia—exhibited consistently elevated SM levels, aligning with expected hydrological behavior. Conversely, arid and semi-arid zones including the Sahara Desert, the Australian Outback, and the Middle East showed markedly lower SM values. Pronounced seasonal variations were observed in the retrieved SM results. In China's Yangtze River Plain, the subtropical monsoon climate was found to drive concentrated precipitation during spring and summer, leading to substantially higher SM values compared to autumn and winter. The Tibetan Plateau represents a unique case due to its high-altitude permafrost environment, where SM values during frozen periods (ST < 273 K) are physically suppressed and masked as invalid in our retrieval. However, during the thawing season (May to September), the eastern plateau region, characterized by lower elevation and denser vegetation, exhibited markedly higher SM levels compared to its western counterpart. The observed spatial and temporal patterns of SM distribution align strongly with established environmental principles and expected SM dynamics. The retrieved SM results exhibit spatial distributions that are consistent with broadly recognized geographic and climatic patterns. This alignment serves as indirect evidence of the method's reliability and its capability to reflect the intricate relationships among climatic conditions, vegetation cover, and SM dynamics.

The following table shows the complete lookup table for model complexity

	x	x^2	$1/x$	$log(x)$	e^x
Complexity	1	2	2	2	2

Data availability

No data was used for the research described in the article.

References

Ambadan, J., MacRae, H., Colliander, A., Tetlock, E., Helgason, W., Gedalof, Z.e., Berg, A., 2022. Evaluation of SMAP soil moisture retrieval accuracy over a boreal forest region. IEEE Trans. Geosci. Remote Sens. 60, 4414611.

Bozorgasl, Z., Chen, H., 2024. Wav-KAN: Wavelet Kolmogorov-Arnold Networks.

Brocca, L., Tarpanelli, A., Filippucci, P., Dorigo, W., Zaussinger, F., Gruber, A., Fernández-Prieto, D., 2018. How much water is used for irrigation? A new approach exploiting coarse resolutionsatellite soil moisture products. Int. J. Appl. Earth Obs. Geoinf. 73, 752–766.

Chai, L., Zhu, Z., Shaomin, L., Xu, Z., Jin, R., Li, X., Kang, J., Che, T., Zhang, Y., Zhang, J., Cui, H., Gao, T., Xu, T., Zhao, S., Pan, X., Guo, G., 2024. QLB-NET: a dense soil moisture and freeze–thaw monitoring network in the Qinghai Lake Basin on the Qinghai–Tibetan Plateau. Bull. Am. Meteorol. Soc. 105, E584–E604.

Dorigo, W.A., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Drusch, M., S, M., Oevelen, P., Robock, A., Jackson, T., 2011. The international soil moisture network: a data hosting facility for global in situ soil moisture measurements. Hydrol. Earth Syst. Sci. 15, 1675–1698.

Engstrom, R., Hope, A., Kwon, H., Stow, D., 2013. The relationship between soil moisture and NDVI near Barrow, Alaska. Phys. Geogr. 29, 38–53.

Entekhabi, D., Njoku, E., O'Neill, P., Kellogg, K.H., Crow, W., Edelstein, W.N., Entin, J., Goodman, S., Jackson, T., Johnson, J., Kimball, J., Piepmeier, J., Koster, R., Martin, N., McDonald, K., Moghaddam, M., Moran, S., Reichle, R., Shi, J., van Zyl, J., 2010. The soil moisture active and passive (SMAP) mission. Proc. IEEE 98, 704–716.

Friedl, M.A., McIver, D., Hodges, J., Zhang, X., Muchoney, D., Strahler, A.H., Woodcock, C., Gopal, S., Schneider, A., Cooper, A., Baccini, A., Gao, F., Schaaf, C., 2002. Global land cover mapping from MODIS: algorithms and early results. Remote Sens. Environ. 83, 287.

Granata, F., Zhu, S., Di Nunno, F., 2024. Advanced streamflow forecasting for central European Rivers: the cutting-edge Kolmogorov-Arnold networks compared to transformers. J. Hydrol. 645, 132175.
Humphrey, V., Berg, A., Ciais, P., Gentine, P., Jung, M., Reichstein, M., Seneviratne, S., &Frankenberg, C., 2021. Soil moisture–atmosphere feedback dominates land carbon uptake variability. Nature 592, 65–69.
Imaoka, K., Kachi, M., Fujii, H., Murakami, H., Hori, M., Ono, A., Igarashi, T., Nakagawa, K., Oki, T., Honda, Y., Shimoda, H., 2010. Global change observation mission (GCOM) for monitoring carbon, water cycles, and climate change. Proc. IEEE 98, 717–734.
Karthikeyan, L., Pan, M., Wanders, N., Kumar, D.N., Wood, E.F., 2017. Four decades of microwave satellite soil moisture observations: part 1. A review of retrieval algorithms. Adv. Water Resour. 109, 106–120.
Kerr, Y.H., Waldteufel, P., Wigneron, J.P., Delwart, S., Cabot, F., Boutin, J., Escorihuela, M.J., Font, J., Reul, N., Gruhier, C., Juglea, S.E., Drinkwater, M.R., Hahne, A., Martín-Neira, M., Mecklenburg, S., 2010. The SMOS Mission: new tool for monitoring key elements ofthe global water cycle. Proc. IEEE 98, 666–687.
Kolassa, J., Reichle, R.H., Liu, Q., Alemohammad, S.H., Gentine, P., Aida, K., Asanuma, J., Bircher, S., Caldwell, T., Colliander, A., Cosh, M., Collins, C.H., Jackson, T.J., Martinez-Fernandez, J., McNairn, H., Pacheco, A., Thibeault, M., Walker, J.P., 2018. Estimating surface soil moisture from SMAP observations using a neural network technique. Remote Sens. Environ. 204, 43–59.
Konings, A., McColl, K., Piles, M., Entekhabi, D., 2015. How many parameters can be maximally estimated from a set of measurements? IEEE Geosci. Remote Sens. Lett. 12, 1081–1085.
Lee, J., Sun, X., Errington, E., Guo, M., 2024. A KAN-based Interpretable Framework for Process-informed Prediction of Global Warming Potential.
Lee, J., Im, J., Son, B., Cosio, E., Salinas, N., 2025. Improved SMAP soil moisture retrieval using a deep neural network-based replacement of radiative transfer and roughness model. IEEE Trans. Geosci. Remote Sens. 63, 4513119.
Li, W., Migliavacca, M., Forkel, M., Denissen, J., Reichstein, M., Yang, H., Duveiller, G., Weber, U., Orth, R., 2022. Widespread increasing vegetation sensitivity to soil moisture. Nat. Commun. 13, 1234567890.
Li, Z., Yang, Q., Li, J., Jin, T., Yuan, Q., Shen, H., Zhang, L., 2025. Global multi-scale surface soil moisture retrieval coupling physical mechanisms and machine learning in the cloud environment. Remote Sens. Environ. 329, 114928.
Liu, S., Zhu, Z., Xu, Z., Jin, R., Chai, L., 2024. In: Dataset of Tianjun dense soil moisture and Freeze–Thaw monitoring network in the Qinghai Lake Basin (QLB-NET)(2019-2023). National Tibetan Plateau / Third Pole Environment Data Center.
Ma, H., Zeng, J., Chen, N., Zhang, X., Cosh, M.H., Wang, W., 2019. Satellite surface soil moisture from SMAP, SMOS, AMSR2 and ESA CCI: a comprehensive assessment using global ground-based observations. Remote Sens. Environ. 231, 111215.
Mao, K., Wang, H., Shi, J., Heggy, E., Wu, S., Bateni, S., Du, G., 2023. A general paradigm for retrieving soil moisture and surface temperature from passive microwave remote sensing data based on artificial intelligence. Remote Sens. 15, 1793.
McColl, K.A., Alemohammad, S.H., Akbar, R., Konings, A.G., Yueh, S., Entekhabi, D., 2017. The global distribution and dynamics of surface soil moisture. Nat. Geosci. 10, 100–104.
Mo, T., Choudhury, B., Schmugge, T., Jackson, T.J., 1982. A model for microwave emission from vegetation-covered fields. J. Geophys. Res. Atmos. 87, 11229–11237.
Njoku, E., li, l., 1999. Retrieval of land surface parameters using passive microwave measurements at 6-18 GHz. IEEE Trans. Geosci. Remote Sens. 37, 79–93.
O'Neill, P., Bindlish, R., Chan, S., Chaubell, J., Njoku, E., Jackson, T., 2020. Algorithm Theoretical Basis Document. Level 2 & 3 Soil Moisture (Passive) Data Products.
Park, C.-H., Jagdhuber, T., Colliander, A., Berg, A., Cosh, M., Lee, J., Boo, K.-O., 2024. Retrieving forest soil moisture from SMAP observations considering a microwave polarization difference index (MPDI) to -ω model. Sci. Remote Sens. 9, 100131.
Peng, J., Loew, A., Merlin, O., Verhoest, N., 2017. A review of spatial downscaling of satellite remotely sensed soil moisture. Rev. Geophys. 55, 341–366.
Qu, Y., Zhu, Z., Chai, L., Liu, S., Montzka, C., Liu, J., Yang, X., Lu, Z., Jin, R., Li, X., Guo, Z., Zheng, J., 2019. Rebuilding a microwave soil moisture product using random forest adopting AMSR-E/AMSR2 brightness temperature and SMAP over the Qinghai–Tibet Plateau, China. Remote Sens. 11, 683.
Seneviratne, S., Corti, T., Davin, E., Hirschi, M., Jaeger, E., Lehner, I., Orlowsky, B., Teuling, A., 2010. Investigating soil moisture-climate interactions in a changing climate: a review. Earth Sci. Rev. 99, 125–161.
Tong, C., Deng, X., Shangguan, Y., Dong, B., Chen, Y., Huang, C., Zhu, L., Li, S., Ye, Y., Wang, H., 2025. The passive microwave remote sensing in soil moisture retrieval: products, models, applications and challenges. Int. Soil Water Conserv. Res. 13, 843–859.
Topp, G.C., Davis, J.L., Annan, P., 1980. Electromagnetic determination of soil water content: measurements in coaxial transmission lines. Water Resour. Res. 16, 574–582.
Wang, H., Magagi, R., Goita, K., Wang, K., 2020. Soil moisture retrievals using ALOS2-ScanSAR and MODIS synergy over Tibetan plateau. Remote Sens. Environ. 251, 112100.
Wigneron, J.P., Jackson, T.J., O'Neill, P., De Lannoy, G., de Rosnay, P., Walker, J.P., Ferrazzoli, P., Mironov, V., Bircher, S., Grant, J.P., Kurum, M., Schwank, M., Munoz-Sabater, J., Das, N., Royer, A., Al-Yaari, A., Al Bitar, A., Fernandez-Moran, R., Lawrence, H., Mialon, A., Parrens, M., Richaume, P., Delwart, S., Kerr, Y., 2017. Modelling the passive microwave signature from land surfaces: A review of recent results and application to the L-band SMOS & SMAP soil moisture retrieval algorithms. Remote Sens. Environ. 192, 238–262.
Yuan, Q., Xu, H., Li, T., Shen, H., Zhang, L., 2020. Estimating surface soil moisture from satellite observations using a generalized regression neural network trained on sparse ground-based measurements in the continental U.S. J. Hydrol. 580, 124351.
Zhang, C., Zeng, J., Shi, P., Ma, H., Letu, H., Zhang, X., Wang, P., Bi, H., Rong, J., 2025. Global-scale gap filling of satellite soil moisture products: methods and validation. J. Hydrol. 653, 132762.
Zhao, J., Sima, O., 2024. A nonlinear Split-window algorithm for retrieving land surface temperatures from Fengyun-4B thermal infrared data. IEEE Trans. Geosci. Remote Sens. 62, 5000612.