https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

Deep Learning-Augmented Soil Moisture Data Assimilation under Temporally Sparse and Irregular Observations: A Case Study over the Tibetan Plateau

Yajie Zhu[1,2], Jinliang Hou[1], Chunlin Huang[1,3], Weizhen Wang[1], Ying Zhang[1], Yuanhong You[4]

[1]Heihe Remote Sensing Experimental Research Station, State Key Laboratory of Cryospheric Science and Frozen Soil Engineering, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China
[2]University of Chinese Academy of Sciences, Beijing, 100094, China
[3]Faculty of Geomatics, Lanzhou Jiaotong University, Lanzhou 730070, China
[4]School of Geography and Tourism, Anhui Normal University, Wuhu 241002, China

Correspondence to: Jinliang Hou (jlhours@lzb.ac.cn)

Abstract. Soil moisture (SM) data assimilation is often challenged by temporally sparse and irregular observations, particularly in data-scarce regions such as the Tibetan Plateau, where observational gaps and uneven sampling frequencies can substantially degrade assimilation performance. Conventional approaches, including the ensemble Kalman filter (EnKF), require repeated covariance estimation and matrix inversion, leading to considerable computational cost and reduced effectiveness under intermittent observation conditions. To address these limitations, we develop a deep learning-augmented SM data assimilation framework that explicitly accounts for temporal irregularity while improving computational efficiency. The framework consists of two complementary components. First, a fully connected neural network (FCNN) is trained to emulate the nonlinear analysis-update process of EnKF, thereby reducing computational overhead by avoiding explicit covariance calculations. Second, a gated recurrent unit with decay mechanism (GRU-D) is introduced to capture temporal dependencies from irregularly sampled observations through explicit incorporation of elapsed-time information. The framework is evaluated using observations from eight sites across the Tibetan Plateau. Results show that the proposed approach consistently outperforms the conventional EnKF under sparse and irregular observation scenarios. Across all sites, the framework reduces RMSE by more than 60% and increases correlation coefficients by over 80% relative to EnKF, while reducing computational time by more than 30%. The largest improvements are observed under highly irregular observation conditions, demonstrating the importance of explicitly accounting for temporal sparsity in data assimilation. These findings highlight the potential of deep learning to enhance both the robustness and efficiency of soil moisture data assimilation and provide a practical solution for hydrological state estimation in observation-limited environments.

Keywords: Soil moisture; Data assimilation; Deep learning-augmented assimilation; Ensemble Kalman Filter; Temporal irregularity

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

1 Introduction

As a critical state variable in land surface processes, soil moisture (SM) plays a fundamental role in regulating land–atmosphere interactions by controlling evapotranspiration, energy exchange, and hydrological partitioning (Seneviratne et al., 2010). Through its influence on precipitation infiltration, runoff generation, root-zone water availability, and biogeochemical cycling, SM affects a wide range of environmental and hydrological processes (Legates et al., 2011; Liu et al., 2024). Consequently, accurate characterization of SM spatiotemporal dynamics is essential for a wide range of applications, including weather forecasting, drought monitoring, agricultural irrigation, and water resources management (Li et al., 2023; Zhu et al., 2025). However, reliable estimation of SM remains challenging because in situ observations are often spatially sparse and temporally irregular, while remote sensing retrievals may suffer from limited temporal coverage and retrieval uncertainties. These observational constraints hinder the continuous characterization of SM dynamics and necessitate the use of physically based models to provide spatially and temporally continuous estimates of soil moisture states.

Land Surface Models (LSMs) serve as the primary tools for simulating the spatiotemporal evolution of SM by explicitly representing land surface hydrological and energy exchange processes. Over the past several decades, LSMs have evolved from simplified bucket-type formulations (Manabe, 1969) to advanced Earth system components that couple water, energy, and carbon cycles, and are widely integrated into climate and hydrological modeling systems (Dai et al., 2003; Lawrence et al., 2019; He et al., 2023). Despite these substantial advances, the predictive performance of LSMs remains limited due to mismatches between model representations and real-world land surface heterogeneity, as well as scale inconsistencies between model structures and natural processes (Li et al., 2022). These fundamental mismatches manifest within modeling systems as multiple forms of uncertainty, including structural deficiencies in process representation, parameter sensitivities, uncertainties in initial conditions, and errors in meteorological forcing data (Huang et al., 2016; Huang et al., 2023), all of which propagate through simulations and degrade soil moisture estimates. These limitations are particularly evident in complex and heterogeneous environments. For instance, in the Noah LSM, underestimation of the thermal roughness length leads to biases in surface temperature and sensible heat flux simulations over arid regions of the Tibetan Plateau (Chen et al., 2015). At larger scales, deficiencies in representing deep root-zone soil moisture processes in Earth system modeling frameworks can result in biased evapotranspiration responses under moisture-limited conditions (Dong et al., 2022). In addition, unresolved sub-grid-scale heterogeneity and complex physical processes remain difficult to parameterize explicitly, further limiting model fidelity in representing land surface dynamics (Fisher and Koven, 2020). Collectively, these uncertainties indicate that LSM-based simulations alone are insufficient to fully constrain soil moisture evolution, as structural and parametric limitations cannot be effectively eliminated through calibration or model refinement alone (Li et al., 2022). Consequently, there is a pressing need to incorporate observational information to systematically reduce model uncertainty and improve the reliability of soil moisture estimation (MacBean et al., 2022).

To reduce these uncertainties, data assimilation (DA) techniques have been widely adopted in hydrological and land surface modeling. DA provides a rigorous framework for optimally combining model simulations with observations to estimate system

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

states and reduce uncertainties arising from imperfect model structures, parameters, and forcing data (Kalnay, 2007; Carrassi et al., 2018). Over the past decades, DA has become a core methodology in hydrological forecasting and land surface state estimation. Representative approaches include variational methods, such as three-dimensional variational assimilation (3D-Var) and four-dimensional variational assimilation (4D-Var), as well as sequential filtering techniques, including the Ensemble Kalman Filter (EnKF) and Particle Filter (PF) (McLaughlin, 1995; Reichle, 2001). Among these methods, EnKF has gained widespread popularity because it estimates background error covariance using an ensemble of model realizations, thereby avoiding the explicit construction of tangent linear and adjoint models required by variational approaches (Chen et al., 2017; Houtekamer and Mitchell, 2005). This characteristic makes EnKF particularly suitable for high-dimensional and nonlinear land surface systems, and it has been extensively applied in SM DA (Burgers et al., 1998). Previous studies have demonstrated its ability to improve SM estimation by effectively integrating model forecasts with observations (Reichle et al., 2002). Nevertheless, EnKF remains subject to important limitations. Its covariance-based update mechanism relies on Gaussian assumptions and linear error propagation, which may lead to degraded performance under strongly nonlinear conditions and complex error structures (Clark et al., 2008). Moreover, repeated covariance estimation and matrix inversion can become computationally demanding in large-scale applications. Importantly, the effectiveness of EnKF and other DA approaches depends not only on the characteristics of the assimilation algorithm itself, but also on the availability and temporal continuity of observations used to constrain model states.

In practice, observational limitations constitute another major challenge for SM DA. SM observations are frequently affected by missing records, intermittent measurements, and irregular sampling intervals, resulting in temporally discontinuous observational time series. Such limitations may arise from sensor malfunctions, communication interruptions, harsh environmental conditions, and restricted observation frequencies for both ground-based and satellite observations (Bradley et al., 2007; Turlapaty et al., 2010). Consequently, the amount of effective information available for constraining model states is substantially reduced, weakening the corrective role of observations within the assimilation process. Under these conditions, DA systems become increasingly dependent on model forecasts, which can amplify the influence of prior model biases and uncertainty in the analysis estimates. This issue is particularly critical for SM because its temporal dynamics can vary rapidly in response to precipitation, evapotranspiration, and land-atmosphere interactions. When observations are sparse or irregularly distributed in time, important temporal variations may not be adequately captured, leading to reduced assimilation accuracy and stability. Therefore, effectively exploiting incomplete and temporally irregular observations has become a fundamental challenge in land surface data assimilation and calls for new approaches capable of extracting meaningful temporal information from sparse and irregular observation sequences (Fan et al., 2022; Ding et al., 2025).

To address the two key challenges identified in soil moisture data assimilation systems, namely, the methodological limitations of traditional EnKF-based approaches and the constraints imposed by temporally sparse and irregular observations, deep learning (DL) has recently emerged as a promising complement to conventional land surface modeling and data assimilation framework (Fisher & Koven, 2020). However, purely data-driven approaches often require large amounts of high-quality labeled data and lack explicit physical constraints, making them difficult to apply as standalone replacements for process-

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

based models (Brajard et al., 2021; Li & Liu, 2025). As a result, hybrid DA frameworks that integrate data-driven learning components with physically based assimilation workflows have attracted increasing attention (Li et al., 2023), including approaches that learn surrogate representations of nonlinear operators, improve the estimation of error statistics, and enhance the handling of incomplete and irregular observations, while preserving physical consistency and computational efficiency (Boucher et al., 2020). From the perspective of addressing methodological limitations in EnKF-based assimilation, DL has been explored as a surrogate for computationally expensive components within assimilation workflows. By learning nonlinear mappings between model forecasts, observations, and analysis states, neural networks can reduce the need for explicit covariance estimation and matrix operations while preserving key statistical relationships required for state updating. For example, Chattopadhyay et al. (2023) proposed a hybrid EnKF framework in which a neural network surrogate is used to improve the estimation of background error statistics during state evolution. Cai et al. (2024) further developed a deep assimilation network (DANet) that learns mappings among model forecasts, observations, and true states, enabling efficient assimilation under sparse observational conditions. From the perspective of observational limitations, SM DA is fundamentally constrained by temporally sparse observations. Reichle et al. (2017) demonstrated the assimilation of SMAP satellite observations within an EnKF-based land surface modeling framework, highlighting the capability of DA systems to partially compensate for the temporal sparsity and discontinuity of SM observations. Despite these advances, existing studies have primarily focused on either improving assimilation algorithms or enhancing the utilization of observational information in isolation. Although DL has shown potential in learning from sparse and irregular observational data in broader data-driven and hybrid modeling contexts, its application to explicitly address temporally irregular SM observations within land surface DA frameworks remains limited. Consequently, there is still a need for approaches that can simultaneously alleviate the computational and nonlinear limitations of conventional data assimilation methods and effectively exploit temporally sparse and irregular observations while maintaining physical consistency and robustness (Li & Liu, 2025).

Building on above context, this study proposes a DL-augmented DA framework for SM estimation under temporally sparse and irregular observations. The proposed framework is designed to address both the computational limitations of conventional EnKF systems and the challenges associated with incomplete and irregular observational time series. A fully connected neural network (FCNN) is employed to learn the nonlinear mapping between model forecasts, observations, and analysis states, providing a surrogate representation of the EnKF analysis update step. In this way, the FCNN avoids explicit covariance estimation and matrix inversion while preserving the functional relationship between system states and observational constraints, thereby improving the computational efficiency of the assimilation process. In parallel, a gated recurrent unit with decay (GRU-D) is adopted to model temporal dependencies under missing and irregularly sampled observations. By explicitly encoding time intervals and missingness patterns, the GRU-D module provides a continuous representation of observational information, enabling more effective utilization of temporally sparse SM observations. The proposed framework integrates these two components in a complementary manner. The FCNN component focuses on improving the efficiency of the assimilation update step, while the GRU-D component enhances the representation of temporally irregular observational inputs. Together, they enable robust SM state estimation under conditions of limited, missing, and irregular observations.

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

The remainder of this paper is organized as follows. Section 2 describes the study area and the datasets employed. Section 3 details the proposed framework, experimental design and evaluation metrics. Section 4 presents and analyzes the results of assimilation experiments. Section 5 provides a discussion of the findings, including their implications and limitations. Finally, Section 6 concludes the study and outlines future research directions.

2 Study Area and Data

2.1 Study Area

The Tibetan Plateau (73°19′~104°47′ E, 26°00′~39°47′ N), characterized by its vast and complex terrain, plays a critical role in the Eurasian climate system and serves as a crucial driver for the Asian monsoon (Zhang et al., 2015). Its pronounced topographic heterogeneity and diverse climatic conditions make it an ideal region for investigating SM dynamics under complex environmental. To capture SM variability across different land surface conditions and to evaluate the proposed DA framework, eight representative observation sites were selected (Fig.1 and Table 1), distributed across the central, eastern, and western Tibetan Plateau. These sites span a range of elevations and soil texture conditions and represent two dominant land cover types in the region, including grasslands and barren or sparsely vegetated surfaces, thereby providing a representative characterization of SM dynamics across the study area. Based on land cover classification data (Zhu, 2022), the selected sites can be broadly categorized into two dominant surface types: grassland (Maqu, Yeniugou, Naqu, and Maduo) and barren or sparsely vegetated areas (Golmud, Tsedang, Shiquanhe, and Tuotuohe). This diversity in underlying surface conditions provides a robust basis for evaluating the performance and generalizability of the proposed framework under heterogeneous environmental conditions.

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

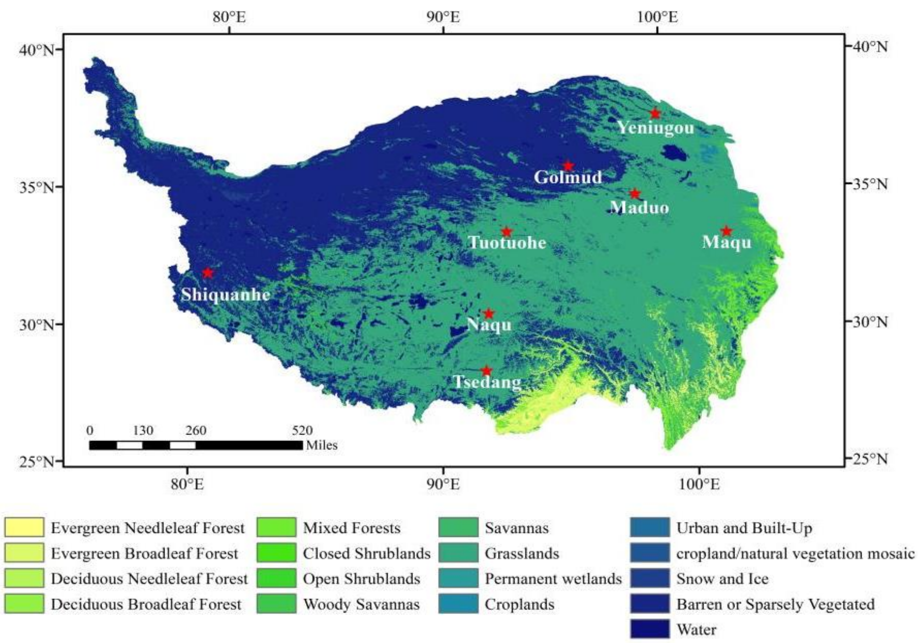


Figure 1: Land-cover-type map of the study area and the location of the eight sites on the Tibetan Plateau.

Table 1. Summary of eight sites and their environmental characteristics

	Coordinates	Elevation (m)	Land Cover	Soil Texture
Maqu	102°08'E, 33°51'N	3474	Grasslands	Clay
Yeniugou	99°34'E, 38°25'N	3320	Grasslands	Loam
Naqu	91°54'E, 31°22'N	4509	Grasslands	Sandy Loam
Maduo	98°22'E, 34°92'N	4272	Grasslands	Sandy Loam
Golmud	94°91'E, 36°42'N	2805	Barren or Sparsely Vegetated	Playa
Tsedang	91°46'E, 29°16'N	3560	Barren or Sparsely Vegetated	Loam
Shiquanhe	80°05'E, 32°30'N	4278	Barren or Sparsely Vegetated	Loam
Tuotuohe	92°44'E, 34°22'N	4533	Barren or Sparsely Vegetated	Sandy Loam

2.2 Data

2.2.1 Forcing Dataset

In this study, the high-resolution near-surface meteorological dataset for the Third Pole region (TPMFD) was employed as the input forcing data to drive the Noah-MP model. The dataset provides seven key forcing variables, including air temperature,

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

surface pressure, relative humidity, wind speed, downward shortwave radiation, downward longwave radiation, and precipitation rate. These variables are available at an hourly temporal resolution and a spatial resolution of 1/30° (Yang et al., 2023). TPMFD offers long-term, high-quality gridded forcing data covering China since 1979, supporting its use in land surface and hydrological modeling applications. Previous evaluations (Jiang et al., 2023, 2025) have shown that most TPMFD variables outperform several widely used reanalysis products, such as ERA5, ERA5-Land, and the Global Land Data Assimilation System (GLDAS), demonstrating its reliability as a high-resolution forcing dataset. In this study, the land surface model was driven at an hourly time step using the TPMFD forcing.

2.2.2 Land surface vegetation Parameters

In addition to the meteorological forcing data, two satellite-derived products were used as vegetation parameters in the land surface model to represent vegetation dynamics. Specifically, Leaf Area Index (LAI) was obtained from the MODIS Aqua MYD15A2H product (Ranga et al., 2015), while Fractional Vegetation Cover (FVC) was derived from the Global Land Surface Satellite (GLASS) product developed by Beijing Normal University (Jia et al., 2019). These two remotely sensed datasets provide complementary information on vegetation structure and are used to characterize vegetation conditions in the model.

2.2.3 Soil Moisture

In this study, the SMCI1.0 dataset (Soil Moisture of China by in situ data, version 1.0) was employed as the reference for validating SM estimates. This dataset, available from the National Tibetan Plateau Data Center (https://data.tpdc.ac.cn/home), was developed using a Random Forest model trained on 1789 in situ stations and provides spatiotemporally continuous daily soil moisture estimates at a 1 km spatial resolution for the period 2000-2020. Previous evaluations (Li et al., 2022) have demonstrated the high accuracy of SMCI1.0 dataset, with unbiased root mean square errors (ubRMSE) ranging from 0.041 to 0.052 and Spearman correlation coefficients (R) from 0.883 to 0.919 at the temporal scale, and ubRMSE values of 0.045 to 0.051 and R values of 0.866 to 0.893 at the spatial scale. By effectively mitigating the data gaps and temporal discontinuities inherent in raw station observations, SMCI1.0 provides a reliable and spatially continuous reference for evaluating soil moisture dynamics. SM data at four depths (10, 30, 60, and 100 cm) were extracted from SMCI1.0 and used as benchmark values for assessing the performance of DA experiments.

3 Methodology

3.1 Land surface model

Noah-MP is an augmented version of the Noah land surface model with multi-physics options (Niu et al., 2011; Yang et al., 2011; He et al., 2023). Currently maintained and released by the National Center for Atmospheric Research (NCAR), its latest version, v5.0, was released in March 2023, focusing on adopting modern Fortran standards and a modular architecture to enhance model scalability and interoperability. Physically, the model employs a tile scheme that divides each grid cell into separate vegetation canopy and bare soil components to solve energy and water processes independently. Fluxes and states are then area-weighted based on the vegetation fraction, while radiative transfer is handled by an improved two-stream scheme to characterize canopy shading and gap effects.

In this study, the Noah-MP model was employed to conduct dynamic simulations of SM. The model was configured in offline mode, with the soil column discretized into four layers at thicknesses of 10, 30, 60, and 100 cm from the surface downward. To ensure experimental consistency and reduce variability arising from differences in model physics, a uniform set of physical parameterization schemes was applied across all simulation runs within the Noah-MP land surface model framework. This configuration helps to minimize the confounding influence of internal parameterization choices, thereby enhancing the comparability of simulated soil moisture responses across experiments. As a result, differences in the simulation outputs can be more confidently attributed to variations in meteorological forcing data, although residual uncertainties may still persist due to model initialization, parameter specification, and inherent structural limitations. Detailed model configurations and the rationale for the selected parameterization schemes are provided in Section 3.4.

3.2 Data assimilation and learning methods

3.2.1 Ensemble Kalman filter (EnKF)

The EnKF is a widely used sequential data assimilation method that is particularly effective for nonlinear systems due to its ensemble-based approximation of error statistics (Evensen, 2003). In hydrological studies, it has been widely used to improve model simulations through state correction (Sun et al., 2016). In this study, the EnKF-derived analysis is used as a reference assimilation benchmark to assess the performance of the proposed framework. Furthermore, it is used as supervisory information to train the FCNN, which learns the nonlinear mapping from model states to EnKF-based analysis estimates. In the EnKF, the forward evolution of each ensemble member can be expressed as follows:

$$\mathbf{x}_{i,t+1}^{f} = \mathbf{M}(\mathbf{x}_{i,t}^{a}, \boldsymbol{\alpha}_{i,t}, \boldsymbol{\beta}_{i,t}) \tag{1}$$

Where $\mathbf{M}(\cdot)$ denotes the model operator (Noah-MP in this study) used to forecast the state variables (i.e., SM). $\mathbf{x}_{i,t+1}^{f}$ represents the forecast state variables (model simulations) of the i-th ensemble member at time t+1, while $\mathbf{x}_{i,t}^{a}$ is the ith ensemble member of analysis state variables (after update/assimilation) at time t. The terms $\boldsymbol{\alpha}_{i,t}$ and $\boldsymbol{\beta}_{i,t}$ represent the meteorological forcing factors and parameters, respectively.

During the evolution of the dynamic system, when the SM observation $\mathbf{Y}_{t+1}$ becomes available at time t+1, a random perturbation $\mathbf{v}_i$ is first added to the original observation to generate the perturbed observation ensemble and account for the error associated with the observation.

$$\mathbf{Y}_{i,t+1} = \mathbf{Y}_{t+1} + \mathbf{v}_i \tag{2}$$

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

Then, the optimal estimate of SM at time t+1 can be obtained by weighing the model simulations and the observations. The specific update process is described as follows:

$$\mathbf{P}_{t+1}^{f}=\frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_{i,t+1}^{f}-\overline{\mathbf{x}_{t+1}^{f}})(\mathbf{x}_{i,t+1}^{f}-\overline{\mathbf{x}_{t+1}^{f}})^{T} \tag{3}$$

$$\mathbf{K}_{t+1}=\mathbf{P}_{t+1}^{f}\mathbf{H}^{T}(\mathbf{H}\mathbf{P}_{t+1}^{f}\mathbf{H}^{T}+\mathbf{R})^{-1} \tag{4}$$

$$\mathbf{y}_{i,t+1}=\mathbf{H}(\mathbf{x}_{i,t+1}^{f}) \tag{5}$$

$$\mathbf{x}_{i,t+1}^{a}=\mathbf{x}_{i,t+1}^{f}+\mathbf{K}_{t+1}(\mathbf{Y}_{i,t+1}-\mathbf{y}_{i,t+1}) \tag{6}$$

Where $\mathbf{P}_{t+1}^{f}$ denotes the forecast error covariance matrix at time t+1. $\overline{\mathbf{x}_{t+1}^{f}}$ represents the ensemble mean of the model simulations at time t+1. $\mathbf{K}_{t+1}$ is the Kalman gain matrix at time t+1. $\mathbf{H}(\cdot)$ is the observation operator to establish a relationship between the model states and observations. $\mathbf{y}_{i,t+1}$ represents the i-th ensemble member of the simulated observations, while $\mathbf{Y}_{i,t+1}$ denotes the i-th ensemble member of perturbed observations at time t+1. Finally, $\mathbf{x}_{i,t+1}^{a}$ is the i-th ensemble member of the analysis state variables after the update.

3.2.2 Fully Connected Neural Network (FCNN)

The FCNN is a feedforward neural network designed for nonlinear function approximation. It consists of an input layer, three hidden layers, and an output layer, providing a flexible architecture for mapping high-dimensional inputs to target variables. The network takes SM forecasts from four soil layers together with surface SM observations as input variables. The output is the analyzed SM state. The Rectified Linear Unit (ReLU) activation function is adopted in the hidden layers to introduce nonlinearity and improve the stability and efficiency of model training. Given the single-point modeling setting and relatively compact input structure, the FCNN provides a simple yet effective framework for capturing nonlinear relationships between input variables and target states, while maintaining computational efficiency and ease of optimization. The detailed FCNN architecture and parameter settings are provided in Table 2.

Table 2. Neural network architecture in FCNN.

Neural network architecture	
Input size	5
Output size	4
Number of layers	3
Activation function	ReLU
Optimizer	Adam

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

3.2.3 Gated Recurrent Unit with Decay (GRU-D)

The GRU-D is a recurrent neural network variant designed to model temporal dependencies in sequential data with missing or irregular observations (Che et al., 2018). It extends the standard Gated Recurrent Unit (GRU) by explicitly incorporating missingness information to better handle incomplete time series data. Compared with traditional interpolation/imputation methods, the GRU-D model encodes missing patterns directly into trainable decay mechanisms through end-to-end joint training, simultaneously capturing variable correlations and temporal dependencies, thereby avoiding the limitations of conventional approaches that ignore missingness information and struggle with high missing rates. As illustrated in Fig. 2, the GRU-D processes sequential inputs by integrating observation values with missingness-related information, enabling time-aware state updates under irregular sampling conditions. Specifically, GRU-D introduces two auxiliary components to characterize missing data: a masking vector $m_t \in \{0, 1\}$, indicating whether each variable is observed, and a time interval $\delta_t \in \mathbb{R}$, characterizing the elapsed time from the last observation. Given a SM observation vector Y_t at time step t, these two components are defined as follows:

$$m_t = \begin{cases} 1, & if\ Y_t\ is\ observed \\ 0, & otherwise \end{cases} \tag{7}$$

$$\delta_t = \begin{cases} s_t - s_{t-1} + \delta_{t-1}, & t > 1, m_{t-1} = 0 \\ s_t - s_{t-1}, & t > 1, m_{t-1} = 1 \\ 0, & t = 1 \end{cases} \tag{8}$$

where s_t denotes the timestamp of the observation at time step t, with the initial timestamp assumed to be $s_1 = 0$.

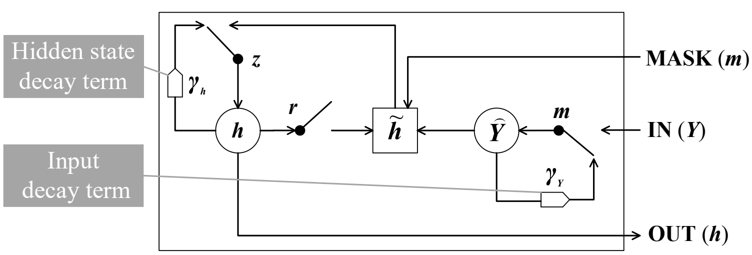


Figure 2: Architecture of the GRU-D model used for reconstructing temporally continuous observation sequences (adapted from Che et al., 2018).

By incorporating the aforementioned masking vector and time interval, the GRU-D controls the hidden state h_t at step t through a reset gate r_t and an update gate z_t, with the specific update functions defined as follows:

$$Y_t = m_t Y_t + (1 - m_t)(\gamma_{Y_t} Y_t + (1 - \gamma_{Y_t}) Y) \tag{9}$$

$$h_{t-1} = \gamma_{h_t} \odot h_{t-1} \tag{10}$$

$$r_t = \sigma(W_r Y_t + U_r h_{t-1} + V_r m_t + b_r) \tag{11}$$

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

$$z_t = \sigma(W_z Y_t + U_z h_{t-1} + V_z m_t + b_z) \tag{12}$$

$$\tilde{h}_t = \tanh(W Y_t + U(r_t \odot h_{t-1}) + V m_t + b) \tag{13}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{14}$$

where $\boldsymbol{\gamma_Y}$ and $\boldsymbol{\gamma_h}$ denote the input decay and hidden state decay, respectively, which are utilized to update the observation vector and the hidden state. $\boldsymbol{W_r}$, $\boldsymbol{W_z}$, $\boldsymbol{W}$, $\boldsymbol{U_r}$, $\boldsymbol{U_z}$, $\boldsymbol{U}$, V_r, V_z and $\boldsymbol{V}$ represent the weight matrices, while $\boldsymbol{b_r}$, $\boldsymbol{b_z}$ and $\boldsymbol{b}$ are the term. $\boldsymbol{\sigma}$ denotes the element-wise sigmoid function, and ⊙ represents the element-wise multiplication.

3.3 Soil Moisture Data Assimilation Framework

A DL-augmented SM DA framework was developed by integrating GRU-D and FCNN within an EnKF-based assimilation system coupled with Noah-MP (Fig. 3). The framework is designed to address two major challenges in conventional SM assimilation: the limited availability of temporally continuous observations and the computational burden associated with the EnKF analysis update.

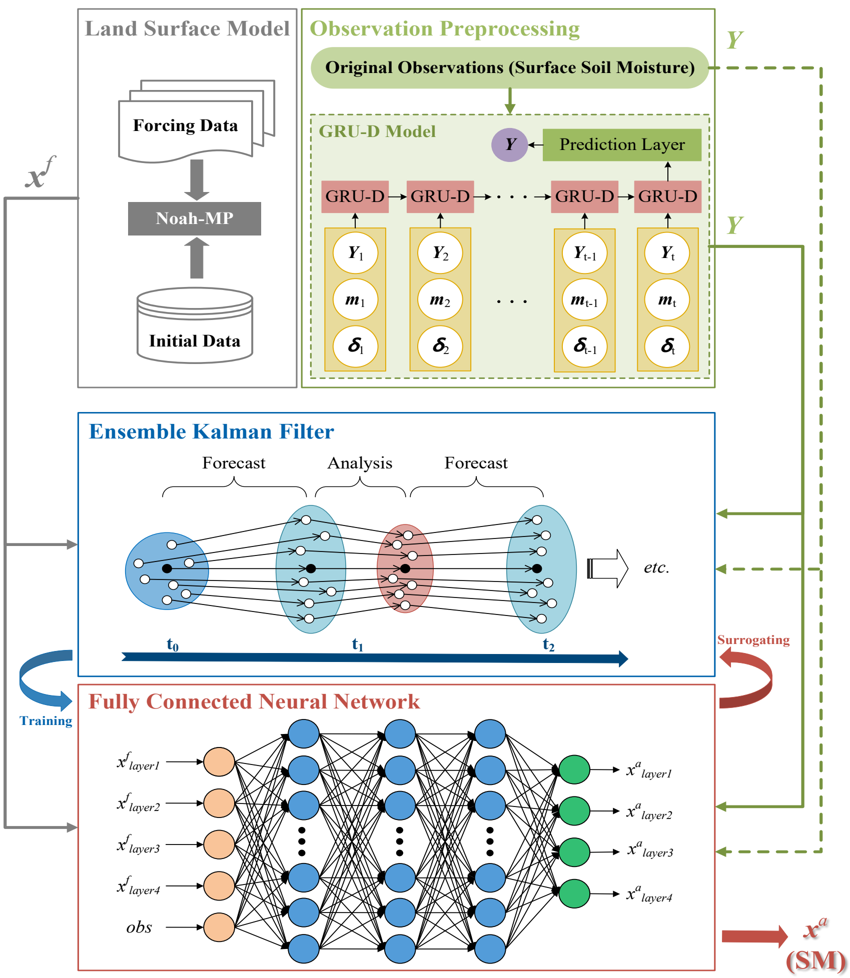
GRU-D is first used to characterize temporal dependencies in incomplete observation sequences and reconstruct missing surface SM observations. By explicitly accounting for elapsed time and missingness patterns, GRU-D generates temporally continuous observational information that can be utilized by the assimilation system. The reconstructed observations are subsequently used as observational constraints for state updating.

FCNN is employed as a surrogate representation of the EnKF analysis update. The network is trained using forecast–analysis state pairs generated from the conventional EnKF assimilation system. Specifically, the inputs consist of four-layer SM forecasts from Noah-MP together with the GRU-D reconstructed observations, while the outputs correspond to the EnKF analysis states. Through supervised learning, FCNN learns the nonlinear mapping between model forecasts, observational information, and updated states.

During assimilation, FCNN directly maps model forecasts and reconstructed observations to updated SM states without requiring explicit background-error covariance estimation, Kalman gain calculation, or matrix inversion. The resulting analysis states are then used to initialize the next model integration cycle, forming a sequential assimilation framework.

Compared with the conventional EnKF workflow, the proposed framework preserves the forecast–analysis structure of data assimilation while replacing the computationally expensive analysis update with a learned surrogate operator. Meanwhile, the incorporation of GRU-D enables more effective utilization of temporally sparse and irregular observations. Together, these two components provide a unified framework for addressing both methodological and observational limitations in SM data assimilation

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026



275 **Figure 3:** Schematic of the proposed SM DA framework based on the EnKF paradigm, with the FCNN (red box) approximating the analysis update and the GRU-D module reconstructing missing observations. The EnKF component (blue box) provides analysis states for training, while the GRU-D module (green box) ensures temporally continuous observational inputs. Solid and dashed green lines denote original and reconstructed observations, respectively.

3.4 Experimental Design

280 #### 3.4.1 Perturbation Settings and Parameterization Scheme

To verify the feasibility of the proposed SM assimilation framework, simulation and assimilation experiments were conducted from May 1 to October 31, 2020, covering the summer and wet season with a daily time step. Prior to the experiments, a one-year forcing dataset was repeatedly cycled eight times for model spin-up (Cai et al., 2014), allowing the system to reach a quasi-equilibrium state and reduce the influence of initial conditions.

285 Random perturbation fields are required for input meteorological variables to characterize the uncertainty of the forcing dataset and generate a state ensemble. In this study, the EnKF is implemented with an ensemble size of 50. Given the implicit physical

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

correlations among meteorological variables, such as the typical association between increased precipitation and decreased temperature, it is more reasonable to employ multivariate random fields synergistic perturbations (Reichle et al., 2007; Kwon et al., 2019). Following the configurations in Table 3, this study applies additive perturbations with a normal distribution to air temperature and longwave radiation, while multiplicative perturbations with a log-normal distribution are applied to precipitation and shortwave radiation (Chen et al., 2021).

Table 3. Perturbation settings for forcing data and state variables.

Description	Perturbation Type	Standard Deviation	Cross-Correlations			
			P	SW	LW	T
Precipitation (P)	M	0.5	/	-0.8	0.5	-0.1
Shortwave radiation (SW)	M	0.3	-0.8	/	-0.5	0.3
Longwave radiation (LW)	A	50 W m^{-2}	0.5	-0.5	/	0.6
Air temperature (T)	A	2 K	-0.1	0.3	0.6	/
Soil moisture	M	0.1				

Note: A indicates normally distributed additive perturbations; M indicates lognormally distributed multiplicative perturbations.

The parameterization schemes of the Noah-MP 5.0 model encompass core modules such as energy balance, hydrological cycle, and biogeochemistry, providing multiple options for several key physical processes (He et al., 2023). In this study, the combination of parameterization schemes (Table S1 in the Supplement) recommended by Chen et al. (2021) is adopted. Based on this configuration, the ensemble of model simulations generated by driving the model with perturbed meteorological forcing is defined as the open-loop (OL) run.

3.4.2 Observation Interval Settings

Considering the limited availability of continuous in situ SM observations, five assimilation scenarios with different observation intervals are designed to examine the sensitivity of the assimilation performance to observation frequency. The scenarios include daily, 3-day, 5-day, 7-day, and randomly sampled observation intervals. These configurations are used to represent varying degrees of temporal sparsity in observational constraints, enabling a systematic evaluation of assimilation behavior under different data availability conditions. In addition, the computational efficiency of all assimilation methods is assessed by recording the total runtime in a CPU-based environment.

3.4.3 Data Assimilation Scheme

To evaluate the respective and combined contributions of the FCNN-based assimilation operator and the GRU-D-based observation reconstruction module, three assimilation schemes are designed in this study, using the standard EnKF-based assimilation system as a benchmark. The first scheme is the baseline EnKF assimilation, in which raw SM observations are directly assimilated whenever available. In the DL-based assimilation framework, two comparative cases are defined: P1, in

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

which the FCNN-based update operator is used to replace the EnKF analysis step, while still assimilating raw SM observations with missing values; and P2, in which the full proposed framework is employed, where missing observations are first reconstructed using the GRU-D model before being input into the FCNN-based assimilation system. This experimental design enables a systematic evaluation of (i) the performance gain from replacing the EnKF update step with FCNN, and (ii) the additional benefit of incorporating GRU-D-based observation reconstruction. All assimilation experiments are conducted under identical model configurations, forcing data, and computational environment to ensure fair comparison of both accuracy and computational efficiency.

3.5 Evaluation Indicators

To more clearly demonstrate the predictive performance of the proposed framework, the mean bias error (MBE), the root-mean-square error (RMSE), and the correlation coefficient (R) are used as evaluation measures in this study.

$$MBE = \frac{1}{T}\sum_{t=1}^{T}(x_t - x_{obs,t}) \tag{15}$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_t - x_{obs,t})^2} \tag{16}$$

$$R = \frac{\sum_{t=1}^{T}(x_t - \overline{x})(x_{obs,t} - \overline{x_{obs}})}{\sqrt{\sum_{t=1}^{T}(x_t - \overline{x})^2}\sqrt{\sum_{t=1}^{T}(x_{obs,t} - \overline{x_{obs}})^2}} \tag{17}$$

where T is the total number of time steps, x_t denotes the SM values estimated from model simulations or assimilation experiments, and $x_{obs,t}$ represents the in situ SM observations at time t.

4 Results

The performance of three assimilation strategies (EnKF, P1, and P2) is comprehensively evaluated through analyses of time series behavior, quantitative metrics, and computational runtime. The main text presents result for three representative scenarios: continuous observations, observations available every three days, and randomly distributed observation intervals. Evaluation metrics calculated over the entire study period for these scenarios are presented in Figs. S1, S2, and S13 in the Supplement. Additional results for scenarios with observation intervals of five and seven days, together with their corresponding evaluation metrics, are presented in Figs. S3-S12 in the Supplement.

4.1 Continuous observations

Fig. 4 presents the averaged SM time series and corresponding scatter plots for the OL simulation and three assimilation experiments (EnKF, P1, and P2) at depths of 10, 30, 60, and 100 cm, based on observations from the eight sites. Overall, the OL results show obvious underestimation at all four soil layers. In contrast, the assimilation results significantly improve the

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

SM estimates and bring them closer to the observations, indicating that DA effectively corrects the systematic bias in the model background field. In this scenario, the observation data are continuously available at a daily time step, resulting in completely consistent outputs of P1 and P2. During the training period, the FCNN assimilation results are highly consistent with the EnKF analysis results, indicating that the network can effectively capture the nonlinear mapping from the model forecast and observation inputs to the corresponding analysis states. This consistency is also maintained in the testing phase, and the FCNN outputs are still highly consistent with the EnKF results, indicating that the learned model has good generalization ability for unseen data and successfully approximates the EnKF update operator, thus realizing efficient assimilation without explicit covariance calculation. The scatter distributions in the right column of Fig. 4 further support the above findings. Compared with OL, all assimilation experiments exhibit distributions that are more closely aligned with the 1:1 reference line, indicating improved agreement with the observations and reduced systematic bias. Among the three assimilation schemes, P1 and P2 show scatter distributions that are highly comparable to those of EnKF, suggesting that FCNN successfully reproduces the statistical characteristics of the EnKF analysis states. These results indicate that the learned surrogate update operator can effectively capture the relationship between model forecasts, observations, and analysis states while maintaining assimilation performance comparable to that of the conventional EnKF. Comparisons across soil depths show that the magnitude of assimilation improvement gradually decreases with increasing depth. The largest improvement is observed at 10 cm, whereas the improvements at 60-100 cm are considerably smaller. This depth-dependent behavior is primarily attributable to the limited capability of surface SM observations to constrain deeper soil layers. Because the observations directly inform near-surface soil moisture states, their influence on deeper layers relies mainly on vertical soil water transport and model dynamics. As a result, observational information gradually attenuates as it propagates downward through the soil profile.

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

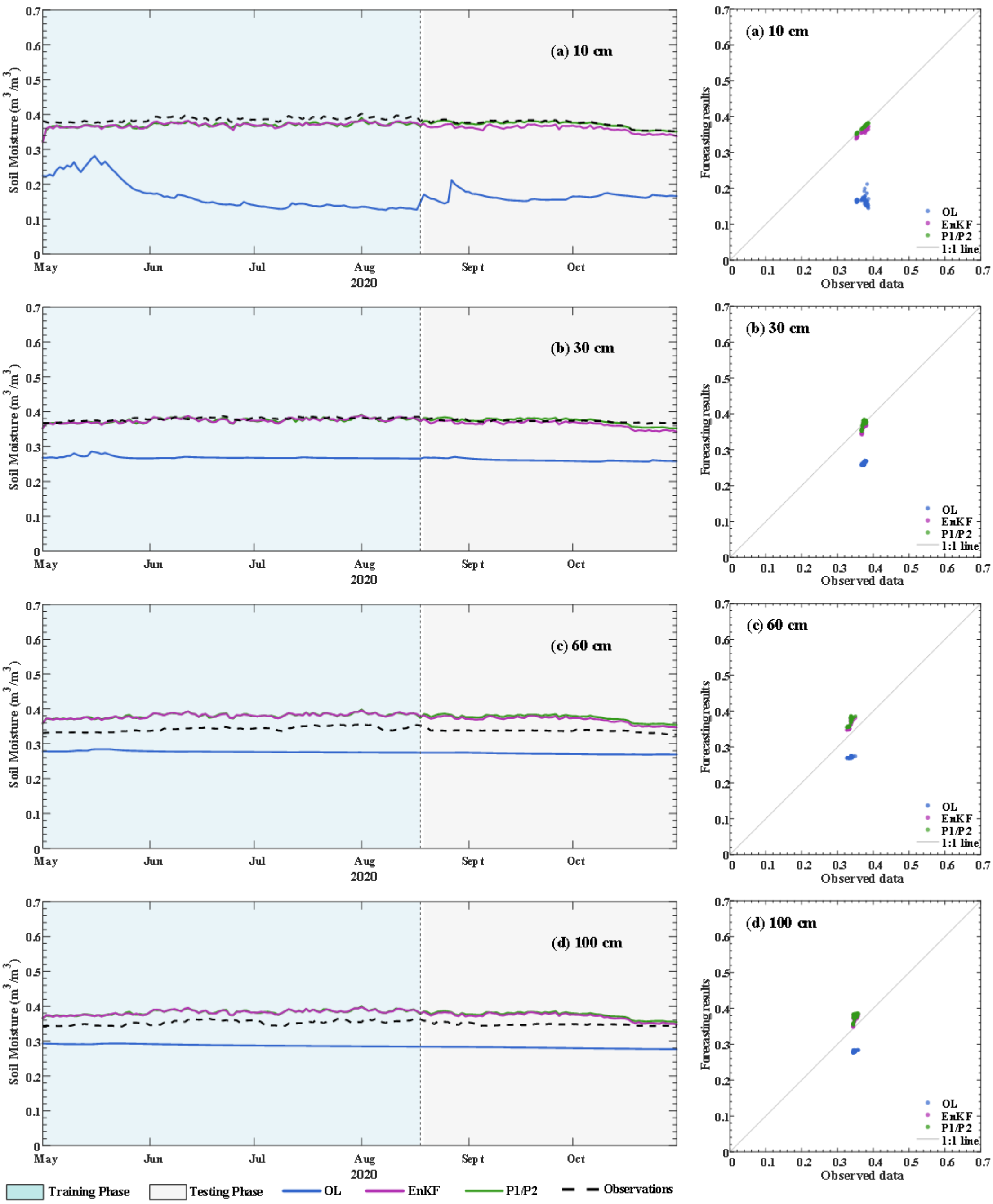


Figure 4: Averaged daily SM time series (left panels) and scatter plots (right panels) across the eight sites for OL, EnKF, P1, P2, and observations at depths of 10, 30, 60, and 100 cm ((a)-(d)). The gray solid line denotes the 1:1 reference line.

Fig. 5 and Fig. 6 present the RMSE, MBE and R at the eight observation sites during the training and testing periods, respectively. Overall, EnKF, P1, and P2 exhibit similar performance across all sites, whereas OL generally underestimates SM and shows larger errors and weaker correlations with observations. After assimilation with FCNN or EnKF, the simulation

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

accuracy of the surface soil layer (0-10 cm) improves markedly, with RMSE value decreasing substantially from 0.238 to 0.016/0.018 for the training/testing periods. Estimates of deep SM also improved, with RMSE values in the 10-30 cm, 30-60 cm and 60-100 cm layers decreasing from 0.152, 0.108 and 0.078 to 0.029/0.029, 0.057/0.059 and 0.060/0.059, respectively. In addition, MBE values from both assimilation methods approach zero, while the R increases significantly, indicating substantially improved agreement between assimilated simulations and observations. Overall, both the FCNN-based and EnKF assimilation schemes effectively enhance SM estimation accuracy, further supporting the reliability of the proposed FCNN-based assimilation framework as a surrogate representation of the EnKF update process.

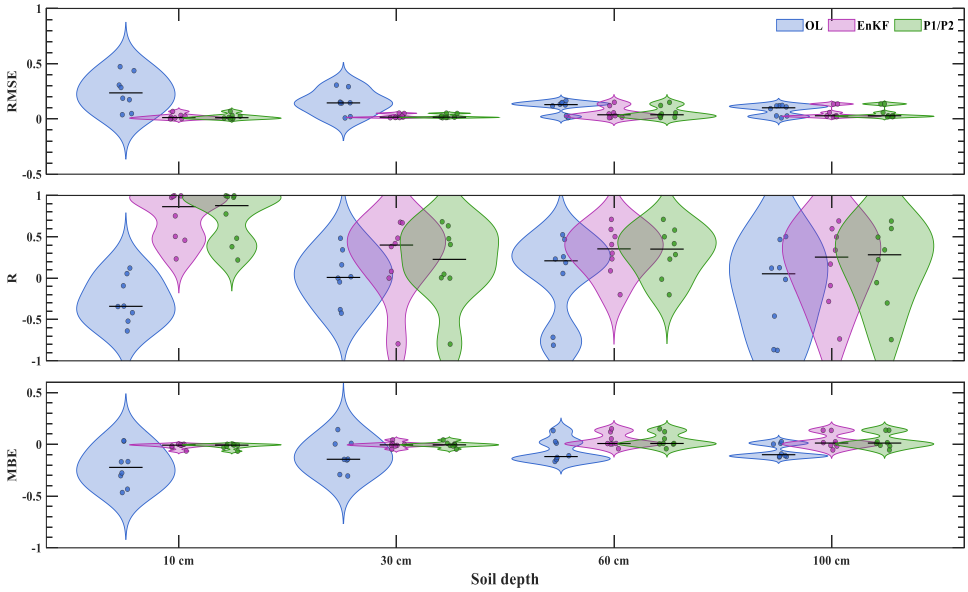


Figure 5: Comparison of RMSE, MBE, and R for OL simulations and different assimilation strategies across four soil layers at eight sites during the training period under continuous observations.

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

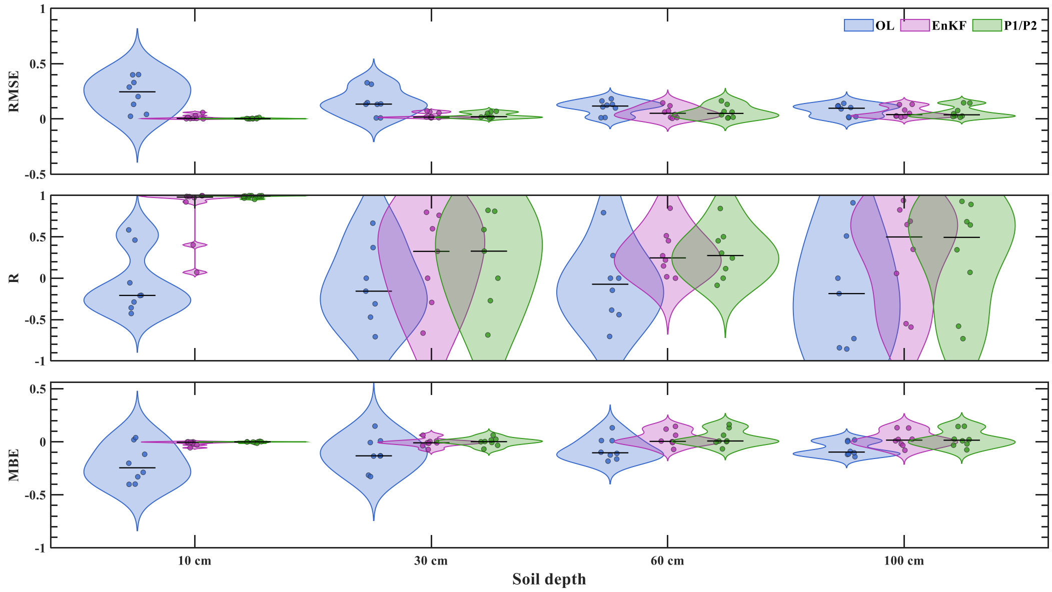


Figure 6: Same evaluation as Fig. 5, but for the testing period under continuous observations.

According to the runtime statistics, the average computation time over the entire simulation period is 293.70 s for EnKF and 186.73 s for FCNN, corresponding to a runtime reduction of approximately 36.4%. The higher computational cost of EnKF mainly arises from the need to estimate background error covariance and perform matrix-based analysis updates at each assimilation cycle. In contrast, FCNN learns the EnKF update relationship during the offline training stage, and the analysis states can subsequently be obtained through efficient forward propagation during assimilation. Consequently, FCNN significantly reduces computational cost while achieving assimilation performance comparable to that of EnKF.

4.2 3-day observation interval

Fig. 7 presents the original SM observations and the reconstructed SM time series generated by the GRU-D model. The GRU-D reconstruction effectively fills in missing values, producing a temporally continuous observation record that provides more complete inputs for the subsequent DA process. Fig. 8 presents the averaged SM time series and scatter plots across the eight study sites under the 3-day observation interval scenario. Compared with the continuous-observation experiment, clearer differences emerge between P1 and P2, particularly in the surface soil layer. Although all assimilation schemes improve the OL simulation by incorporating observational information, the advantage of P2 becomes more evident when observations are temporally sparse. Under this scenario, SM estimation relies more heavily on model forecasts during periods without observations, allowing model biases to gradually accumulate between successive assimilation updates. By reconstructing missing observations, GRU-D provides temporally continuous observational constraints that can be incorporated into the assimilation process. As a result, P2 produces smoother and more stable SM estimates than P1 and exhibits closer agreement with the observations, particularly during periods with prolonged observational gaps. These results demonstrate the importance

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

of explicitly accounting for temporal sparsity in SM data assimilation and highlight the benefit of incorporating reconstructed observations under limited observation conditions.

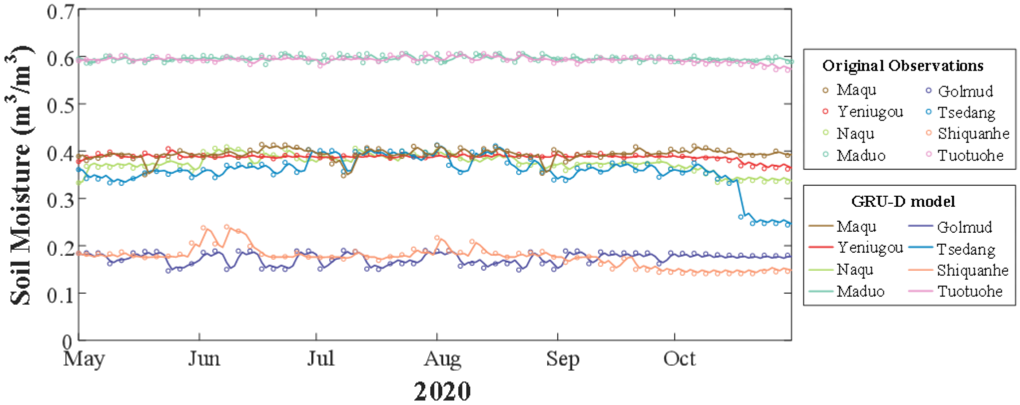


Figure 7: Comparison between original SM observations and the GRU-D reconstructed series under a 3-day observation interval at eight sites. Dots represent the original observations, while the solid line represents the time series reconstructed by the GRU-D model.

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

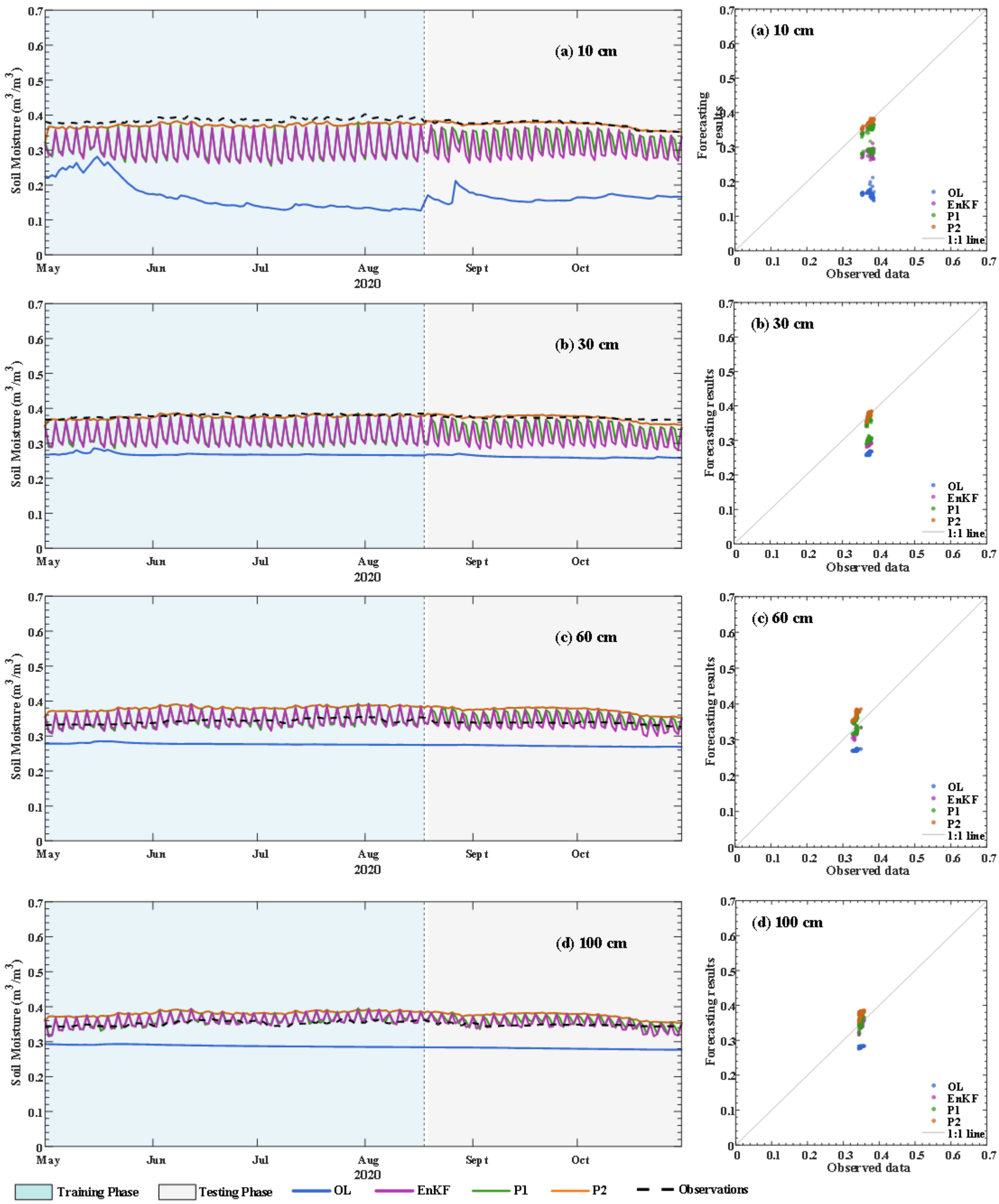


Figure 8: Averaged daily SM time series (left panels) and scatter plots (right panels) across the eight study sites for OL, EnKF, P1, P2, and observations under the 3-day observation interval scenario. Panels (a)-(d) correspond to soil depths of 10, 30, 60, and 100 cm, respectively. The gray solid line denotes the 1:1 reference line.

Fig. 9 and Fig. 10 summarize the RMSE, MBE, and R values at the eight observation sites during the training and testing periods under the 3-day observation interval scenario. Overall, all assimilation schemes outperform OL, demonstrating the

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

benefit of incorporating observational information into the SM estimation process. Compared with EnKF and P1, P2 generally achieves lower RMSE and MBE values and higher R values at most sites, indicating that the incorporation of reconstructed observations provides additional skill under temporally sparse observation conditions. The assimilation improvements are most pronounced in the upper two soil layers, where EnKF, P1, and particularly P2 substantially improve SM estimation. Specifically, the mean R of the first layer increases from -0.067 to approximately 0.364/0.311, while in the second layer it increases from -0.076 to about 0.156/0.131. Meanwhile, both the mean MBE and RMSE decrease markedly, with the average RMSE reductions exceeding 60% in the upper layers. With the P2 scheme, further improvements are achieved, with the mean R value in the first layer reaching 0.757 and that in the second layer increasing to 0.275, while the corresponding RMSE reductions exceed 80%. In contrast, improvements in the deeper layers are relatively limited, and the differences among the assimilation schemes become less evident. This depth-dependent behavior reflects the limited ability of surface observations to constrain subsurface states. Because observational information is introduced at the surface and propagated downward primarily through vertical soil water transport processes, its influence gradually weakens with increasing depth, resulting in reduced assimilation impacts in deeper soil layers.

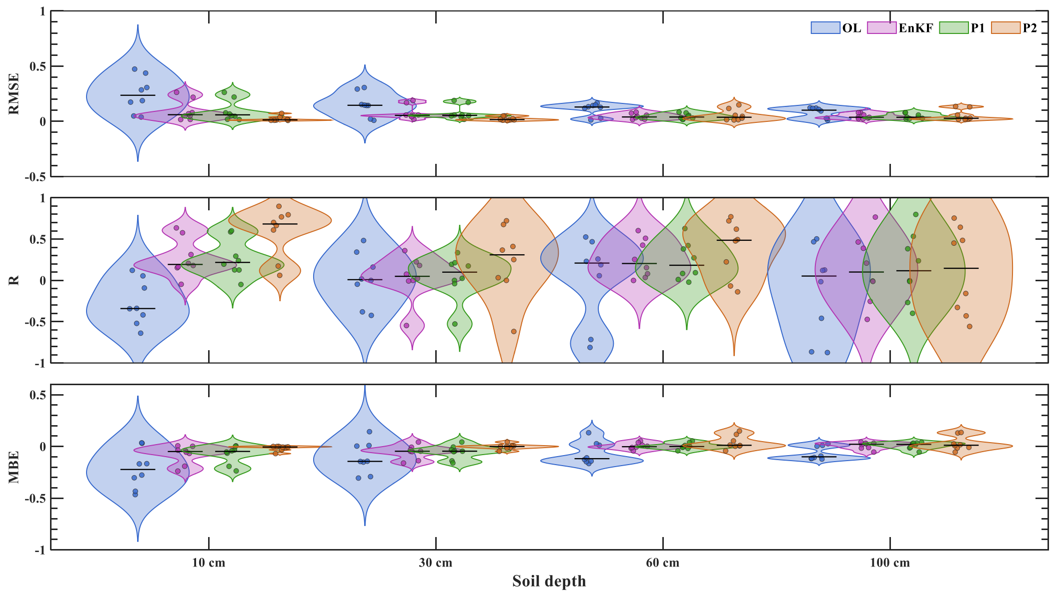


Figure 9: Same evaluation as Fig. 5, but for the training period under a 3-day observation interval.

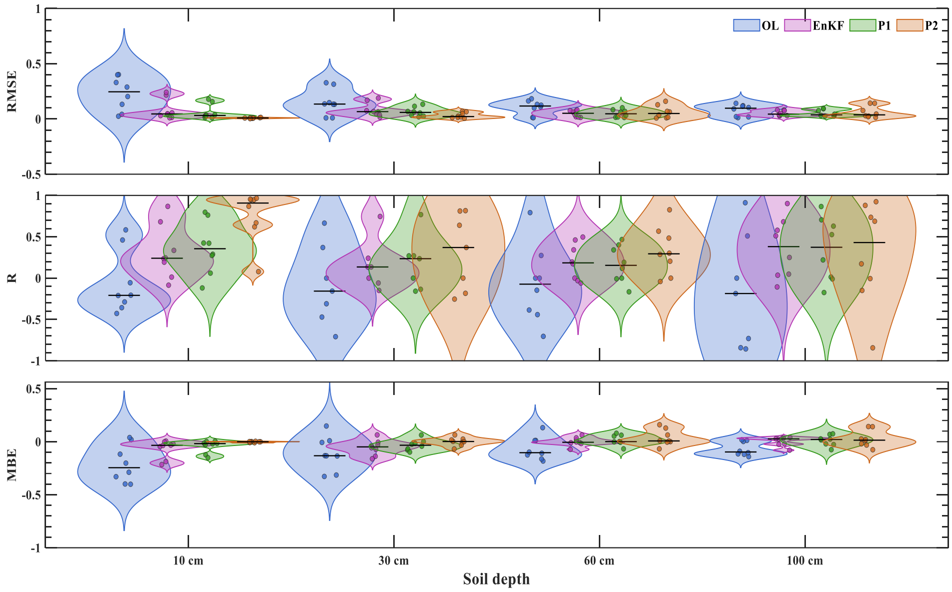


Figure 10: Same evaluation as Fig. 5, but for the testing period under a 3-day observation interval.

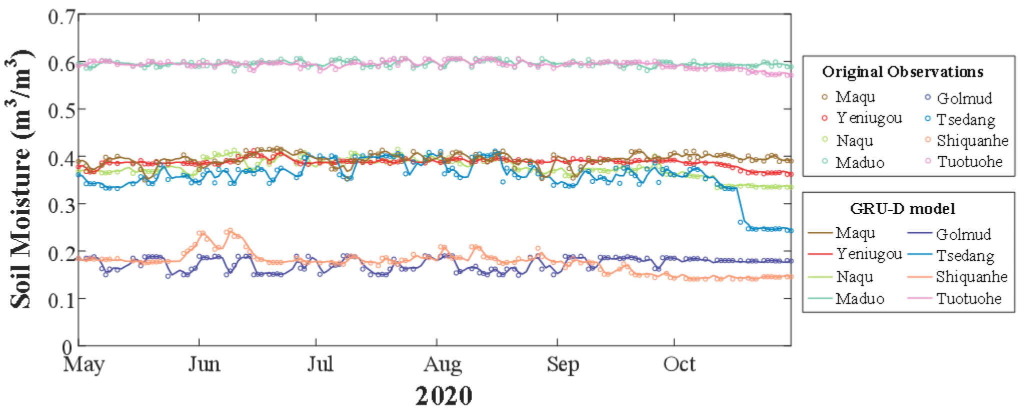
According to the runtime statistics, the average computational time of EnKF over the entire simulation period is 284.76 s, while P1 and P2 require 189.63 s and 186.97 s, respectively. This corresponds to a reduction in computational cost of approximately 33.4% for P1 and 34.4% for P2 relative to EnKF. Both P1 and P2 substantially reduce computational cost compared with EnKF. In particular, P1 achieves a significant reduction while maintaining assimilation accuracy comparable to EnKF, whereas P2 exhibits a similar computational cost to P1 while providing improved consistency with observations, as demonstrated in the statistical evaluation results.

4.3 Random observation intervals

Under fixed observation intervals, the P2 has been shown to effectively reconstruct missing observations and improve the SM estimates. This section further evaluates the robustness of the proposed framework under irregular observation timing conditions.

Fig. 11 presents the GRU-D reconstructed observation series under random observation intervals. Fig. 12 compares the OL simulations with results of the three assimilation schemes. Both the EnKF and P1 schemes effectively correct model biases once observational information becomes available. However, as the observation interval increases, short-term sharp deviations are still observed, indicating that model forecast errors tend to accumulate during periods without observational constraints, leading to reduced temporal stability in the simulated SM series. In contrast, the P2 scheme, which incorporates GRU-D-based observation reconstruction, provides continuous observational constraints and maintains more stable assimilation updates under irregular observation conditions, resulting in improved simulation consistency.

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026



435

Figure 11: Same as Fig. 7, but for irregular observation intervals.

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

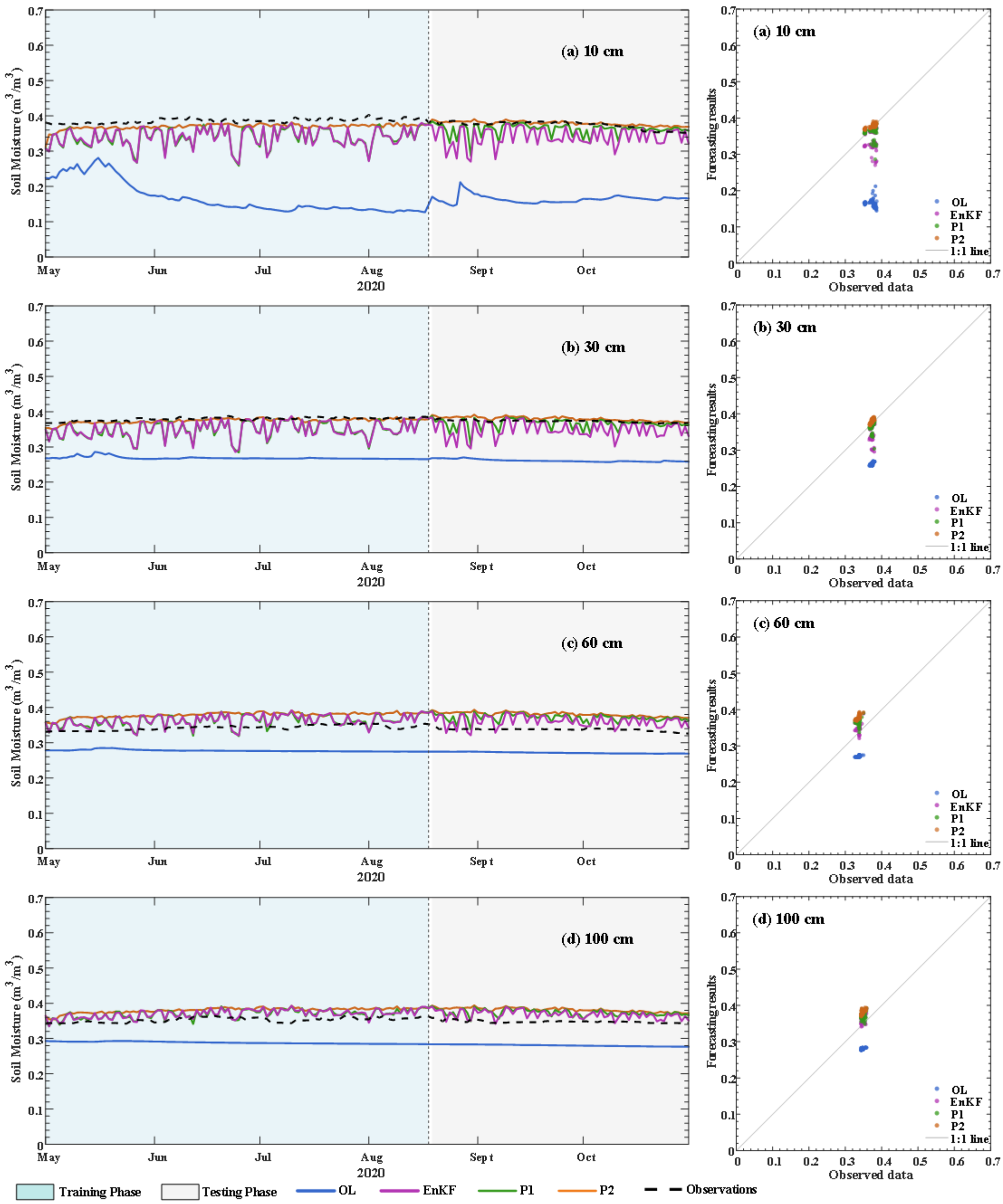


Figure 12: Same as Fig. 8, but for irregular observation intervals.

Fig. 13 shows the RMSE, MBE, and R during the training period, indicating that the FCNN can still effectively approximate the EnKF analysis outputs under irregular observation scenarios. Consistent with the testing results on Fig. 14, the P2 scheme generally outperforms the other methods in surface SM simulation across all sites. For deeper soil layers, P2 still performs

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

better in most cases. However, at several sites (e.g., Maduo, Golmud, and Shiquanhe), the EnKF, P1, or OL simulations show slightly better performance for certain metrics. This suggests that when model structural uncertainties or soil parameter errors are relatively large, the benefits of DA for deep SM remain limited.

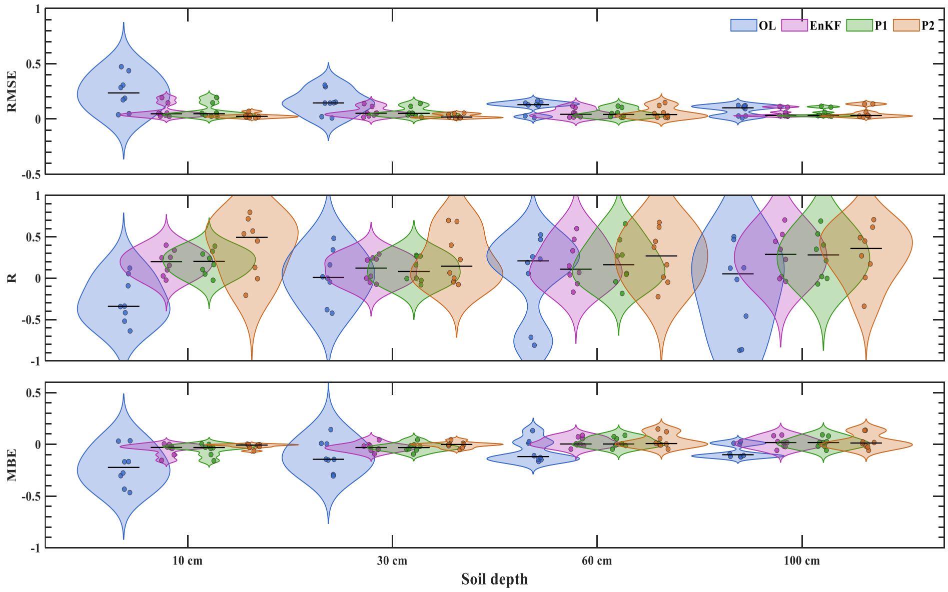


Figure 13: Same evaluation as Fig. 5, but for the training period under randomly distributed observation intervals.

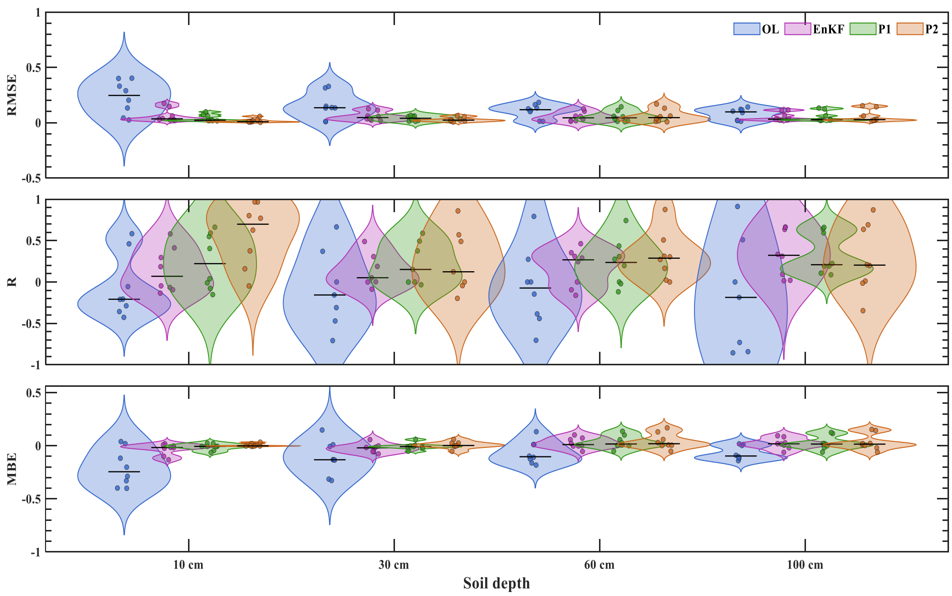


Figure 14: Same evaluation as Fig. 5, but for the testing period under randomly distributed observation intervals.

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

Over the entire simulation period, the conventional EnKF-based assimilation strategy exhibits the highest computational cost, with an average runtime of 282.69 s. In comparison, the FCNN-based assimilation requires 187.59 s, corresponding to a runtime reduction of approximately 33.6%. The P2 scheme shows a similar computational cost of 193.33 s, representing a reduction of approximately 31.6% relative to EnKF. These results indicate that the proposed DL-augmented assimilation frameworks substantially reduce computational cost while maintaining assimilation performance comparable to that of the conventional EnKF approach.

5 Discussion

5.1 Rationale for FCNN and GRU-D Integration

This study proposes an SM DA framework that integrates FCNN and GRU-D components, in which SM states are estimated by combining Noah-MP model forecasts with surface SM observations. Previous studies have demonstrated that neural network-based approaches can effectively emulate the EnKF update step, particularly in handling nonlinear and non-Gaussian characteristics (Boucher et al., 2020; Cai et al., 2024; Zhang et al., 2025). By learning nonlinear mappings between model forecasts, observations, and analysis states, these methods relax the restrictive assumptions of linearity and Gaussian error statistics inherent in traditional DA formulations. However, existing approaches are often based on relatively deep and complex network architectures, which, while expressive, may introduce challenges in computational efficiency and generalization stability, particularly in large-scale or frequently updated land surface applications. To address this issue, a lightweight FCNN architecture is adopted in this study to balance nonlinear representation capability and computational efficiency, making it more suitable for operational land surface assimilation. In addition, most conventional studies treat observational preprocessing in a simplified manner, typically relying on interpolation or smoothing techniques that fail to adequately represent temporal discontinuities and missing-data patterns in observational records. To overcome this limitation, a GRU-D-based reconstruction module is introduced to explicitly model missingness patterns and temporal intervals, thereby enabling more physically consistent and temporally continuous observational constraints within the assimilation system. Overall, the proposed framework maintains the structural logic of the EnKF-based assimilation paradigm while extending it into a hybrid data-driven formulation. This is achieved by replacing the explicit Kalman update with a learned nonlinear operator and embedding a missing-data-aware observational reconstruction mechanism, thereby improving both robustness under incomplete observations and computational efficiency in soil moisture estimation.

5.2 Advantages of the Proposed DA Framework

The proposed framework improves both computational efficiency and assimilation robustness by integrating neural operator learning with missing-data-aware observation reconstruction. By replacing the explicit Kalman filter update with a FCNN-based nonlinear mapping, the framework provides a data-driven approximation of the EnKF analysis operator, substantially reducing the computational cost associated with covariance propagation and gain estimation in conventional assimilation

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

schemes. This reformulation enables efficient application to high-resolution LSMs and large-scale or long-term simulations under limited computational resources. Meanwhile, from a dynamical systems perspective, DA relies on continuous observational constraints to regulate model state evolution; however, intermittent or missing observations weaken this constraint, allowing model errors to accumulate during temporal integration and degrading system stability. To address this issue, the GRU-D module reconstructs temporally continuous observational sequences by explicitly modeling missingness patterns and temporal dependencies, thereby restoring observation-driven constraints throughout the assimilation process. Overall, the framework establishes a unified hybrid paradigm that couples neural approximation of the assimilation operator with temporally consistent observational reconstruction, enhancing robustness under irregular observation conditions while maintaining computational efficiency.

5.3 Limitations and Prospects

Although the proposed framework demonstrates notable advantages in handling observation sequences with missing values and improving computational efficiency, several limitations remain. The current evaluation is conducted primarily at the site scale, which does not fully represent the spatial heterogeneity and process complexity of large-scale land surface systems. Future work should extend the framework to regional and continental scales to further assess its robustness and generalization capability under diverse hydro-meteorological conditions. It is important to clarify the methodological positioning of the proposed approach. The framework is developed within the EnKF-based DA paradigm, where the FCNN is designed to emulate the analysis update step rather than to replace the overall assimilation theory. In this context, the GRU-D module acts as an auxiliary component that improves the continuity of observational constraints by reconstructing missing observations, but the overall framework remains grounded in the EnKF formulation. Consequently, while the proposed method improves computational efficiency and robustness under missing observations, it does not constitute a fully independent alternative to traditional DA theory, and future research should explore more general assimilation paradigms beyond the EnKF-based framework. In addition, limitations in the physical representation of the underlying Noah-MP model remain an important source of uncertainty, particularly in the simulation of deep SM, where structural and parameterization biases are evident. Addressing these limitations will require targeted calibration and process-level refinement of LSMs (Raoult et al., 2025). Finally, the current observation reconstruction strategy does not explicitly incorporate additional land surface energy variables such as shortwave radiation, longwave radiation, and latent heat flux, which may provide complementary physical constraints for SM estimation under sparse observation conditions. Future work should consider integrating multi-source observational information to enhance the physical consistency and stability of reconstructed SM observation sequences (Wang et al., 2025).

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

6 Conclusion

This study developed a DL-augmented SM DA framework by integrating FCNN and GRU-D with the Noah-MP land surface model, enabling efficient SM estimation under temporally incomplete observational conditions. The framework was evaluated across five experimental scenarios at eight sites representing different land surface conditions over the Tibetan Plateau. Results show that the FCNN-based assimilation reproduces EnKF-based state updates with comparable temporal dynamics and statistical performance, indicating that the learned surrogate operator can effectively capture the nonlinear relationship between model forecasts, observations, and analysis states. This suggests the potential of replacing computationally expensive ensemble-based update procedures with data-driven nonlinear mappings. The inclusion of GRU-D improves the utilization of temporally incomplete observations by generating continuous observational inputs under missing-data conditions, thereby enhancing the stability of the assimilation process under sparse observation scenarios. From a computational perspective, the proposed framework reduces the average runtime by approximately 33.6% compared with the conventional EnKF method, demonstrating the efficiency of neural surrogate-based assimilation.

Overall, the proposed framework provides a hybrid DA paradigm that combines physical land surface modeling with data-driven operator learning and missing-data-aware observation processing. It improves computational efficiency and maintains reliable assimilation performance under limited observational availability, offering a scalable pathway for land surface data assimilation in data-sparse environments. Future work will focus on extending the framework to larger spatial domains and further investigating parameters and structural uncertainties within the assimilation system.

Code and data availability

All datasets used in this study are publicly accessible via the corresponding links provided in the manuscript. The source code of the Noah-MP model is available at https://github.com/NCAR/noahmp . Alternatively, the Noah-MP code can also be obtained as a submodule within the HRLDAS framework at https://github.com/NCAR/hrldas . All scripts developed to reproduce the experiments presented in this paper are available from the corresponding author upon reasonable request.

Author contributions

YZ: Conceptualization, Methodology, Software, validation, formal analysis, Writing-original draft. JL: Conceptualization, Methodology, Writing-review and editing, Supervision. CL: Writing-review and editing, Supervision. WZ: Writing-review and editing. YZ: Writing-review and editing, Supervision. YH: Writing-review and editing.

Competing interests

The authors declare no conflict of interest.

Acknowledgements

The authors would like to thank the project team for many inspiring discussions. A great thank goes also to the stakeholders involved in project activities, who provided their knowledge and expertise on the area.

Financial support

This work was supported by National Natural Science Foundation of China (Grant No. 42130113, 42371398, 42471434).

References

Abowarda, A. S., Bai, L., Zhang, C., Long, D., Li, X., Huang, Q., and Sun, Z.: Generating surface soil moisture at 30 m spatial resolution using both data fusion and machine learning toward better water resources management at the field scale, Remote Sens. Environ., 255, 112301, https://doi.org/10.1016/j.rse.2021.112301, 2021.

Boucher, M.-A., Quilty, J., and Adamowski, J.: Data assimilation for streamflow forecasting using extreme learning machines and multilayer perceptrons, Water Resour. Res., 56, e2019WR026226, https://doi.org/10.1029/2019WR026226, 2020.

Bradley, B. A., Jacob, R. W., Hermance, J. F., and Mustard, J. F.: A curve fitting procedure to derive inter-annual phenologies from time series of noisy satellite NDVI data, Remote Sens. Environ., 106, 137–145, https://doi.org/10.1016/j.rse.2006.08.002, 2007.

Burgers, G., van Leeuwen, P. J., and Evensen, G.: Analysis scheme in the ensemble Kalman filter, Mon. Wea. Rev., 126, 1719–1724, https://doi.org/10.1175/1520-0493(1998)126<<1719:ASITEK>2.0.CO;2, 1998.

Cai, S., Fang, F., and Wang, Y.: Advancing neural network-based data assimilation for large-scale spatiotemporal systems with sparse observations, Phys. Fluids, 36, 096616, https://doi.org/10.1063/5.0228384, 2024.

Cai, X., Yang, Z.-L., David, C. H., Niu, G.-Y., and Rodell, M.: Hydrological evaluation of the Noah-MP land surface model for the Mississippi River Basin, J. Geophys. Res.-Atmos., 119, 23–38, https://doi.org/10.1002/2013JD020792, 2014.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G.: Data assimilation in the geosciences: An overview of methods, issues, and perspectives, WIREs Clim. Change, 9, e535, https://doi.org/10.1002/wcc.535, 2018.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y.: Recurrent neural networks for multivariate time series with missing values, Sci. Rep., 8, 6085, https://doi.org/10.1038/s41598-018-24271-9, 2018.

Chen, H., Nan, Z., Zhao, L., Ding, Y., Chen, J., and Pang, Q.: Noah modelling of the permafrost distribution and characteristics in the West Kunlun area, Qinghai–Tibet Plateau, China, Permafrost Periglac. Process., 26, 160–174, https://doi.org/10.1002/ppp.1841, 2015.

Chen, W., Huang, C., and Yang, Z.-L.: More severe drought detected by the assimilation of brightness temperature and terrestrial water storage anomalies in Texas during 2010–2013, J. Hydrol., 603, 126802, https://doi.org/10.1016/j.jhydrol.2021.126802, 2021.

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

Chen, W., Shen, H., Huang, C., and Li, X.: Improving soil moisture estimation with a dual ensemble Kalman smoother by jointly assimilating AMSR-E brightness temperature and MODIS LST, Remote Sens., 9, 273, https://doi.org/10.3390/rs9030273, 2017.

Chattopadhyay, A., Nabizadeh, E., Bach, E., and Hassanzadeh, P.: Deep learning-enhanced ensemble-based data assimilation for high-dimensional nonlinear dynamical systems, J. Comput. Phys., 477, 111918, https://doi.org/10.1016/j.jcp.2023.111918, 2023.

Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P., Slater, A. G., Schmidt, J., and Uddstrom, M. J.: Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model, Adv. Water Resour., 31, 1309–1324, https://doi.org/10.1016/j.advwatres.2008.06.005, 2008.

Dai, Y., Zeng, X., Dickinson, R. E., Baker, I., Bonan, G. B., Bosilovich, M. G., Denning, A. S., Dirmeyer, P. A., Houser, P. R., Niu, G., Oleson, K. W., Schlosser, C. A., and Yang, Z.-L.: The Common Land Model, B. Am. Meteorol. Soc., 84, 1013–1024, https://doi.org/10.1175/BAMS-84-8-1013, 2003.

Ding, L., Bai, Y.-L., Fan, M.-H., Song, W., and Ren, H.-H.: Using a snow ablation optimizer in an autonomous echo state network for the model-free prediction of chaotic systems, Nonlinear Dynam., 112, 11483–11500, https://doi.org/10.1007/s11071-024-09656-y, 2024.

Ding, L., Bai, Y., Zheng, D., Pan, X., Fan, M., and Li, X.: Chaotic climate system forecasting using an improved echo state network with sparse observations, Sci. China Earth Sci., 68, 2346–2360, https://doi.org/10.1007/s11430-024-1593-9, 2025.

Dong, J., Lei, F., and Crow, W. T.: Land transpiration-evaporation partitioning errors responsible for modeled summertime warm bias in the central United States, Nat. Commun., 13, 336, https://doi.org/10.1038/s41467-021-27938-6, 2022.

Evensen, G.: The ensemble Kalman filter: Theoretical formulation and practical implementation, Ocean Dynam., 53, 343–367, https://doi.org/10.1007/s10236-003-0036-9, 2003.

Fan, Y., Wang, X., Funk, T., Rashid, I., Herman, B., Bompoti, N., Mahmud, M. S., Chrysochoou, M., Yang, M., Vadas, T. M., Lei, Y., and Li, B.: A critical review for real-time continuous soil monitoring: Advantages, challenges, and perspectives, Environ. Sci. Technol., 56, 13546–13564, https://doi.org/10.1021/acs.est.2c03562, 2022.

Fisher, R. A., and Koven, C. D.: Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems, J. Adv. Model. Earth Syst., 12, e2018MS001453, https://doi.org/10.1029/2018MS001453, 2020.

He, C., Valayamkunnath, P., Barlage, M., Chen, F., Gochis, D., Cabell, R., Schneider, T., Rasmussen, R., Niu, G.-Y., Yang, Z.-L., Niyogi, D., and Ek, M.: Modernizing the open-source community Noah with multi-parameterization options (Noah-MP) land surface model (version 5.0) with enhanced modularity, interoperability, and applicability, Geosci. Model Dev., 16, 5131–5151, https://doi.org/10.5194/gmd-16-5131-2023, 2023.

Houtekamer, P. L., and Mitchell, H. L.: Ensemble Kalman filtering, Q. J. Roy. Meteorol. Soc., 131, 3269–3289, https://doi.org/10.1256/qj.05.135, 2005.

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

Huang, C., Chen, W., Li, Y., Shen, H., and Li, X.: Assimilating multi-source data into land surface model to simultaneously improve estimations of soil moisture, soil temperature, and surface turbulent fluxes in irrigated fields, Agr. Forest Meteorol., 230–231, 142–156, https://doi.org/10.1016/j.agrformet.2016.03.013, 2016.

Huang, C., Hou, J., Li, W., Gu, J., Zhang, Y., Han, W., Wang, W., Wen, X., and Zhu, G.: Data assimilation in terrestrial hydrology based on deep learning fusing remote sensing big data: Research advances and key scientific issues, Adv. Earth Sci., 38, 441-452, https://doi.org/10.11867/j.issn.1001-8166.2023.022, 2023.

Jia, K., Yang, L., Liang, S., Xiao, Z., Zhao, X., Yao, Y., Zhang, X., Jiang, B., and Liu, D.: Long-term Global Land Surface Satellite (GLASS) fractional vegetation cover product derived from MODIS and AVHRR data, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 12, 508–518, https://doi.org/10.1109/JSTARS.2018.2854293, 2019.

Jiang, Y., Yang, K., Qi, Y., Zhou, X., He, J., Lu, H., Li, X., Chen, Y., Li, X., Zhou, B., Mamtimin, A., Shao, C., Ma, X., Tian, J., and Zhou, J.: TPHiPr: A long-term (1979–2020) high-accuracy precipitation dataset (1/30°, daily) for the Third Pole region based on high-resolution atmospheric modeling and dense observations, Earth Syst. Sci. Data, 15, 621–638, https://doi.org/10.5194/essd-15-621-2023, 2023.

Jiang, Y., Tang, W., Yang, K., He, J., Shao, C., Zhou, X., Lu, H., Chen, Y., Li, X., and Shi, J.: Development of a high-resolution near-surface meteorological forcing dataset for the Third Pole region, Sci. China Earth Sci., 68, 1274–1290, https://doi.org/10.1007/s11430-024-1507-6, 2025.

Kalnay, E., Li, H., Miyoshi, T., Yang, S.-C., and Ballabrera-Poy, J.: 4-D-Var or ensemble Kalman filter?, Tellus A, 59, 758–773, https://doi.org/10.1111/j.1600-0870.2007.00261.x, 2007.

Kwon, Y., Forman, B. A., Ahmad, J. A., Kumar, S. V., and Yoon, Y.: Exploring the utility of machine learning-based passive microwave brightness temperature data assimilation over terrestrial snow in High Mountain Asia, Remote Sens., 11, 2265, https://doi.org/10.3390/rs11192265, 2019.

Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., Collier, N., Ghimire, B., van Kampenhout, L., Kennedy, D., Kluzek, E., Lawrence, P. J., Li, F., Li, H., Lombardozzi, D., Riley, W. J., Sacks, W. J., Shi, M., Vertenstein, M., Wieder, W. R., Xu, C., Ali, A. A., Badger, A. M., Bisht, G., van den Broeke, M., Brunke, M. A., Burns, S. P., Buzan, J., Clark, M., Craig, A., Dahlin, K., Drewniak, B., Fisher, J. B., Flanner, M., Fox, A. M., Gentine, P., Hoffman, F., Keppel-Aleks, G., Knox, R., Kumar, S., Lenaerts, J., Leung, L. R., Lipscomb, W. H., Lu, Y., Pandey, A., Pelletier, J. D., Perket, J., Randerson, J. T., Ricciuto, D. M., Sanderson, B. M., Slater, A., Subin, Z. M., Tang, J., Thomas, R. Q., Val Martin, M., and Zeng, X.: The Community Land Model version 5: Description of new features, benchmarking, and impact of forcing uncertainty, J. Adv. Model Earth Sy., 11, 4245–4287, https://doi.org/10.1029/2018MS001583, 2019.

Legates, D. R., Mahmood, R., Levia, D. F., DeLiberty, T. L., Quiring, S. M., Houser, C., and Nelson, F. E.: Soil moisture: A central and unifying theme in physical geography, Prog. Phys. Geogr., 35, 65–86, https://doi.org/10.1177/0309133310386514, 2011.

Li, L., Bisht, G., and Leung, L. R.: Spatial heterogeneity effects on land surface modeling of water and energy partitioning, Geosci. Model Dev., 15, 5489–5510, https://doi.org/10.5194/gmd-15-5489-2022, 2022.

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

Li, Q., Shi, G., Shangguan, W., Nourani, V., Li, J., Li, L., Huang, F., Zhang, Y., Wang, C., Wang, D., Qiu, J., Lu, X., and Dai, Y.: A 1 km daily soil moisture dataset over China using in situ measurement and machine learning, Earth Syst. Sci. Data, 14, 5267–5286, https://doi.org/10.5194/essd-14-5267-2022, 2022.

Li, R., Zhang, X., Liu, B., and Zhang, B.: Review on methods of remote sensing time-series data reconstruction, National Remote Sensing Bulletin, 13, 335–341, https://doi.org/10.11834/jrs.20090257, 2009.

Li, X. and Liu, F.: A mathematical comparison of data assimilation and machine learning in Earth system state estimation from a Bayesian inference viewpoint, Inf. Geogr., 1, 100001, https://doi.org/10.1016/j.infgeo.2025.100001, 2025.

Li, X., Feng, M., Ran, Y., Su, Y., Liu, F., Huang, C., Shen, H., Xiao, Q., Su, J., Yuan, S., and Guo, H.: Big data in Earth system science and progress towards a digital twin, Nat. Rev. Earth Environ., 4, 319–332, https://doi.org/10.1038/s43017-023-00409-w, 2023.

Liu, Y., Cui, W., Ling, Z., Fan, X., Dong, J., Luan, C., Wang, R., Wang, W., and Liu, Y.: The impact of satellite soil moisture data assimilation on the hydrological modeling of SWAT in a highly disturbed catchment, Remote Sens., 16, 429, https://doi.org/10.3390/rs16020429, 2024.

MacBean, N., Bacour, C., Raoult, N., Bastrikov, V., Koffi, E., Kuppel, S., Maignan, F., Ottlé, C., Peaucelle, M., Santaren, D., and Peylin, P.: Quantifying and reducing uncertainty in global carbon cycle predictions: Lessons and perspectives from 15 years of data assimilation studies with the ORCHIDEE terrestrial biosphere model, Global Biogeochem. Cy., 36, e2021GB007177, https://doi.org/10.1029/2021GB007177, 2022.

Manabe, S.: Climate and the ocean circulation: I. The atmospheric circulation and the hydrology of the Earth's surface, Mon. Weather Rev., 97, 739–774, https://doi.org/10.1175/1520-0493(1969)097<0739:CATOC>2.3.CO;2, 1969.

McLaughlin, D.: Recent developments in hydrologic data assimilation, Rev. Geophys., 33, 977–984, https://doi.org/10.1029/95RG00740, 1995.

Myneni, R., Knyazikhin, Y., and Park, T.: MYD15A2H MODIS/Aqua Leaf Area Index/FPAR 8-day L4 global 500 m SIN grid, NASA LP DAAC [data set], https://doi.org/10.5067/MODIS/MYD15A2H.006, 2015.

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and Xia, Y.: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements, J. Geophys. Res.-Atmos., 116, D12109, https://doi.org/10.1029/2010JD015139, 2011.

Raoult, N., Douglas, N., MacBean, N., Kolassa, J., Quaife, T., Roberts, A. G., Fisher, R. A., Fer, I., Bacour, C., Dagon, K., Hawkins, L., Carvalhais, N., Cooper, E., Dietze, M. C., Gentine, P., Kaminski, T., Kennedy, D., Liddy, H. M., Moore, D. J. P., Peylin, P., Pinnington, E., Sanderson, B. M., Scholze, M., Seiler, C., Smallman, T. L., Vergopolan, N., Viskari, T., Williams, M., and Zobitz, J.: Parameter estimation in land surface models: Challenges and opportunities with data assimilation and machine learning, J. Adv. Model Earth Sy., 17, e2024MS004733, https://doi.org/10.1029/2024MS004733, 2025.

Reichle, R. H., McLaughlin, D. B., and Entekhabi, D.: Variational data assimilation of microwave radiobrightness observations for land surface hydrology applications, IEEE T. Geosci. Remote, 39, 1708–1718, https://doi.org/10.1109/36.942549, 2001.

https://doi.org/10.5194/egusphere-2026-3207
Preprint. Discussion started: 22 July 2026

Reichle, R. H., McLaughlin, D. B., and Entekhabi, D.: Hydrologic data assimilation with the ensemble Kalman filter, Mon. Weather Rev., 130, 103–114, https://doi.org/10.1175/1520-0493(2002)130<0103:HDAWTE>2.0.CO;2, 2002.

Reichle, R. H., Koster, R. D., Liu, P., Mahanama, S. P. P., Njoku, E. G., and Owe, M.: Comparison and assimilation of global soil moisture retrievals from the Advanced Microwave Scanning Radiometer for the Earth Observing System (AMSR-E) and the Scanning Multichannel Microwave Radiometer (SMMR), J. Geophys. Res.-Atmos., 112, D09108, https://doi.org/10.1029/2006JD008033, 2007.

Reichle, R. H., De Lannoy, G. J. M., Liu, Q., Koster, R. D., Kimball, J. S., Crow, W. T., Ardizzone, J. V., Chakraborty, P., Collins, D. W., Conaty, A. L., Girotto, M., Jones, L. A., Kolassa, J., Lievens, H., Lucchesi, R. A., and Smith, E. B.: Global assessment of the SMAP Level-4 surface and root-zone soil moisture product using assimilation diagnostics, J. Hydrometeorol., 18, 3217–3237, https://doi.org/10.1175/JHM-D-17-0130.1, 2017.

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

Shen, Z., Ramirez-Lopez, L., Behrens, T., Cui, L., Zhang, M., Walden, L., Wetterlind, J., Sudduth, K. A., Baumann, P., Song, Y., Catambay, K., and Viscarra Rossel, R. A.: Deep transfer learning of global spectra for local soil carbon monitoring, ISPRS J. Photogramm. Remote Sens., 188, 190–200, https://doi.org/10.1016/j.isprsjprs.2022.04.009, 2022.

Sun, L., Seidou, O., Nistor, I., and Liu, K.: Review of the Kalman-type hydrological data assimilation, Hydrol. Sci. J., 61, 2348–2366, https://doi.org/10.1080/02626667.2015.1127376, 2016.

Turlapaty, A. C., Younan, N. H., and Anantharaj, V. G.: Interpolation of missing values in AMSR-E soil moisture data using modified SSA, IEEE Geosci. Remote S., 8, 322–325, https://doi.org/10.1109/LGRS.2010.2071852, 2010.

Wang, W., Cui, W., Wang, X., and Chen, X.: Evaluation of GLDAS-1 and GLDAS-2 forcing data and Noah model simulations over China at the monthly scale, J. Hydrometeorol., 17, 2815–2833, https://doi.org/10.1175/JHM-D-15-0191.1, 2016.

Wang, W., Brönnimann, S., Zhou, J., Li, S., and Wang, Z.: Near-surface air temperature estimation for areas with sparse observations based on transfer learning, ISPRS J. Photogramm. Remote Sens., 220, 712–727, https://doi.org/10.1016/j.isprsjprs.2025.01.021, 2025.

Yang, K., Chen, Y.-Y., and Qin, J.: Some practical notes on the land surface modeling in the Tibetan Plateau, Hydrol. Earth Syst. Sci., 13, 687–701, https://doi.org/10.5194/hess-13-687-2009, 2009.

Yang, K., Jiang, Y., Tang, W., He, J., Shao, C., Zhou, X., Lu, H., Chen, Y., Li, X., and Shi, J.: A high-resolution near-surface meteorological forcing dataset for the Third Pole region (TPMFD, 1979–2023), National Tibetan Plateau/Third Pole Environment Data Center [data set], https://doi.org/10.11888/Atmos.tpdc.300398, 2023.

Yang, Z.-L., Niu, G.-Y., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Longuevergne, L., Manning, K., Niyogi, D., Tewari, M., and Xia, Y.: The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins, J. Geophys. Res.-Atmos., 116, D12110, https://doi.org/10.1029/2010JD015140, 2011.

Zhang, R., Jiang, D., Zhang, Z., and Yu, E.: The impact of regional uplift of the Tibetan Plateau on the Asian monsoon climate, Palaeogeogr. Palaeocl., 417, 137–150, https://doi.org/10.1016/j.palaeo.2014.10.030, 2015.

Zhang, T., Chang, J., Zhang, J., Shao, Q., Sarukkalige, R., Luo, Z., Bi, C., Wang, J., Hu, Y., Li, B., and Liang, Z.: Enhanced sequential soil moisture estimation using an ensemble deep learning filter without linear-Gaussian constraints, J. Hydrol., 666, 134810, https://doi.org/10.1016/j.jhydrol.2025.134810, 2026.

Zhu, J.: Dataset of land cover over the Tibetan Plateau from 2001 to 2020, National Tibetan Plateau/Third Pole Environment Data Center [data set], https://doi.org/10.5067/MODIS/MCD12Q1.006, 2022.

Zhu, S., Zha, G., Wang, Q., Ma, S., and Qin, H.: A high-performance assimilation of surface soil moisture based on a hybrid framework of machine learning and a physical hydrological model, J. Hydrol., 664, 134513, https://doi.org/10.1016/j.jhydrol.2025.134513, 2026.