JGR | Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2025JH001121

Special Collection:
Advancing Interpretable AI/ML Methods for Deeper Insights and Mechanistic Understanding in Earth Sciences: Beyond Predictive Capabilities

Improving Global Surface Soil Moisture Prediction Through Physics-Guided Deep Learning and Cluster-Based Regionalization

Xuan Xi[1] **and Qianlai Zhuang**[1,2]

[1]Department of Earth, Atmospheric, and Planetary Sciences, Purdue University, West Lafayette, IN, USA, [2]Department of Agronomy, Purdue University, West Lafayette, IN, USA

Key Points:

- A physics-guided deep learning model outperforms both process-based and deep learning models in global surface soil moisture predictions
- An optimized clustering strategy enhances generalization by creating environmentally consistent training samples across the globe
- The physics-guided deep learning model shows stronger water balance consistency and stays robust after excluding uncertain observations

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

Q. Zhuang,
qzhuang@purdue.edu

Citation:

Xi, X., & Zhuang, Q. (2026). Improving global surface soil moisture prediction through physics-guided deep learning and cluster-based regionalization. *Journal of Geophysical Research: Machine Learning and Computation*, *3*, e2025JH001121. https://doi.org/10.1029/2025JH001121

Received 5 NOV 2025
Accepted 7 MAY 2026

Abstract Surface soil moisture (SSM) is essential to the hydrological cycle and land–atmosphere interactions, and its accurate simulation is crucial for climate prediction and resource management. This study developed an innovative modeling framework for global SSM prediction by integrating physics-guided deep learning (PGDL) and clustering-based regionalization. The PGDL model combines the physical knowledge from the Terrestrial Ecosystem Model (TEM) and the temporal learning capacity of long short-term memory (LSTM) networks. By introducing a clustering strategy based on multi-source features, the global land was divided into subregions with consistent characteristics. Within this framework, cluster-specific models were trained using in situ observations and evaluated at the global scale against independent satellite observations. This clustering approach enhanced model generalization across diverse climatic and geographic conditions, yielding more robust predictions based on environmentally consistent samples. Results show that the PGDL model (RMSE: 0.081, r: 0.55) outperformed both the process-based (PB) model (RMSE: 0.167, r: 0.43) and the purely deep learning (DL) model (RMSE: 0.085, r: 0.40) at the global scale, while also exhibiting stronger physical consistency with water balance diagnostics. After excluding regions with high uncertainty in SSM observations, the performance of all models improved, with PGDL maintaining the best performance (masked RMSE: 0.064, masked r: 0.58). Overall, this study demonstrates the superiority of the PGDL model and highlights the importance of clustering strategies in model construction and evaluation for achieving more accurate and robust SSM predictions across heterogeneous environments.

Plain Language Summary Soil surface moisture plays a key role in the water cycle, weather, and agriculture, but it is difficult to predict accurately across the globe. In this study, we employed a hybrid modeling framework that combines knowledge of hydrological processes and modern machine learning methods. The global land was divided into regions with similar environmental conditions and models were trained for each region with ground observations and evaluated against satellite data at the global scale. Compared with traditional models, our framework produced more accurate and physically consistent soil moisture estimates, especially when accounting for uncertainties in satellite observations. These results show that integrating physical science with machine learning can lead to more reliable soil moisture prediction, improving global resource management and our understanding of land–climate interactions.

1. Introduction

Surface soil moisture (SSM) plays a critical role in the terrestrial hydrological cycle by regulating surface water and energy fluxes and influencing vegetation growth, agricultural productivity, and carbon cycle feedback (Green et al., 2019; Méndez-Barroso et al., 2009; Rosenzweig et al., 2002). Accurate simulation of spatiotemporal dynamics of SSM is essential for understanding land-atmosphere interactions and supporting water resource management and climate change research (Berg & Sheffield, 2018; Dobriyal et al., 2012; McColl et al., 2017; Sun et al., 2025). However, the complex controlling factors and strong regional heterogeneity of SSM make global-scale modeling a long-standing challenge (Seneviratne et al., 2010; Vereecken et al., 2014).

Two main approaches have been developed for soil moisture simulation: process-based (PB) models and deep learning (DL) models. PB models rely on mathematical representations of physical processes and provide strong interpretability, but their complex structures, large number of parameters, and high sensitivity to input data quality limit their applicability at large scales (Clark et al., 2016; Wood et al., 2011). In contrast, DL models can effectively capture complex dynamic patterns by learning nonlinear relationships from large data sets, but they

often lack physical constraints, which can undermine their credibility and generalizability (Karpatne et al., 2017; Reichstein et al., 2019). To address these limitations, a new modeling paradigm—physics-guided deep learning (PGDL)—has emerged in recent years. PGDL integrates physical knowledge into deep neural network structures, aiming to combine physical consistency with data-fitting capability. This approach has demonstrated promising performance across a range of domains (Willard et al., 2022). Recent studies have also explored the potential of PGDL for soil moisture prediction (Breen et al., 2020; Li et al., 2024; Zhang et al., 2025). However, these efforts have largely been limited to regional scales and lack global-scale applications, thus constraining improvements in generalization ability.

In recent years, remote sensing products have played a crucial role in enabling large-scale SSM monitoring and have supported the development of global SSM modeling efforts. Examples such as NNsm (Yao et al., 2021), SoMo.ml (O & Orth, 2021), and GSSM1km (Han et al., 2023) have made significant progress in producing long-term, high-resolution, and spatially extensive SSM data sets, thereby laying a solid foundation for data-driven modeling. However, these methods primarily focus on applying machine learning techniques to reconstruct or fit SSM time series, often without explicitly incorporating physical processes into the modeling framework. While they achieve good predictive performance, limitations remain in terms of physical interpretability and consistency. Methodologically, NNsm uses satellite data as training features and relies on remote sensing observations as targets, lacking explicit representation of physical processes and thorough validation of transferability across different data sources. Additionally, both NNsm and GSSM1km adopt temporal train–test splits, making it difficult to assess spatial generalization performance. Therefore, explicitly incorporating physical information with flexible model structures across diverse environmental conditions is required to provide a more adaptive global-scale SSM modeling.

The distribution of SSM and its responses to environmental drivers exhibit pronounced spatial heterogeneity. This heterogeneity is linked not only to static attributes such as soil type and surface conditions but also to the nonlinear response of SSM to dynamic drivers (Lal et al., 2023; Sun & Niu, 2019). Previous studies have shown that, compared to unified modeling strategies, structured clustering approaches that partition the modeling domain into relatively homogeneous subregions based on environmental characteristics are better suited to capture regional variability, thereby improving both model performance and interpretability (Dunkl & Ließ, 2022; Kratzert et al., 2019). In practice, clustering has been widely adopted to support regional segmentation and model optimization for improving SSM predictions. For example, a recent study developed a cluster-based local modeling method that integrates clustering with neural network architectures, significantly improving the accuracy of SSM estimation (Moosavi et al., 2024). Another study employed cluster-based sampling strategies to balance training data distributions and effectively mitigate sample bias (Li et al., 2024). Collectively, these findings highlight that clustering-based regionalization enables more representative training structures and enhances model generalization.

Building upon our previously developed PGDL framework for SSM prediction (Xi et al., 2025), this study advances a global SSM modeling approach that integrates clustering-based regionalization with PGDL. Specifically, global land grid cells are first clustered based on static and dynamic features to identify environmentally and climatically similar regions and form structurally consistent clusters. A dedicated PGDL model is then trained for each cluster. To evaluate the spatial generalization capability of the models, we employ high-quality in situ SSM data from the International Soil Moisture Network (ISMN; Dorigo et al., 2011, 2013) during 2016–2020 to conduct leave-one-site-out (LOSO) cross-validation and apply the trained models to all grid cells within the corresponding cluster. For independent global-scale evaluation, we further compare the model outputs against the Soil Moisture Active Passive (SMAP; Entekhabi et al., 2010) SSM product, which is used as an independent observational reference to evaluate model performance. This design ensures that SMAP is used only to assess the large-scale performance and transferability of the models. The PGDL framework explicitly incorporates key intermediate physical processes governing SSM dynamics, providing a physically informed model structure that links data-driven learning with process understanding. We systematically compare the predictive performance of the PB, DL, and PGDL models, and further examine the physical consistency of the latter two models to enhance scientific interpretability. Overall, this study proposes a modeling framework that integrates PGDL with clustering-based regionalization, combining physical constraints with spatial generalization capability and offering a novel pathway for simulating global SSM dynamics from sparse ground-based observations.

10.1029/2025JH001121

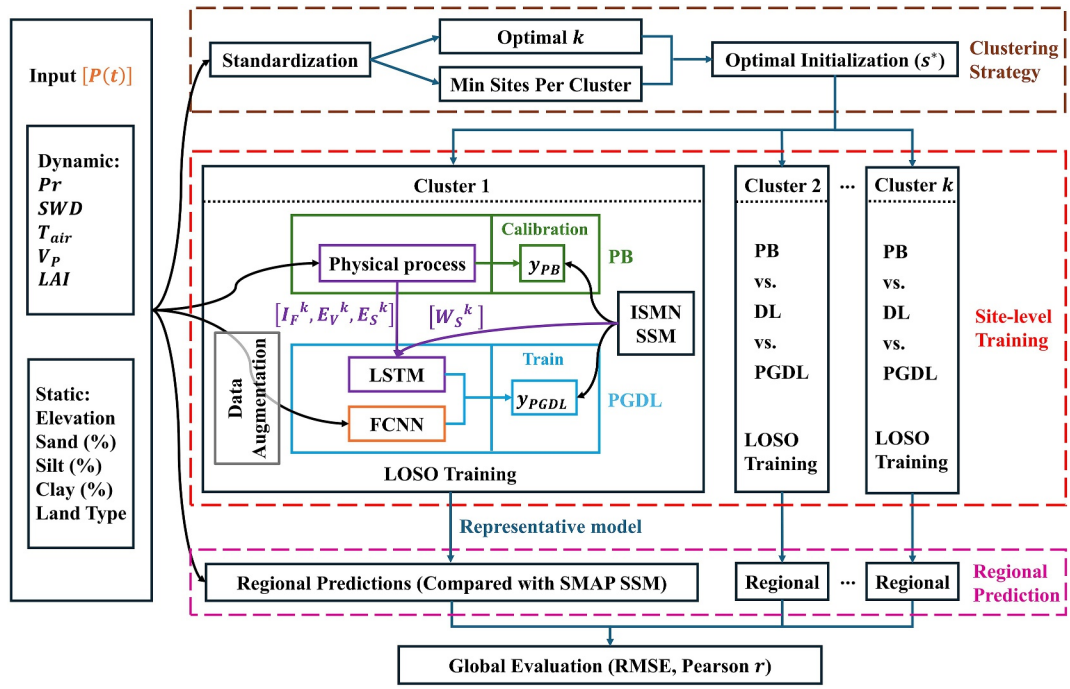

Figure 1. Workflow of the proposed methodology for global SSM prediction and evaluation. The framework integrates multi-source inputs, clustering-based regionalization, and comparative analysis across PB, DL, and PGDL models. Ten input features are first standardized and employed to derive the optimal clustering structure, defined by the number of clusters (k) and a minimum number of sites per cluster, with the initialization subsequently selected based on the Silhouette Score. Within each cluster, DL and PGDL models are trained and evaluated in a cluster-specific manner, with a data augmentation step applied prior to site-level training when multiple ISMN sites share identical inputs within the same grid cell. The schematic of Cluster 1 presents the PGDL model architecture and its linkages to the PB model. The green section displays the PB process flow, and the blue section corresponds to the PGDL process. The purple box denotes the physical process represented in the PB model, with solid purple lines showing the coupling to the PGDL model through the transfer of $I_F{}^k$, $E_V{}^k$, and $E_S{}^k$ from PB outputs and $W_S{}^k$ from observations into the LSTM branch. The orange box represents the FCNN module processing the external inputs $P(t)$ in the PGDL model. Both PB and PGDL models generate outputs (y_{PB} and y_{PGDL}) with their respective parameterization, and these outputs are evaluated against ISMN SSM observations through their respective calibration or training procedures. For each cluster, model performance is assessed under a LOSO scheme, and the best-performing model is selected as the representative model for regional prediction and validated against SMAP SSM. Finally, results across all clusters are aggregated to obtain a global evaluation.

Section 2 introduces the data collection, including SSM observations and input data, followed by the clustering strategy along with the PB, DL, and PGDL models and the experimental design for model evaluation. Section 3 presents the clustering distribution and then reports the model performance from both site-level and global perspectives. Finally, Section 4 provides detailed discussions in three aspects, focusing, respectively, on the effectiveness of the clustering strategy, the physical consistency of DL and PGDL predictions, and the influence of SMAP uncertainties on model assessments.

2. Methods

The overall workflow of the proposed methodology is summarized in Figure 1, covering data collection (Section 2.1), clustering strategy (Section 2.2), model development (Section 2.3), and model training and evaluation followed by regional prediction and global assessment (Section 2.4).

2.1. Data Collection

2.1.1. Observation of Surface Soil Moisture

2.1.1.1. ISMN Soil Moisture

This study employs in situ SSM observations from the ISMN as supervised labels for model training and testing. ISMN is a globally integrated database that provides quality-controlled soil moisture measurements with broad spatial coverage and high data reliability. It has been widely used in soil moisture research and validation of satellite-derived products.

We selected SSM observations from 1 January 2016, to 31 December 2020 (1,827 days), retaining only records of good quality. The original hourly observations were first aggregated to daily resolution to match the temporal scale of the models. To ensure data completeness and reliability, we excluded sites with more than 182 missing days (approximately half a year) or continuous gaps longer than 30 days. Missing values were then filled using linear interpolation, and the proportion of missing data in the selected sites was small (approximately 0.4% of the total time series). We further removed invalid interpolated values as well as sites exhibiting unrealistic temporal trends. In addition, to ensure spatial consistency, only sites located within the coverage of the SMAP satellite product were retained. After screening, a total of 484 ISMN sites were selected for site-level training. The complete list of selected sites is provided in Table S1 of Supporting Information S1.

2.1.1.2. SMAP Soil Moisture

This study utilizes SSM products from the SMAP satellite launched by the National Aeronautics and Space Administration (NASA) in 2015 (Entekhabi et al., 2010). Equipped with an L-band microwave radiometer, SMAP provides global observations of near-surface (0–5 cm) soil moisture.

We used the SMAP Level 3 daily passive microwave soil moisture product (Version 8), which is generated by interpolating and resampling Level 2 swath data onto a global grid with a native spatial resolution of approximately 36 km (O'Neill et al., 2021). For direct comparison with the PB, DL, and PGDL model outputs at 0.5° resolution, the SMAP data were aggregated to a 0.5° × 0.5° global grid using a nearest-neighbor method. This spatial aggregation was performed solely for resolution matching and does not aim to modify the SMAP product. The temporal coverage was matched to the ISMN data (2016–2020) for model training and evaluation. After resampling, each SMAP grid cell provides daily SSM values that can be directly compared with model predictions. The final data set consists of 56,112 valid land grid cells, which define the global prediction domain in this study.

2.1.2. Input Data

The input data include dynamic and static variables representing soil properties, climatic dynamics, and landscape conditions. The dynamic input comprises five variables: precipitation (Pr, units: mm), surface downwelling shortwave radiation (SWD, units: W m^{-2}), near-surface air temperature (T_{air}, units: °C), vapor pressure (V_P, units: hPa), and leaf area index (LAI). Vapor pressure was derived from air temperature and relative humidity. The climate data were obtained from the 20CRv3-ERA5 data set provided by the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP), with a spatial resolution of 0.5° × 0.5° and a daily time step (Lange et al., 2022). LAI data were sourced from the ERA5 hourly single-level data set (Hersbach et al., 2023) at an original spatial resolution of 0.25° × 0.25°. We converted the same LAI data to monthly scales for the PB model and daily scales for the DL and PGDL models. The static inputs describe spatial attributes for each grid cell and include elevation, soil sand content, soil silt content, soil clay content, and land type (1 indicates wetlands and 0 indicates uplands) (Zhuang et al., 2011). These features vary spatially but remain temporally invariant, providing an environmental context for regional generalization.

All models in this study employed the same input variables to ensure consistency and comparability of results. In addition, the features used for clustering were identical to those used in the models, thereby maintaining alignment between spatial partitioning and model inputs.

2.1.3. Data Augmentation

When mapping ISMN sites to their corresponding 0.5° × 0.5° grid cells, multiple sites may fall within the same cell. In such cases, these sites share identical dynamic inputs, which constrains the ability to fully exploit SSM observations from each site during model training.

To overcome this limitation, we applied a data augmentation scheme for sites located within the same grid cell. Specifically, for each cell containing multiple ISMN sites, we first adopted a spatial distance-based optimization procedure to identify the site closest to the grid center and assigned it the original dynamic inputs (i.e., Pr, SWD, T_{air}, V_P, and LAI). For the remaining sites, we introduced controlled perturbations to the dynamic inputs to treat them as independent samples while preserving physical plausibility.

Among the dynamic inputs, Pr typically exhibits a highly skewed distribution with frequent near-zero values and occasional extremes. To account for this property, we applied multiplicative Gaussian perturbations to more realistically capture its inherent variability and uncertainty. For the other four variables with relatively stable distributions, additive Gaussian perturbations were applied. The perturbation intensity for each variable was scaled by its site-specific mean to ensure a reasonable range. Non-negativity constraints were imposed on all variables except T_{air} to avoid physically inconsistent samples.

This data augmentation strategy produces additional training samples while maintaining the original temporal structure and statistical properties. Consequently, all 484 ISMN sites were incorporated into the training as independent and valid samples, thereby maximizing the utility of available SSM observations.

2.2. Clustering Strategy

To enable global-scale SSM prediction using limited in situ observations, this study applied k-means clustering of all valid global land grid cells (56,112 in total) to support regional model training and spatial generalization.

K-means is a widely used unsupervised learning algorithm that partitions data into k clusters by minimizing within-cluster variance (Hu et al., 2023; MacQueen, 1967). The algorithm begins with the random initialization of k centroids. Each data point is then assigned to the nearest centroid based on distance, after which centroids are updated as the mean of the points assigned to them. This assignment–update process is repeated iteratively until convergence, typically defined by centroid stabilization or the attainment of a maximum number of iterations.

The following subsections describe the clustering process in detail, including the clustering input features (Section 2.2.1), the optimal number of clusters (Section 2.2.2), the minimum number of ISMN sites per cluster (Section 2.2.3), the clustering initialization (Section 2.2.4), and the evaluation of clustering effectiveness (Section 2.2.5).

2.2.1. Clustering Input Features

The features used for clustering are consistent with those used as model inputs, comprising five dynamic variables (Pr, SWD, T_{air}, V_P, and LAI) and five static variables (elevation, soil sand content, soil silt content, soil clay content, and land type) (see Section 2.1.2 for details). Given that the k-means algorithm requires each sample to be represented by a fixed-dimensional input vector, we used the multi-year mean values of the dynamic variables over 2016–2020 as clustering inputs to represent regional climate characteristics while reducing the influence of short-term variability and extremes. This feature combination provides a physically interpretable characterization of the environment and ensures consistency with the model inputs. All input features were standardized prior to clustering to eliminate differences in scale and units across variables.

2.2.2. Optimal Number of Clusters

To achieve reasonable partitioning that ensures both internal consistency within clusters and clear separation between clusters, an appropriate number of clusters (k) is required for constructing the spatial clustering structure. To this end, we clustered all sample points for each integer value of k from 2 to 20. The MiniBatch k-means algorithm (Sculley, 2010) was applied to improve computational efficiency for large-scale data. To mitigate the instability caused by the random centroid initialization, each k value was evaluated across 300 different initialization settings.

After each clustering run, we calculated the corresponding Silhouette Score (Rousseeuw, 1987) as a measure of clustering quality. The Silhouette Score integrates both intra-cluster cohesion and inter-cluster separation and is defined as

$$\text{Silhouette Score} = \frac{1}{N}\sum_{i=1}^{N}\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{1}$$

where $a(i)$ is the average distance from sample i to all other points within the same cluster, $b(i)$ is the minimum average distance from sample i to all points in any other cluster, and N is the total number of samples. A higher Silhouette Score indicates a more well-defined clustering structure.

For each initialization, we recorded the k value that achieved the highest Silhouette Score and then counted the number of times each k was identified as optimal across the 300 independent runs. This count-based strategy reduces reliance on any single clustering outcome and enhances the robustness of the optimal k selection. As shown in Figure S1 of Supporting Information S1, $k = 11$ had the highest count and was therefore chosen as the optimal number of clusters in this study. The subsequent initialization selection and related analyses were based on this configuration.

2.2.3. Minimum Number of ISMN Sites per Cluster

To ensure feasible and stable model training within each cluster and support the subsequent LOSO strategy, the minimum number of ISMN sites required per cluster was determined. Although LOSO theoretically requires only two sites (one for training and one for testing), such a minimal training set is insufficient for representative and robust model evaluation in practice. Therefore, we designed a systematic experiment to examine how model performance varies with sample size and empirically determine a reasonable lower bound for the number of required sites. This threshold was then used as a constraint in the selection of clustering initialization.

We first constructed a representative sample set from the 484 ISMN sites to evaluate the empirical lower bound of required sample size under ideal conditions with highly consistent SSM characteristics. Specifically, based on the SSM time series at each site, we calculated three statistical metrics: correlation, mean, and standard deviation. A composite similarity distance was then defined as a weighted combination of these metrics with weights of 0.4, 0.3, and 0.3, respectively. Using this metric, we applied a greedy algorithm to iteratively construct a candidate set of 15 sites: starting with a randomly selected site, we successively added the site with the highest average similarity (i.e., the smallest average distance) to the current set in each iteration until 15 sites were included. The choice of 15 sites was based on the maximum number of available sites per cluster under the current conditions of clustering inputs and ISMN data. To enhance robustness, the selection process was repeated with 10 different initial seeds, and the final set with the highest overall similarity score was retained. This procedure resulted in a candidate set with high internal consistency in statistical characteristics.

After obtaining this representative set of 15 sites, we systematically evaluated how model performance varied with sample sizes $n \in [3, 15]$. For each value of n, we randomly selected the corresponding number of sites from the candidate pool (with $n = 15$ representing the full set). To ensure robustness, the training and evaluation process for each n was independently repeated across multiple randomly sampled subsets. In each run, both the DL and PGDL models were trained using a LOSO strategy, where $n - 1$ sites were used for training and the remaining site for testing, rotating until each site served once as the test set. For each test site, we calculated the Root Mean Square Error (RMSE) and Pearson correlation coefficient (r) for the DL and PGDL models and then aggregated the results across each n to analyze performance trends.

As shown in Figure S2 of Supporting Information S1, within the selected site set, RMSE decreases and r increases as n increases, with both metrics tending to stabilize around $n = 10$. This indicates that model training becomes sufficiently stable and representative when $n \geq 10$. Based on this finding, we set $n = 10$ as the empirical lower bound for the number of ISMN sites required per cluster and used as a constraint in clustering initialization.

2.2.4. Clustering Initialization

We determined the optimal number of clusters as $k = 11$ in Section 2.2.2 and established that each cluster must contain at least 10 ISMN sites in Section 2.2.3. Building on these foundations, this section identifies a well-

structured clustering initialization based on large-scale trials. Since centroid initialization in k-means clustering is inherently random, different initialization settings can lead to slight variations in the resulting cluster structures. To ensure a robust configuration, we designed a systematic selection strategy to identify the optimal initialization from a pool of candidates.

Specifically, with the number of clusters fixed at $k = 11$, we applied the MiniBatch k-means algorithm with standardized input features to cluster all 56,112 land grid cells across numerous initialization settings. For each result, cluster labels were mapped to the 484 ISMN sites to determine the number of sites within each cluster. Configurations in which any cluster contained fewer than 10 sites were excluded. Among the remaining candidates, we calculated the Silhouette Score of each candidate and selected the initialization with the highest score as the optimal clustering configuration. This process is formally described as follows:

$$s^* = \underset{s \in S'}{\text{argmax}} \text{Silhouette}(s) \tag{2}$$

where s^* denotes the optimal clustering initialization, s denotes an individual initialization corresponding to a clustering attempt, and S' denotes the set of all initializations that satisfy the minimum site count constraint, defined as:

$$S' = \left\{ s \in S \middle| \begin{array}{c} \min\limits_{1 \le j \le k} N_j^{(s)} \ge n_{\min} \\ \text{Label}_s = \text{KMeans}(\mathbf{X}; k, s) \end{array} \right\} \tag{3}$$

where $\mathbf{X}$ denotes the standardized feature matrix used for clustering, $N_j^{(s)}$ denotes the number of ISMN sites in the j-th cluster under initialization s, $n_{\min}$ denotes the predefined minimum number of required sites per cluster (set to 10 in this study), and Label_s denotes the cluster labels generated by the k-means algorithm with initialization s. This strategy ensures that each selected cluster at least contains the required number of samples while prioritizing structural separation.

To implement the above selection process, we executed a total of 100,000 clustering runs with different initialization settings. This setting was motivated by two considerations: first, such many runs provide broad coverage of the initialization space enabling a comprehensive exploration of potential clustering structures; second, additional tests with 20,000 new initializations did not identify any configuration that outperformed the existing best under the Silhouette Score and sample distribution constraints. This suggests that the 100,000 attempts provided sufficient coverage to ensure stable initialization selection while maintaining computational efficiency. Finally, the clustering result with the highest Silhouette Score (s^*) among those that met the constraints was selected for subsequent model training and global simulation.

2.2.5. Evaluation of Clustering Effectiveness

To verify the effectiveness of the constructed clustering structure with respect to the target variable, we evaluated both intra-cluster consistency and inter-cluster distinctiveness. Intra-cluster consistency measures the similarity in dynamic SSM behavior among sites within the same cluster, whereas inter-cluster distinctiveness examines whether significant differences in SSM exist between clusters. A joint evaluation of these two aspects provides a comprehensive assessment of the interpretability and physical relevance of the clustering structure.

2.2.5.1. Intra-Cluster Consistency

To evaluate intra-cluster consistency in SSM dynamics, we calculated the average pairwise r of SSM time series among sites within each cluster. As a baseline, we calculated the pairwise r across all 484 ISMN sites and obtained an overall mean of 0.27, reflecting substantial global heterogeneity in SSM dynamics.

Building on this, for each cluster, we calculated the average pairwise correlation of SSM time series among all included sites based on the selected clustering initialization. To reflect the overall clustering performance, we further defined a weighted average correlation coefficient ($\bar{r}_w$) as follows:

$$\bar{r}_w = \frac{\sum_{i=1}^{k} n_i \cdot \bar{r}_i}{\sum_{i=1}^{k} n_i} \quad \text{where} \quad n_i = \frac{N_i(N_i - 1)}{2} \tag{4}$$

In this formula, k denotes the total number of clusters ($k = 11$), $\bar{r}_i$ is the average Pearson correlation coefficient within cluster i, N_i is the number of sites in cluster i, and n_i is the number of site pairs in that cluster. This metric weights the contribution of each cluster by its sample size, providing a comprehensive measure of consistency in SSM dynamics across the clustering structure.

The results show that $\bar{r}_i$ ranges from 0.26 to 0.53 (see Table S2 in Supporting Information S1 for the detailed value of each $\bar{r}_i$), indicating that the spatial grouping based on input features also exhibits strong internal consistency in SSM dynamics. The weighted average correlation across the 11 clusters ($\bar{r}_w = 0.39$) is higher than the global average correlation (0.27) among all 484 ISMN sites, indicating the effectiveness of the clustering in capturing regional SSM coherence.

2.2.5.2. *Inter-Cluster Distinctiveness*

To evaluate whether the constructed clustering structure effectively distinguishes dynamic SSM behaviors, we performed a one-way analysis of variance (ANOVA) to assess the statistical differences in SSM characteristics among clusters. ANOVA partitions the total variance into inter-cluster and intra-cluster components and calculates their ratio (the F-statistic) to test whether differences among clusters are statistically significant. In this study, the mean and standard deviation of the SSM time series were used to represent the overall level and variability of SSM, respectively. Based on the selected clustering configuration, we calculated these statistics for each ISMN site and conducted separate ANOVA tests to assess differences among clusters.

The results show significant differences among clusters in both the mean ($F = 18.35$; $p < 0.001$) and standard deviation ($F = 5.76$; $p < 0.001$) of SSM, indicating distinct SSM dynamics across clusters under the selected clustering configuration.

Together, the intra-cluster consistency and inter-cluster distinctiveness analyses indicate that the selected clustering structure exhibits strong internal coherence while capturing meaningful differences in SSM dynamics across clusters. These findings provide a solid foundation for subsequent cluster-based modeling and regional prediction.

2.3. Model Description

2.3.1. Process-Based (PB) Model

The process-based simulation in this study is conducted using the Terrestrial Ecosystem Model (TEM), a modeling framework that quantifies water, carbon, and nitrogen exchanges between the atmosphere and terrestrial ecosystems (Zhuang et al., 2004, 2011). Specifically, we use its coupled Water Balance Model (WBM; Zhuang et al., 2002) to simulate daily soil moisture dynamics on the global scale. The WBM applies a physically based water balance approach, accounting for infiltration, transpiration, evaporation, and drainage processes.

The WBM represents soil moisture (W_S) dynamics through a unified water balance equation at each grid cell:

$$\frac{\partial W_S}{\partial t} = I_F - E_V - E_S - D_R \tag{5}$$

where t means time, I_F is infiltration, E_V is evapotranspiration from the vegetation, E_S is evaporation from the soil surface, and D_R is drainage. These processes jointly determine temporal changes in soil moisture.

To account for hydrological differences between soil types, the WBM employs two parameterization schemes within the unified water balance framework, one for upland soils and the other for wetland soils. In the WBM, SSM is defined as the average volumetric water content in the top 10 cm of soil, with the appropriate scheme selected based on the inundation conditions of each grid cell.

For upland soils, the model divides the vertical soil profile into six layers, each with distinct thicknesses and hydraulic parameters. In the daily simulation, soil moisture is first updated using the water balance equation and

then redistributed vertically through the Richards equation, with the tridiagonal matrix algorithm used to numerically solve the SSM dynamics. For wetland soils, the model employs a two-zone structure defined by the groundwater level: an unsaturated upper zone and a saturated lower zone. It first updates the total water volume based on the water balance equation, then determines the water table depth, and finally estimates SSM as a weighted average of the moisture profile in the topsoil layers, governed by the water table depth. Despite the structural differences between upland and wetland representations, the SSM states in both cases are governed by the same water balance equation, ensuring a consistent physical foundation for unified modeling.

2.3.2. Deep Learning (DL) Model

In this study, the Long Short-Term Memory (LSTM) network (Hochreiter & Schmidhuber, 1997) is employed to capture the temporal dynamics of SSM. LSTM incorporates an internal memory mechanism that enables learning long-term dependencies in time series data, effectively addressing the information loss problem in traditional recurrent neural networks when modeling long sequences.

The core of LSTM lies in its gating mechanism, which dynamically regulates the flow of information through three gates: the input gate controls how new information enters the memory cell, the forget gate determines how much of the previous state is retained, and the output gate governs how much of the cell state contributes to the output. This mechanism enables LSTM to extract relevant features and suppress irrelevant noise in complex time series, making it suitable for modeling the long-term dependencies inherent in SSM dynamics.

This study adopts standard LSTM architecture without altering the internal gating mechanism. LSTM thus serves as the baseline data-driven model and is integrated into the hybrid framework developed for SSM prediction.

2.3.3. Physics-Guided Deep Learning (PGDL) Model

SSM dynamics are influenced by multiple hydrological processes. According to the water balance equation (see Equation 5), SSM is primarily driven by I_F, E_V, E_S, and D_R. Previous studies have also shown that soil moisture exhibits a "memory" effect (Koster & Suarez, 2001), meaning that its current state depends not only on present inputs but also on historical conditions. Based on this understanding, we adopt a hybrid modeling framework—PGDL—that integrates physical mechanisms with temporal dynamics for SSM prediction. This study builds upon the PGDL model structure proposed in our earlier work (Xi et al., 2025) and extends its application to global-scale SSM simulations. In this framework, physical information is incorporated through the model architecture building with physically meaningful intermediate variables, enabling the network to learn physically plausible SSM dynamics through physics-guided model design without explicitly imposing physical constraints on the loss function.

The PGDL model consists of two branches, each designed to handle different types of input variables. One branch is a time-series module based on LSTM, which takes input from the past k days of historical SSM data (denoted as ${W_S}^k$) and the historical sequences of intermediate variables I_F, E_V, and E_S (denoted as ${I_F}^k$, ${E_V}^k$, and ${E_S}^k$). To ensure temporal alignment, the SSM sequence precedes the intermediate variables by 1 day: ${W_S}^k$ spans from $t-1$ to $t-k$, while ${I_F}^k$, ${E_V}^k$, and ${E_S}^k$ span from t to $t-k+1$. The LSTM captures temporal dependencies in SSM dynamics from these input sequences.

Another branch is a fully connected neural network (FCNN), which is designed to account for the impact of uncertainty in the physical model on the learning of time-series features in LSTM. Within the PGDL framework, a data-driven latent correction term U^k is learned through the FCNN and incorporated as an additional input to complement the process-based component by capturing residual information i.e. not explicitly represented in the physical formulation. This term aggregates residual effects associated with hydrological processes that are difficult to represent explicitly (e.g., D_R) as well as structural biases in the intermediate variables, and is inferred from the same physical drivers $P(t)$ as I_F, E_V, E_S. This latent correction complements the process-based component while preserving the physics-guided structure of the framework. Specifically, $P(t)$ denotes the physical driver at the current time step, including five dynamic features (Pr, SWD, T_{air}, V_P, and LAI) and five static features (elevation, soil sand content, soil silt content, soil clay content, and land type). These variables are consistent with those used in the clustering process. The overall PGDL model structure can be formally expressed as

$$W_S(t) = \mathcal{N}\left(W_S{}^k, I_F{}^k, E_V{}^k, E_S{}^k, P(t)\right) \quad (6)$$

Here, $W_S{}^k$ represents the historical SSM states derived from observations, and $I_F{}^k$, $E_V{}^k$, and $E_S{}^k$ denote sequences of intermediate variables obtained from the process-based model.

Finally, the outputs from the LSTM and FCNN branches are concatenated to form a joint feature representation, which is then mapped to the final output (i.e., the predicted SSM) through an additional fully connected layer. By combining these two sources of information, the PGDL leverages the LSTM's strength in capturing temporal dependencies together with the FCNN's ability to represent complex relationships among physical drivers. Further details of the PGDL architecture can be found in our previous work (Xi et al., 2025).

In practice, all dynamic input features were standardized using Z-score normalization to eliminate the effects of differing units and scales. To further stabilize the training labels, a one-dimensional Kalman filter was applied to the observed SSM values used as labels to reduce the influence of observational noise.

Through this physics-guided design, the PGDL model integrates process-based and data-driven approaches to simulate SSM spatiotemporal dynamics. The PGDL model was trained separately within each cluster and subsequently applied to regional predictions for that cluster at the global scale.

2.4. Experiment

2.4.1. For PB Model

To ensure a fair comparison between the PB model and the DL/PGDL models, and to provide a reliable physical foundation for the PGDL framework, we calibrated the PB model by optimizing key parameters that affect SSM simulation accuracy. While the PB model employs a unified water balance equation to describe soil moisture dynamics, it adopts different parameterization schemes for upland and wetland soils.

Specifically, for upland sites, we calibrated seven key parameters: θ_{sat} (saturated volumetric water content), k_{sat} (saturated hydraulic conductivity), ψ_{sat} (saturated matric potential), b (Clapp-Hornberger parameter), I_{RMAX} (canopy rainfall interception parameter), g_{max} (maximum canopy conductance), and E_R (extinction coefficient of radiation). The first four parameters primarily control the soil water-holding capacity, hydraulic conductivity, and matric potential gradient, which are critical for infiltration and runoff. The last three parameters regulate the interception of precipitation and energy fluxes through the canopy, thereby influencing transpiration and evaporation processes closely linked to SSM dynamics.

For wetland sites, the model employs a two-zone structure defined by the water table depth to dynamically estimate the SSM distribution. At these sites, while I_{RMAX}, g_{max}, and E_R were still calibrated to represent transpiration and evaporation processes, the four soil parameters calibrated for uplands were not adjusted. Instead, we calibrated Q_{DRMAX} (maximum drainage rate), which governs the calculation of saturated flow drainage below the maximum water table depth, thereby improving the representation of drainage processes and SSM dynamics under wetland conditions.

All parameter calibration in this study was conducted using the PEST software (Model-Independent Parameter Estimation and Uncertainty Analysis, v17.2 for Linux), which optimizes parameters by minimizing the difference between simulated and observed values (Doherty, 2004). To ensure consistency across models, the observational data used for PB model calibration were identical to those used for training the DL and PGDL models. The calibrated PB outputs were subsequently incorporated as physically guided inputs into the PGDL model.

2.4.2. For DL and PGDL Model

2.4.2.1. Multi-Site Training With ISMN SSM Observations

DL and PGDL models were implemented in PyTorch. We applied Bayesian optimization to tune four key hyperparameters: learning rate, sequence length, number of hidden units in the LSTM, and number of hidden units in the FCNN. Other hyperparameters were fixed, including a batch size of 64, a single LSTM layer, and a dropout rate of 0.5. During training, the Adam optimizer (Kingma & Ba, 2014) was used for gradient descent,

Table 1
Number of Grid Cells and ISMN Sites in Each Cluster and the Total Across All Clusters

Cluster	Number of grid cell	Number of ISMN sites
1	3,937	21
2	5,756	26
3	10,963	31
4	6,653	41
5	6,069	14
6	4,398	15
7	1,470	13
8	2,430	11
9	4,900	268
10	5,880	31
11	3,656	13
Total	56,112	484

with mean squared error (MSE) as the loss function. Early stopping was also employed to prevent overfitting and reduce unnecessary training time.

The models were trained using SSM observations from ISMN sites as labels. Training was organized by cluster, and the LOSO strategy was adopted across sites. Within each cluster, every site was iteratively treated as the test set, while the remaining sites served as the training set, with both sets spanning the full temporal series at each site. For Cluster 9, which contained the largest number of ISMN sites (268 sites, see Table 1), we selected a representative subset of 40 sites for training and evaluation to control computational cost and maintain consistency in training size with other clusters. Site selection followed a structured clustering-based sampling procedure: we first extracted and standardized input features for each site, applied k-means clustering to group the sites into 40 clusters, and then selected the site closest to each cluster center as the representative. This ensured comprehensive coverage of the internal feature distribution within Cluster 9 and improved representativeness while minimizing training bias. This cross-site training and validation workflow presents a more challenging and realistic modeling scenario, as the model is required to learn temporal dynamics from multiple spatially distinct sites and apply this knowledge to predict the full time series at an unseen site within the same cluster.

For SSM predictions on the test set, we employed a single-step iterative forecasting strategy. After each prediction step, the model output was fed back into the input sequence for the next time step, replacing the earliest SSM data point to form a new input. This approach enables continuous forecasting without relying on future SSM observations, thereby avoiding information leakage while preserving essential temporal dependencies. By dynamically updating the input sequence with model predictions at each step, the model effectively tracks SSM evolution over time. Among all LOSO iterations, the model achieving the best performance at the test site was selected as the representative model for that cluster and subsequently applied to regional prediction.

2.4.2.2. Global Prediction Against SMAP SSM Observations

SMAP SSM observations were used as an independent benchmark to assess the global performance of the PB, DL, and PGDL models. During this stage, we applied the same iterative prediction strategy used in the site-level simulations, requiring the model to perform rolling predictions based on a valid input sequence at each time step. To initialize these predictions, the SMAP SSM data set was preprocessed to construct continuous valid input sequences for model prediction. Specifically, for each grid cell, we first identified a valid window at the beginning of the observation period from the original time series, prioritizing windows with fewer missing values. If no valid window was available, we constructed a representative sequence by spatially averaging the values of neighboring grid cells and extracted a valid window from it. This synthetic sequence was used only to fill the missing segment of the target grid cell to retain as much of the original observation as possible. For a small subset of windows with remaining limited missing values, we applied local interpolation to complete the sequence and introduced minor perturbations to avoid numerical homogeneity. The resulting preprocessed SSM observational data set thus provides continuous valid input windows for initialization purposes without altering the original SMAP measurements.

After completing the gap-filling of SMAP SSM initial sequences, we performed a post-calibration procedure to align model predictions to a consistent observational scale. To account for potential systematic scale differences between training labels and global simulations, a simple linear adjustment was applied to the outputs of both DL and PGDL models. This regression-based adjustment aligns the overall magnitude of model simulations with the SMAP reference without modifying the model structure or the SMAP data set. Specifically, for each cluster, we set a series of SMAP valid-ratio thresholds to select reference samples for post-calibration. These thresholds were chosen to balance sample quality and quantity, ensuring both sufficient observations for reliable calibration and broad spatial coverage. The threshold intervals were defined at moderate resolution to provide sufficient representative samples for fitting under varying levels of data completeness. The final result for each cluster was determined based on a comprehensive evaluation across multiple threshold levels to balance accuracy and

coverage. For the reference samples selected at each threshold, separate simple linear regressions were fitted for the DL and PGDL predictions against SMAP observations using the least squares method and then applied to all grid cells to correct systematic biases and produce the final regional simulations.

2.4.3. Aggregation of Evaluation Metrics

To comprehensively assess the predictive performance of the PB, DL, and PGDL models, we conducted model comparisons at both site and regional scales using RMSE and r as the evaluation metrics.

For site-level assessments based on ISMN SSM observations, metrics were calculated under the LOSO strategy, with each site treated as the test site in turn and results averaged within each cluster. For regional-scale assessments based on SMAP SSM observations, metrics were calculated by comparing model predictions with observations at each grid cell. To ensure the stability and reliability of the correlation analysis, only grid cells with a SMAP valid-ratio of at least 0.5 and a standard deviation of no less than 0.02 were included in the r statistics. RMSE was calculated for all grid cells with valid SMAP observations.

To obtain across-cluster mean values for the performance metrics, we computed a weighted average ($\overline{v}_w$) based on the number of samples within each cluster:

$$\overline{v}_w = \frac{\sum_{i=1}^{k} n_i \cdot v_i}{\sum_{i=1}^{k} n_i} \tag{7}$$

where v_i is the mean metric value for cluster i, n_i is the number of samples in cluster i, and k is the total number of clusters. For site-level analysis, n_i denotes the number of ISMN sites used for LOSO strategy in cluster i; for regional analysis, n_i denotes the number of grid cells meeting the evaluation criteria in cluster i. In this way, $\overline{v}_w$ represents the overall model performance at both site and regional scales.

3. Results

3.1. Spatial Distribution and Climatic Characteristics of Clusters

This study conducted a clustering analysis of global land regions using the same features as model inputs and determined the optimal clustering configuration. Figure 2 presents the spatial distribution of clusters and ISMN sites, with panel (a) illustrating the optimal cluster configuration and panel (b) displaying the locations of the selected ISMN sites. Gray regions indicate areas outside the valid SMAP SSM coverage that were excluded from the analysis. Table 1 summarizes the number of grid cells and ISMN sites in each cluster. In total, 56,112 land grid cells were grouped into 11 clusters, with 484 ISMN sites distributed among them.

To interpret the clustering results from a climatic perspective, we incorporated the widely used Köppen–Geiger climate classification, a widely used global climate zoning scheme that characterizes regional climate conditions. We utilized the latest data set based on 1991–2020 historical maps at 0.5° resolution (Beck et al., 2023). For this macro-scale analysis, the original climate types were aggregated into five major categories following the Köppen–Geiger system: tropical, arid, temperate, cold, and polar. The global distribution of these five climate types is shown in Figure S3 of Supporting Information S1. These categories represent the main global climate zones and provide a clear basis for interpreting the clustering results. We calculated the proportion of each climate type within each cluster and summarized the results using bar plots (Figure 3).

The analysis reveals that Clusters 1, 10, and 11 are predominantly characterized by tropical climates. Among them, Cluster 11 exhibits the most pronounced tropical dominance, primarily associated with low-latitude humid regions. In contrast, Clusters 1 and 10 show increasing contributions from temperate or arid climates, indicating transitional climatic characteristics toward subtropical or tropical–arid regimes.

Clusters 2, 5, and 9 are primarily characterized by arid climates, with varying contributions from cold or tropical components. Among them, Cluster 2 contains the most concentrated arid component and is mainly distributed across typical desert regions. In contrast, Clusters 5 and 9 reflect more mixed arid–cold structures associated with subtropical and high-altitude environments.

10.1029/2025JH001121

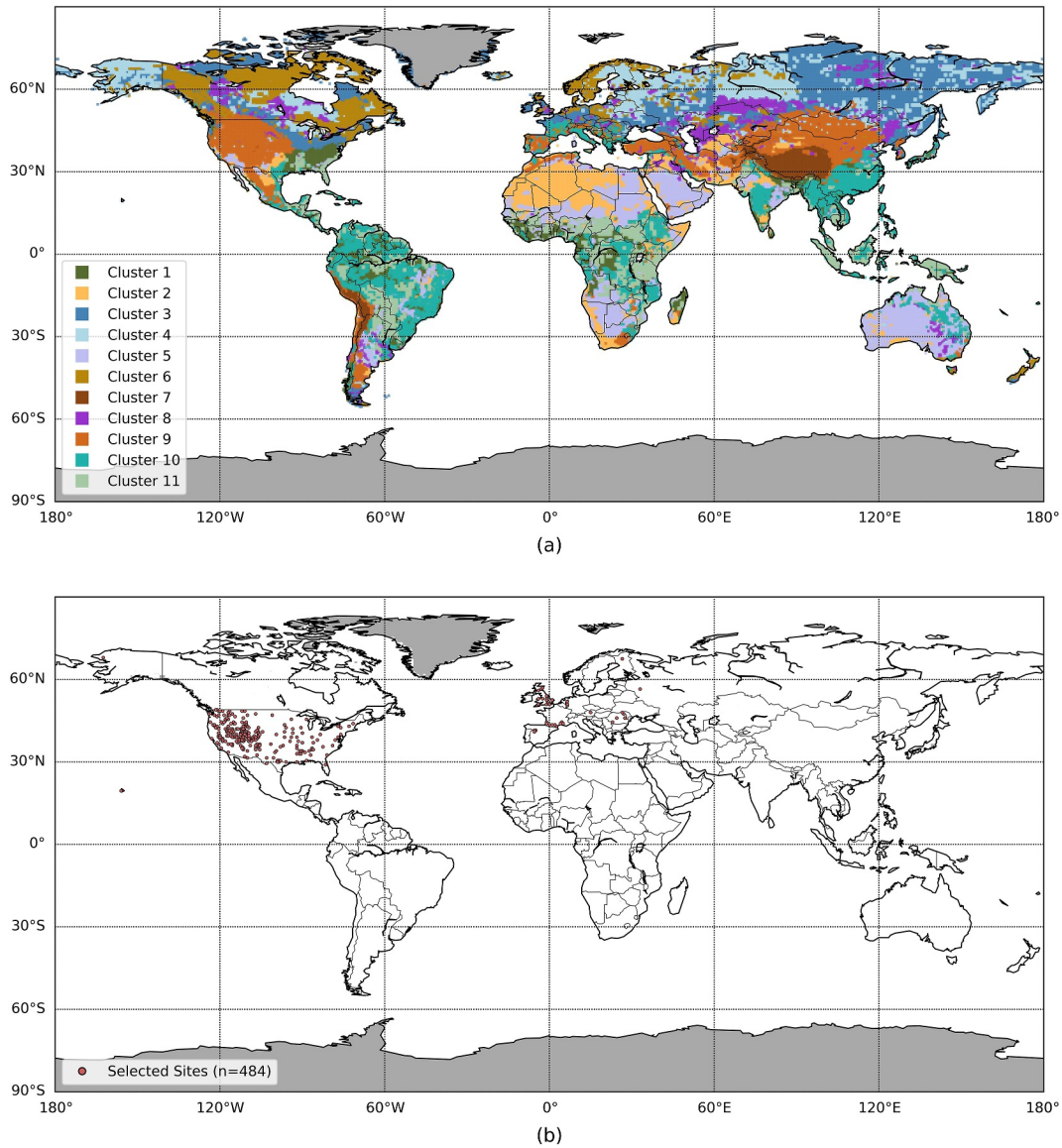

Figure 2. Spatial distribution of clusters and ISMN sites derived from the optimal k-means clustering at the global scale. Panel (a) depicts the cluster distribution, with each cluster represented by a unique color. Panel (b) presents the locations of the selected ISMN sites, shown as red dots. In both panels, gray regions indicate areas outside the valid SMAP SSM coverage that were excluded from the analysis.

Clusters 3, 4, 6, and 8 exhibit cold-climate dominance and are mainly distributed across high-latitude regions. Among them, Cluster 4 has the highest proportion of cold climates. Clusters 3 and 6 also exhibit strong cold-climate dominance but with notable polar contributions, indicating transitional characteristics toward polar conditions. In contrast, Cluster 8 shows a reduced cold component accompanied by increased arid influence, reflecting a mixed cold–arid climatic structure.

Finally, Cluster 7 exhibits the highest proportion of polar climate among all clusters and is primarily associated with high-altitude cold environments, where polar climatic characteristics occur despite relatively low latitudes.

Overall, incorporating the Köppen–Geiger classification clarifies the climatic composition and geographic distribution of each cluster, supporting the physical relevance of the clustering for subsequent regional modeling.

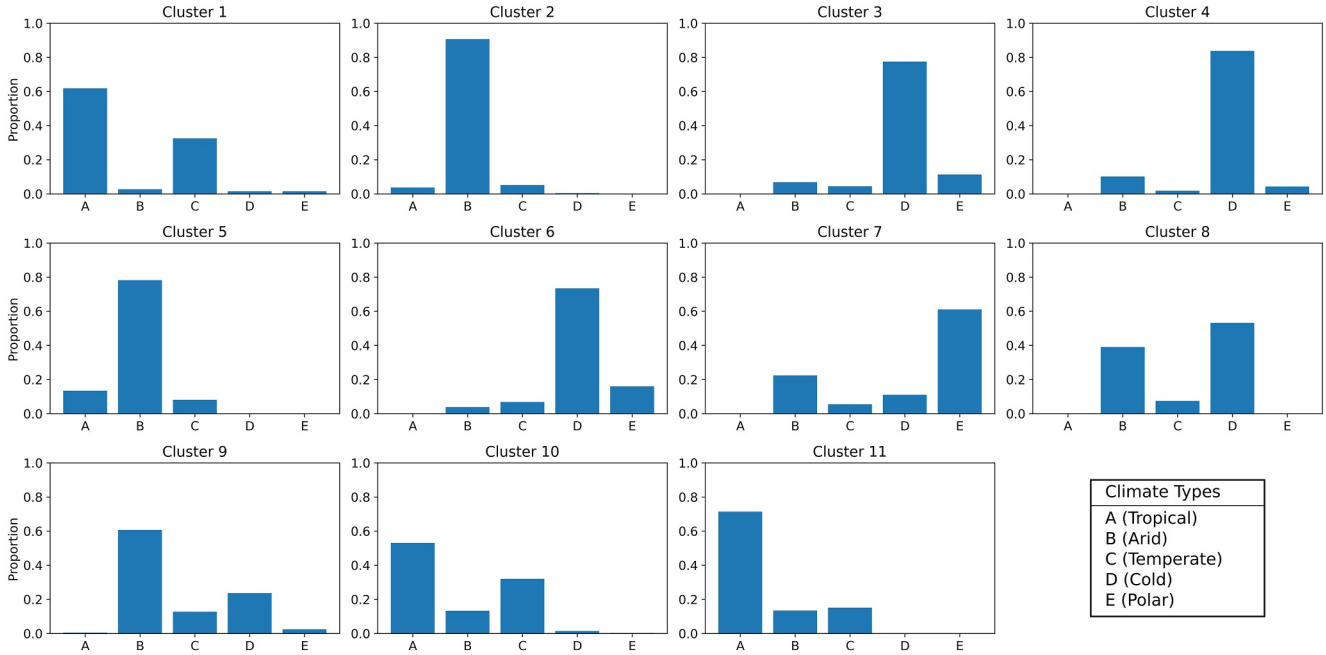

Figure 3. Proportional distribution of the five aggregated Köppen–Geiger climate types (Tropical, Arid, Temperate, Cold, and Polar) within each cluster.

3.2. Model Performance on ISMN Sites

The predictive performance of the PB, DL, and PGDL models was evaluated based on ISMN observations under the LOSO cross-validation strategy. Table 2 presents the average performance of the three models across all clusters, with overall mean values calculated using the weighted averaging scheme described in Equation 7.

Across clusters, the PGDL model achieved an average RMSE = 0.078 and r = 0.61, significantly outperforming both the PB model (RMSE = 0.117; r = 0.45) and the DL model (RMSE = 0.087; r = 0.55). At the cluster level, PGDL generally outperformed the other two models in predictive accuracy and temporal correlation. The DL model outperforms the PB model in most clusters, highlighting the potential of data-driven approaches for cross-site SSM modeling. PGDL further improved upon DL by incorporating physical process guidance, thereby enhancing generalization capability. Notably, even in clusters where the PB model does not perform well (e.g., Cluster 4 and Cluster 11), PGDL still outperforms DL, indicating that integrating physical processes can strengthen deep learning performance even when process-based model accuracy is limited.

Table 2
Averaged LOSO Performance Metrics of the PB, DL, and PGDL Models Across ISMN Test Sites

	PB		DL		PGDL	
Cluster	RMSE	r	RMSE	r	RMSE	r
1	0.094	0.52	0.082	0.60	0.070	0.68
2	0.096	0.65	0.077	0.63	0.059	0.68
3	0.106	0.46	0.105	0.47	0.093	0.56
4	0.198	0.08	0.081	0.59	0.079	0.66
5	0.070	0.48	0.080	0.50	0.073	0.60
6	0.092	0.40	0.087	0.50	0.077	0.49
7	0.097	0.54	0.086	0.46	0.080	0.55
8	0.091	0.50	0.082	0.52	0.072	0.58
9	0.109	0.57	0.091	0.53	0.085	0.55
10	0.106	0.62	0.079	0.61	0.070	0.68
11	0.135	0.13	0.114	0.44	0.103	0.58
Mean	0.117	0.45	0.087	0.55	0.078	0.61

Note. For each cluster, RMSE ($m^3\ m^{-3}$) and r were averaged over all test sites, and the last row summarizes the weighted mean values across all clusters.

Following the cross-site evaluation, the best-performing DL and PGDL models within each cluster were selected based on test results as representative models for subsequent regional-scale simulations. PB regional predictions were not selected based on ISMN site-level test performance but were generated independently from its underlying physical mechanism. Figure 4 shows the representative model simulations at the corresponding test site within each cluster.

As shown in Figure 4, the representative DL and PGDL models exhibit strong fitting performance at the test sites, confirming their reliability for regional simulations. Across all clusters, the PGDL model (average RMSE = 0.044, average r = 0.77) consistently outperforms DL (average RMSE = 0.054, average r = 0.66). This underscores the superior temporal fidelity of PGDL and suggests a higher upper limit for cross-site SSM simulation. While the DL model captures general SSM trends in most clusters, it shows limitations in

10.1029/2025JH001121

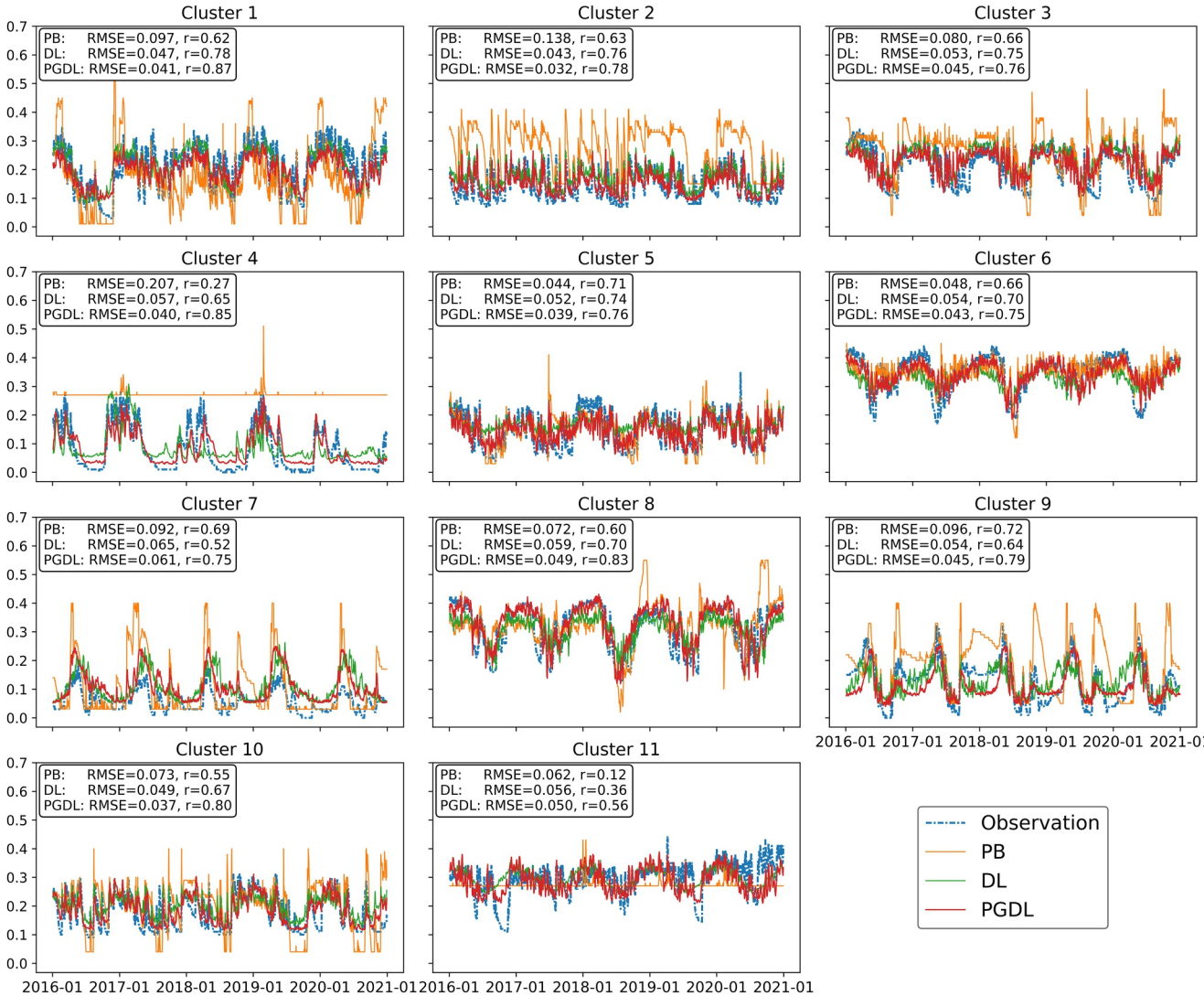

Figure 4. Comparison of daily SSM observations and model predictions (PB, DL, and PGDL) on the test set for representative sites within each cluster. The *y*-axis shows SSM (m^3 m^{-3}), and the *x*-axis shows time. The black text box in each panel reports RMSE and *r* for each model to indicate predictive performance.

reproducing extremes and seasonal fluctuations. In contrast, the PGDL model demonstrates greater sensitivity and expressive capacity in capturing both SSM magnitudes and seasonal dynamics, resulting in more consistent SSM time series fitting across clusters.

In summary, the cross-site performance metrics and predictions at representative test sites demonstrate that the PGDL model improves the learning and generalization of SSM dynamics through physics-guided design, supporting its application for regional simulation.

3.3. Model Evaluation Against SMAP SSM

The generalization performance of the three models was evaluated by comparing their predictions against SMAP SSM observations. Figure 5 shows the spatial distributions of RMSE and *r* for the three models, along with their corresponding latitudinal mean profiles. Table 3 summarizes the performance metrics for each cluster and the across-cluster means derived using Equation 7.

As shown in Figure 5 and Table 3, the PGDL model (RMSE = 0.081; $r = 0.55$) outperforms both PB (RMSE = 0.167; $r = 0.43$) and DL (RMSE = 0.085; $r = 0.40$) models, demonstrating its stronger

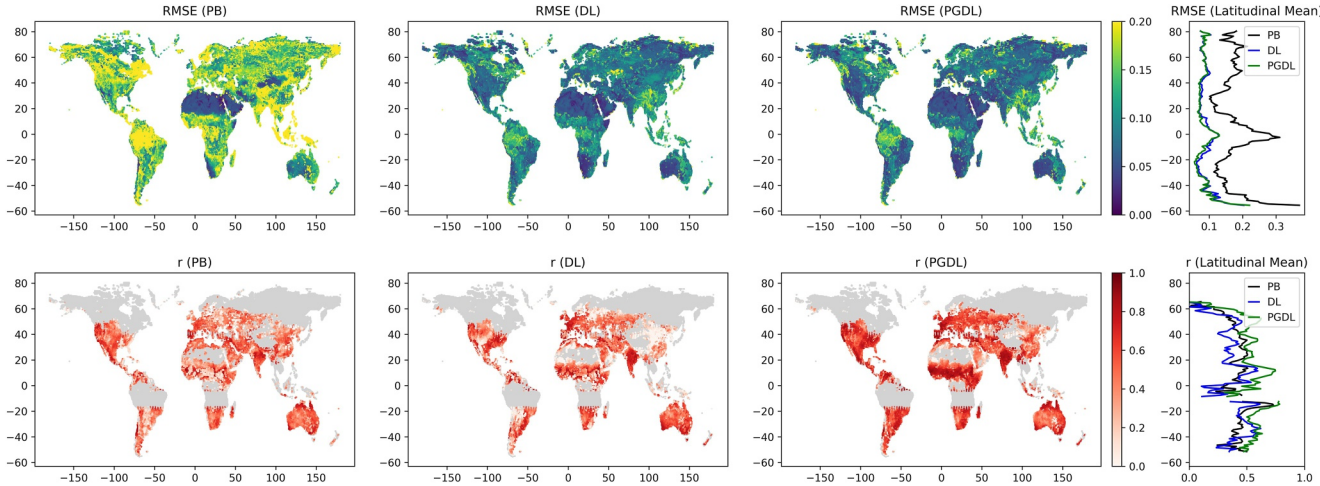

Figure 5. Spatial distributions and latitudinal means of model performance compared to SMAP SSM over the global scale during 2016–2020 at daily resolution. The top row shows the RMSE for the PB, DL, and PGDL models, respectively, while the bottom row presents the corresponding *r*. The fourth column in each row displays the latitudinal mean values for RMSE ($m^3\ m^{-3}$) and *r*. For *r*, only grid cells with a SMAP valid-ratio of at least 0.5 and a standard deviation no less than 0.02 were included to ensure the stability and reliability of the correlation calculation, and only latitudes with at least 10 valid grid cells were used in the latitudinal mean calculation. Gray areas in the *r* maps denote grid cells that did not meet the inclusion criteria and were excluded from the analysis.

generalization capability while balancing numerical accuracy and temporal consistency. The comparison between PB and DL reveals complementary behavior: PB achieves a slightly higher *r* than DL (0.43 vs. 0.40) but has a substantially larger RMSE (0.167 vs. 0.085), reflecting fundamental differences in model mechanisms. The PB model, governed by physical processes, better preserves SSM temporal dynamics and attains higher *r*, but its sensitivity to input errors and reliance on uniform parameterization result in larger magnitude deviations and higher RMSE. The DL model, by contrast, benefits from data-driven learning to achieve better magnitude accuracy, but the absence of physical constraints increases its sensitivity to noise and reduces temporal consistency. By integrating physics-guided representations, the PGDL model achieves both improved temporal fidelity and magnitude accuracy, thereby achieving the best overall performance.

Table 3
Regional Performance of the PB, DL, and PGDL Models Across All Clusters Compared to SMAP SSM

	PB		DL		PGDL	
Cluster	RMSE	*r*	RMSE	*r*	RMSE	*r*
1	0.202	0.50	0.102	0.51	0.090	0.62
2	0.083	0.50	0.052	0.45	0.051	0.54
3	0.175	0.44	0.082	0.43	0.081	0.54
4	0.181	0.11	0.094	0.18	0.089	0.53
5	0.114	0.48	0.061	0.56	0.061	0.55
6	0.206	0.43	0.093	0.37	0.094	0.45
7	0.193	0.50	0.097	0.05	0.084	0.54
8	0.150	0.31	0.087	0.38	0.090	0.35
9	0.155	0.48	0.083	0.20	0.075	0.55
10	0.192	0.51	0.101	0.51	0.096	0.61
11	0.228	0.18	0.109	0.18	0.098	0.63
Mean	0.167	0.43	0.085	0.40	0.081	0.55

Note. RMSE is reported in $m^3\ m^{-3}$. To ensure the stability and reliability of *r*, only grid cells with a SMAP valid data ratio of at least 0.5 and a standard deviation no less than 0.02 were included in *r* statistics. The last row reports the overall performance weighted by the number of eligible grid cells in each cluster.

The latitudinal mean profiles in Figure 5 show that the PGDL model outperforms the other models across all latitude bands, with the largest differences in the mid- and low-latitude regions. In terms of RMSE, the PB model yields higher values than both DL and PGDL across all latitudes, whereas for *r*, PB generally exhibits a more stable trend than DL. Across all three models, SSM simulations show larger deviations near the equator and in high-latitude regions, while overall performance is better in the mid-latitudes. This spatial pattern highlights systematic differences in model performance across latitude bands, with mid-latitude regions exhibiting more stable and consistent prediction accuracy.

To further evaluate model performance across climate regimes, we aggregated the comparisons into four major climate zones (Tropical, Arid, Cold, and Polar). Figure 6 shows that PGDL achieves the best overall performance across all zones, with particularly notable advantages in tropical and polar regions. Across the four climate types, model performance is generally highest in arid regions, followed by tropical regions, while cold and polar regions exhibit poorer results. Overall, PGDL consistently shows smaller errors and higher correlations than the PB and DL models, indicating a more consistent performance under diverse climatic conditions. For the other two models, although DL generally outperforms PB in terms of RMSE, its *r* in arid and polar zones are markedly lower, reflecting limited adaptability to these climate regimes.

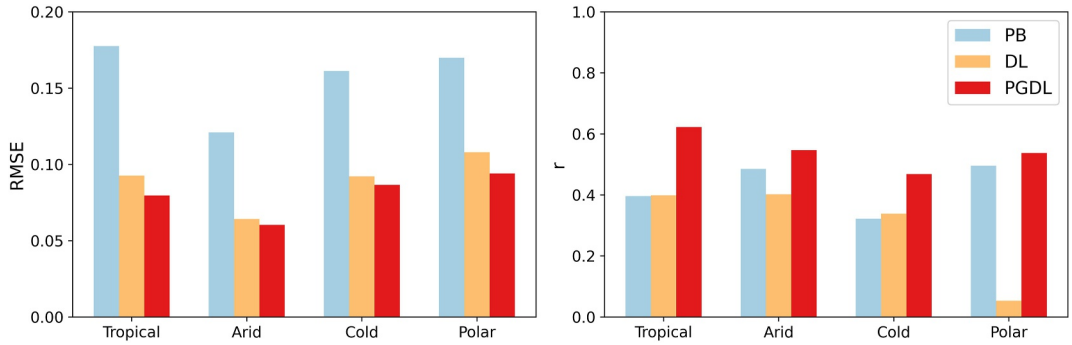

Figure 6. Performance of the PB, DL, and PGDL models across four major Köppen–Geiger climate types (Tropical, Arid, Cold, and Polar) at the global scale. The left panel shows RMSE and the right panel shows *r*, both calculated between model simulations and SMAP SSM. For each climate type, performance metrics were averaged across associated clusters.

In summary, the PGDL model outperforms PB and DL models across spatial scales and climate zones. These results suggest that the physics-guided design provides a more reliable and transferable deep learning framework for regional SSM modeling.

4. Discussions

4.1. Assessment of the Clustering Strategy

4.1.1. Effectiveness of Clustering for SSM Prediction

In this study, we applied a clustering strategy that integrates static and dynamic features to partition the global domain into subregions with relatively homogeneous environmental characteristics and developed independent models within each cluster. To further evaluate the necessity of this strategy for global SSM prediction, we compared the predictive performance of cluster-specific modeling with cross-cluster mixed modeling to test whether the proposed clustering approach improves SSM prediction. This comparison was conducted exclusively for the PGDL model, which demonstrated superior overall performance relative to DL, thereby providing a consistent framework for evaluating the clustering strategy.

The comparison was conducted for three major climate types: tropical, arid, and cold. For each climate type, a reference cluster was selected by jointly considering both the high proportion of the target climate type and the availability of sufficient ISMN sites to support the representative and robust test. Accordingly, clusters 1, 2, and 4 were selected as the reference clusters for the tropical, arid, and cold climates, respectively. Using each reference cluster, we fixed the test site and constructed two models: (a) independent modeling (IM), trained solely on ISMN sites within the reference cluster (Tables 2 and 3); and (b) mixed-sample modeling (MM), where the training set was built by randomly selecting equal numbers of ISMN sites from each cluster within the same climate type and combining them to match the size of the reference cluster training set. For example, cluster 4, the reference cluster in the cold climate type, contained 41 ISMN sites. In MM, 10 sites were randomly selected from each cold-climate cluster (clusters 3, 4, 6, and 8) to form a 40-site training set, trained and tested on the same test site as in IM, and then applied for regional prediction within cluster 4. The configurations for the ISMN-based training and SMAP-based regional prediction were kept identical between the two models to ensure that performance differences arose solely from the training data sources.

The overall performance of the two modeling strategies for each climate type is summarized in Table 4. Site-level metrics represent the mean LOSO-based predictions across all ISMN sites within each cluster, and regional metrics represent the averages over the corresponding clusters relative to SMAP SSM. The last row reports the overall performance across the three climate types.

The results show that, for site-level metrics, the IM approach (RMSE = 0.070; $r = 0.67$) consistently outperformed the MM approach (RMSE = 0.107; $r = 0.57$) across all climate types. This indicates that restricting training sites to a single cluster with more homogeneous environmental conditions enables the model to better capture feature-response relationships, whereas mixed training disrupts this consistency and reduces predictive accuracy. For regional metrics, the IM approach (RMSE = 0.077; $r = 0.56$) likewise outperformed the MM

Table 4
Predictive Performance of Independent Modeling (IM) and Mixed-Sample Modeling (MM) for the PGDL Model Across Tropical, Arid, and Cold Climate Types, Based on Site-Level Predictions Against ISMN SSM and Regional Predictions Relative to SMAP SSM

	Independent modeling (IM)				Mixed-sample modeling (MM)			
	Site level		Regional scale		Site level		Regional scale	
Climate	RMSE	r	RMSE	r	RMSE	r	RMSE	r
Tropical	0.070	0.68	0.090	0.62	0.111	0.65	0.100	0.46
Arid	0.059	0.68	0.051	0.54	0.096	0.56	0.053	0.43
Cold	0.081	0.64	0.089	0.53	0.115	0.49	0.086	0.43
Mean	0.070	0.67	0.077	0.56	0.107	0.57	0.080	0.44

Note. RMSE is reported in $m^3\ m^{-3}$. The last row reports the mean values across the three climate types.

approach (RMSE = 0.080; r = 0.44). These findings demonstrate that clustering benefits both site-level predictions and regional generalization.

Furthermore, this comparison was designed with mixed training conducted within the same climate type, representing a more stringent test. Mixing across different climate types inherently introduces larger environmental differences, where a decline in predictive performance would be expected. However, even under a consistent climatic background, mixed training still performed noticeably worse than cluster-specific independent modeling, underscoring the advantage of clustering. By leveraging multi-source static and dynamic features, this approach constructs more homogeneous training samples. Overall, this comparison confirms the necessity of the clustering strategy by enabling more representative and consistent training structures for global SSM prediction.

It should be noted that the clustering configuration adopted in this study represents a reasonable regionalization under the given research objectives and data conditions. It was selected based on a set of evaluation metrics and training constraints and used to define regionalization and support subsequent cluster-specific model training. The clustering results and associated performance could be influenced by key design choices, including the number of clusters (k), the feature set used for clustering, and the temporal aggregation of dynamic variables. For example, the choice of k reflects a trade-off between environmental homogeneity and sample size; aligning clustering features with model inputs ensures consistency between regionalization and modeling information; and using multi-year means is intended to represent long-term climatic background conditions for regional-scale modeling. Different settings may therefore lead to different regional partitions that are more suitable for different research objectives. Accordingly, the results of this study should be viewed as evidence for the effectiveness of regionalized training based on environmental similarity for the current research problem.

4.1.2. Influence of Observation Density on Model Performance

ISMN sites are unevenly distributed globally, with a large proportion concentrated in the continental United States (CONUS) region. Given this distribution, we further examined the relationship between observation density and model predictive performance. Specifically, we divided the study domain into CONUS and non-CONUS regions, and the PGDL model performance was compared between these two regions. The results show that the RMSE values are similar (CONUS: 0.077; non-CONUS: 0.081), indicating comparable error levels across regions. The r values are also relatively similar (CONUS: 0.63; non-CONUS: 0.54), suggesting no evident degradation in model performance in observation-sparse regions. These results indicate that the model maintains relatively stable predictive performance even under limited observational conditions.

To further characterize the effect of observation density, we conducted a cluster-level analysis. Observation density was defined as the ratio of the number of ISMN sites to the number of grid cells within each cluster. Clusters were then ranked from low to high density, and PGDL model performance (RMSE and r, evaluated against SMAP SSM observations) was compared across clusters (Figure S4 in Supporting Information S1). The results show that there is no clear monotonic relationship between model performance and observation density. Although some variability exists among clusters, no consistent decline in performance is observed with decreasing observation density, indicating that variations in observation density do not lead to systematic changes in model performance.

Overall, these results suggest that the proposed cluster-based PGDL framework remains stable under uneven observation distributions and exhibits good generalization capability across regions with varying observation densities.

4.2. Physical Consistency of DL and PGDL Predictions

In this study, we examined not only the predictive accuracy of the PGDL model but also its ability to maintain stronger physical consistency than the purely data-driven DL model. By incorporating physics-guided design, the

Table 5
Comparison of Violation Days of the Physical Water Balance Mechanism in DL and PGDL Model Predictions at Site and Regional Scales

	Site level		Regional scale	
Cluster	DL	PGDL	DL	PGDL
1	48	39	97	84
2	151	136	492	486
3	57	43	41	38
4	101	75	31	26
5	157	146	418	413
6	53	51	31	27
7	44	38	83	73
8	69	68	130	125
9	105	97	151	135
10	109	89	120	100
11	72	48	157	124
Mean	88	76	159	148

Note. For each cluster, site-level metrics represent the mean number of violation days across all LOSO test sites, and regional-scale metrics represent the mean number of violation days across all grid cells with valid evaluation periods. The last row reports the across-cluster means.

PGDL model is expected to exhibit more physically consistent behavior aligned with hydrological principles. According to the water balance equation (Equation 5), only I_F increases SSM, whereas E_V, E_S, and D_R decrease it. Thus, during periods with zero I_F and nonzero E_V and E_S, SSM should not increase. Based on this diagnostic criterion, we evaluated DL and PGDL predictions by comparing the number of days on which each model predicted an increase in SSM during such conditions (i.e., violations of the water balance principle). This analysis tested whether incorporating physics-guided design enables the PGDL model to produce predictions that are more consistent with the water balance principle.

We performed this evaluation for both site-level and regional-scale predictions. In both cases, evaluation periods were first identified based on the following criteria: I_F remained zero, at least one of E_V or E_S was nonzero, and the period lasted at least 10 days. This threshold of length excluded short-term random fluctuations while retaining a sufficient sample size to capture representative processes. For site-level predictions, if ISMN SSM did not increase during a qualified period, days with predicted SSM increases were counted as violations of the water balance principles. For regional-scale predictions, only periods with no more than 50% missing SMAP SSM and with no more than 50% of days showing SMAP SSM increases were included. These constraints effectively removed cases with high observational uncertainty, thereby avoiding bias in the violation statistics. Finally, for each cluster, violation days were averaged across all LOSO-based test sites at the site level and across all grid cells with valid evaluation periods at the regional scale, with results summarized in Table 5.

Results in Table 5 show that the PGDL model produced fewer violation days than the DL model in almost all clusters at both site and regional scales. This indicates that under the defined evaluation conditions, PGDL is less prone to generating anomalous SSM increases that violate water balance expectations. It should be noted that this diagnostic focuses on the directionality of SSM responses under specific water balance conditions and does not explicitly assess the magnitude or timing of SSM variations. Therefore, the results provide a comparative indication of physical consistency rather than a comprehensive evaluation of all aspects of SSM dynamics. In addition, fewer site-level violations suggest that the model exhibits stronger physical consistency when driven directly by in situ observations. This assessment was conducted under strict screening criteria, which reduced the proportion of extreme violation cases and thereby narrowed the gap between the two models. Thus, the improvements observed for PGDL were achieved under stringent quality controls, highlighting its robustness.

This improvement in physical consistency can be attributed to the integration of physically informed components into the PGDL model architecture, enabling better capture of relationships among controlling variables of SSM dynamics. The LSTM branch identifies temporal dependencies and evolving patterns, while the FCNN branch leverages features linked to physical processes to supplement nonlinear response representation. As a result, the model exhibits SSM responses that are consistent with water balance principles without explicitly enforcing physical constraints during training. Such alignment with physical laws enhances the model stability across diverse climatic and geographic conditions. Compared to the purely DL model, PGDL predictions are more interpretable in terms of physical processes. Adhering to the fundamental mechanism of water balance helps reduce the risk of producing unrealistic predictions under unseen conditions, thereby improving the model's applicability and reliability in global SSM prediction.

4.3. Impact of SMAP SSM Uncertainty on Model Evaluation

In this study, global SSM simulations were evaluated using SMAP SSM as the observational reference. Although SMAP products generally meet their design accuracy, retrievals over densely vegetated regions and areas with water bodies, snow, or frozen soils can exhibit substantial uncertainties (Entekhabi et al., 2014; McColl et al., 2017; Wrona et al., 2017). To reduce their impact on model evaluation, we applied a quality mask that excluded grid cells with vegetation water content greater than 5 kg m^{-2}, surface temperature below 0°C, or water

Table 6
Predictive Performance (RMSE in $m^3\ m^{-3}$ and r) of the PB, DL, and PGDL Models Against SMAP SSM Across Clusters Excluding Regions With High Retrieval Uncertainty

	PB		DL		PGDL	
Cluster	RMSE	*r*	RMSE	*r*	RMSE	*r*
1	0.169	0.57	0.091	0.60	0.066	0.71
2	0.079	0.51	0.048	0.45	0.047	0.55
3	0.170	0.45	0.073	0.48	0.069	0.61
4	0.214	0.11	0.078	0.26	0.070	0.52
5	0.110	0.49	0.056	0.57	0.055	0.55
6	0.175	0.45	0.090	0.49	0.091	0.56
7	0.167	0.47	0.098	0.07	0.084	0.54
8	0.148	0.32	0.086	0.39	0.085	0.36
9	0.149	0.47	0.077	0.21	0.066	0.55
10	0.154	0.56	0.079	0.58	0.069	0.67
11	0.159	0.19	0.099	0.19	0.079	0.69
Mean	0.140	0.45	0.071	0.43	0.064	0.58

Note. The last row reports the mean values across clusters.

body fraction exceeding 5%. Based on the original results (Table 3), predictive performance metrics of the three models were recalculated after removing the masked regions (Table 6). The global distribution of the applied mask is shown in Figure S5 of Supporting Information S1.

As shown in Table 6, after removing these high-uncertainty regions, the PGDL model still achieved the best performance (masked RMSE = 0.064, masked r = 0.58), outperforming both PB (masked RMSE = 0.140, masked r = 0.45) and DL (masked RMSE = 0.071, masked r = 0.43), with consistent advantages across nearly all clusters. Compared with unmasked results (Tables 2 and 3), predictive performance improved for all three models, with average RMSE reductions of 20%, 16%, and 16% for PB, DL, and PGDL, respectively, along with modest increases in r. These findings indicate that uncertainties in SMAP observations can bias global-scale evaluations and lead to underestimation of model performance. Once these high-uncertainty regions were excluded, the results more reliably reflected model behavior under robust observational conditions. Among the three models, PGDL exhibited the most pronounced improvement, suggesting that its performance enhancements are not merely attributable to reduced observational noise but arise from its intrinsic design, which integrates physical mechanisms with data-driven learning. This highlights the superior generalization ability and physical consistency of PGDL in global SSM simulations.

5. Conclusion

This study employed a PGDL model that integrates physical mechanisms with deep learning for global SSM prediction. Using an optimized clustering strategy based on multi-source features, the globe was partitioned into subregions with relatively consistent environmental characteristics. Cluster-specific training and independent modeling were conducted with in situ observations and subsequently extended to global-scale predictions and evaluations using satellite observations.

Compared with the traditional PB model and the purely data-driven DL model, the PGDL model achieved higher accuracy and stability in SSM prediction at both site and regional scales. This advantage stems from its architecture, which combines the DL capability of capturing long-term temporal dependencies with physics-guided design informed by the PB model, thereby enabling a more comprehensive representation of SSM dynamics.

Further analyses reveal that the clustering strategy significantly enhances model generalization across diverse climatic and geographic regions, as constructing environmentally consistent training samples effectively improves predictive performance. In addition, the evaluation using water balance diagnostics shows that the PGDL model produces fewer violations, indicating physically consistent SSM behavior aligned with hydrological principles and improved interpretability. Moreover, after excluding regions with high uncertainties in SMAP SSM observations, all three models exhibited performance improvements, with PGDL showing the largest enhancement. This suggests that SMAP uncertainties can bias global evaluations, whereas excluding these regions yields results that more accurately reflect model behavior under reliable observational conditions.

In summary, the proposed framework provides several advantages for global soil moisture modeling compared with existing approaches. First, the physics-guided design enables the model to explicitly incorporate key water-balance processes, allowing the learning algorithm to utilize physical information while improving predictive performance and maintaining better physical consistency. Second, the clustering-based regionalization framework accounts for environmental heterogeneity at the global scale by training region-specific models within environmentally similar regions, thereby improving model adaptability and stability across diverse conditions. Third, the framework effectively leverages limited ground-based observations to generate spatially continuous soil moisture predictions globally, highlighting its scalability and potential for large-scale applications under sparse observational conditions.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Availability Statement

The codes and data for this study are available at the Purdue University Research Repository (Xi & Zhuang, 2025).

Acknowledgments
This study is supported through a project funded to Qianlai Zhuang by NASA (NNX17AK20G). We acknowledge the Rosen High-Performance Computing Center at Purdue for computing support.

References

Beck, H. E., McVicar, T. R., Vergopolan, N., Berg, A., Lutsko, N. J., Dufour, A., et al. (2023). High-resolution (1 km) Köppen-Geiger maps for 1901–2099 based on constrained CMIP6 projections. *Scientific Data*, *10*(1), 724. https://doi.org/10.1038/s41597-023-02549-6

Berg, A., & Sheffield, J. (2018). Climate change and drought: The soil moisture perspective. *Current Climate Change Reports*, *4*(2), 180–191. https://doi.org/10.1007/s40641-018-0095-0

Breen, K. H., James, S. C., White, J. D., Allen, P. M., & Arnold, J. G. (2020). A hybrid artificial neural network to estimate soil moisture using swat+ and SMAP data. *Machine Learning and Knowledge Extraction*, *2*(3), 16–306. https://doi.org/10.3390/make2030016

Clark, M. P., Schaefli, B., Schymanski, S. J., Samaniego, L., Luce, C. H., Jackson, B. M., et al. (2016). Improving the theoretical underpinnings of process-based hydrologic models. *Water Resources Research*, *52*(3), 2350–2365. https://doi.org/10.1002/2015wr017910

Dobriyal, P., Qureshi, A., Badola, R., & Hussain, S. A. (2012). A review of the methods available for estimating soil moisture and its implications for water resource management. *Journal of Hydrology*, *458*, 110–117. https://doi.org/10.1016/j.jhydrol.2012.06.021

Doherty, J. (2004). *PEST model-independent parameter estimation user manual* (Vol. 3338, p. 3349). Watermark Numerical Computing.

Dorigo, W. A., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Xaver, A., et al. (2011). The international soil moisture network: A data hosting facility for global in situ soil moisture measurements. *Hydrology and Earth System Sciences*, *15*(5), 1675–1698. https://doi.org/10.5194/hess-15-1675-2011

Dorigo, W. A., Xaver, A., Vreugdenhil, M., Gruber, A., Hegyiova, A., Sanchis-Dufau, A. D., et al. (2013). Global automated quality control of in situ soil moisture data from the international soil moisture network. *Vadose Zone Journal*, *12*(3), 1–21. https://doi.org/10.2136/vzj2012.0097

Dunkl, I., & Ließ, M. (2022). On the benefits of clustering approaches in digital soil mapping: An application example concerning soil texture regionalization. *Soil*, *8*(2), 541–558. https://doi.org/10.5194/soil-8-541-2022

Entekhabi, D., Njoku, E. G., O'neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., et al. (2010). The soil moisture active passive (SMAP) mission. *Proceedings of the IEEE*, *98*(5), 704–716. https://doi.org/10.1109/jproc.2010.2043918

Entekhabi, D., Yueh, S., O'Neill, P. E., Kellogg, K. H., Allen, A., Bindlish, R., & West, R. (2014). SMAP handbook–soil moisture active passive: Mapping soil moisture and freeze/thaw from space.

Green, J. K., Seneviratne, S. I., Berg, A. M., Findell, K. L., Hagemann, S., Lawrence, D. M., & Gentine, P. (2019). Large influence of soil moisture on long-term terrestrial carbon uptake. *Nature*, *565*(7740), 476–479. https://doi.org/10.1038/s41586-018-0848-x

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

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., & Rozum, I. (2023). ERA5 hourly data on single levels from 1959 to present—Copernicus climate change service (C3S) climate data store (CDS).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu, H., Liu, J., Zhang, X., & Fang, M. (2023). An effective and adaptable K-means algorithm for big data cluster analysis. *Pattern Recognition*, *139*, 109404. https://doi.org/10.1016/j.patcog.2023.109404

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., et al. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318–2331. https://doi.org/10.1109/tkde.2017.2720168

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv Preprint*. https://doi.org/10.48550/arXiv.1412.6980

Koster, R. D., & Suarez, M. J. (2001). Soil moisture memory in climate models. *Journal of Hydrometeorology*, *2*(6), 558–570. https://doi.org/10.1175/1525-7541(2001)002<0558:smmicm>2.0.co;2

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Lal, P., Shekhar, A., Gharun, M., & Das, N. N. (2023). Spatiotemporal evolution of global long-term patterns of soil moisture. *Science of the Total Environment*, *867*, 161470. https://doi.org/10.1016/j.scitotenv.2023.161470

Lange, S., Mengel, M., Treu, S., & Büchner, M. (2022). ISIMIP3a atmospheric climate input data (v1.0). *ISIMIP Repository*. https://doi.org/10.48364/ISIMIP.982724

Li, Q., Xiao, Q., Zhang, C., Zhu, J., Chen, X., Yan, Y., et al. (2024). Improving global soil moisture prediction through cluster-averaged sampling strategy. *Geoderma*, *449*, 116999. https://doi.org/10.1016/j.geoderma.2024.116999

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. *Nature Geoscience*, *10*(2), 100–104. https://doi.org/10.1038/ngeo2868

McQueen, J. B. (1967). Some methods of classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley symposium on math. Statistics and Probability* (pp. 281–297).

Méndez-Barroso, L. A., Vivoni, E. R., Watts, C. J., & Rodríguez, J. C. (2009). Seasonal and interannual relations between precipitation, surface soil moisture and vegetation dynamics in the North American monsoon region. *Journal of hydrology*, *377*(1–2), 59–70. https://doi.org/10.1016/j.jhydrol.2009.08.009

Moosavi, V., Zuravand, G., & Shamsi, S. R. F. (2024). Cluster-based local modeling (CBLM) paradigm meets deep learning: A novel approach to soil moisture estimation. *Journal of Hydrology*, *635*, 131161. https://doi.org/10.1016/j.jhydrol.2024.131161

O'Neill, P., Chan, S., Njoku, E., Jackson, T., Bindlish, R., & Chaubell, J. (2021). SMAP L3 radiometer global daily 36 km EASE-grid soil moisture, version 8 [Dataset]. *NASA National Snow and Ice Data Center Distributed Active Archive Center*. https://doi.org/10.5067/OMHVSRGFX38O

O, S., & Orth, R. (2021). Global soil moisture data derived through machine learning trained with in-situ measurements. *Scientific Data*, *8*(1), 1–14. https://doi.org/10.1038/s41597-021-00964-1

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat, F. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Rosenzweig, C., Tubiello, F. N., Goldberg, R., Mills, E., & Bloomfield, J. (2002). Increased crop damage in the US from excess precipitation under climate change. *Global Environmental Change*, *12*(3), 197–202. https://doi.org/10.1016/s0959-3780(02)00008-0

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sculley, D. (2010). Web-scale k-means clustering. In *Proceedings of the 19th International Conference on world wide web* (pp. 1177–1178).

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., et al. (2010). Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews*, *99*(3–4), 125–161. https://doi.org/10.1016/j.earscirev.2010.02.004

Sun, W., Zhou, S., Yu, B., Zhang, Y., Keenan, T., & Fu, B. (2025). Soil moisture-atmosphere interactions drive terrestrial carbon-water trade-offs. *Communications Earth & Environment*, *6*(1), 169. https://doi.org/10.1038/s43247-025-02145-z

Sun, Y., & Niu, J. (2019). Regionalization of daily soil moisture dynamics using wavelet-based multiscale entropy and principal component analysis. *Entropy*, *21*(6), 548. https://doi.org/10.3390/e21060548

Vereecken, H., Huisman, J. A., Pachepsky, Y., Montzka, C., Van Der Kruk, J., Bogena, H., et al. (2014). On the Spatio-temporal dynamics of soil moisture at the field scale. *Journal of Hydrology*, *516*, 76–96. https://doi.org/10.1016/j.jhydrol.2013.11.061

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2022). Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*, *55*(4), 1–37. https://doi.org/10.1145/3514228

Wood, E. F., Roundy, J. K., Troy, T. J., Van Beek, L. P. H., Bierkens, M. F., Blyth, E., et al. (2011). Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research*, *47*(5). https://doi.org/10.1029/2010wr010090

Wrona, E., Rowlandson, T. L., Nambiar, M., Berg, A. A., Colliander, A., & Marsh, P. (2017). Validation of the soil moisture active passive (SMAP) satellite soil moisture retrieval in an Arctic tundra environment. *Geophysical Research Letters*, *44*(9), 4152–4158. https://doi.org/10.1002/2017gl072946

Xi, X., & Zhuang, Q. (2025). Improving global surface soil moisture prediction through physics-guided deep learning and cluster-based regionalization [Dataset]. *Purdue University Research Repository*. https://doi.org/10.4231/PX6C-5V49

Xi, X., Zhuang, Q., & Liu, X. (2025). A hybrid physics-guided deep learning modeling framework for predicting surface soil moisture. *Journal of Geophysical Research: Machine Learning and Computation*, *2*(3), e2025JH000682. https://doi.org/10.1029/2025jh000682

Yao, P., Lu, H., Shi, J., Zhao, T., Yang, K., Cosh, M. H., et al. (2021). A long term global daily soil moisture dataset derived from AMSR-E and AMSR2 (2002–2019). *Scientific Data*, *8*(1), 143. https://doi.org/10.1038/s41597-021-00925-8

Zhang, T., Liang, Z., Zhou, J., Shao, Q., Sarukkalige, R., Lü, H., et al. (2025). Multi-layer grid-scale soil moisture estimation using spatiotemporal deep learning methods with physical constraints. *Journal of Hydrology*, *657*, 133086. https://doi.org/10.1016/j.jhydrol.2025.133086

Zhuang, Q., McGuire, A. D., Melillo, J. M., Clein, J. S., Dargaville, R. J., Kicklighter, D. W., et al. (2011). Carbon cycling in extratropical terrestrial ecosystems of the Northern Hemisphere during the 20th century: A modeling analysis of the influences of soil thermal dynamics. *Tellus B: Chemical and Physical Meteorology*, *55*(3), 751–776. https://doi.org/10.1034/j.1600-0889.2003.00060.x

Zhuang, Q., McGuire, A. D., O'neill, K. P., Harden, J. W., Romanovsky, V. E., & Yarie, J. (2002). Modeling soil thermal and carbon dynamics of a fire chronosequence in interior Alaska. *Journal of Geophysical Research*, *107*(D1), FFR-3.

Zhuang, Q., Melillo, J. M., Kicklighter, D. W., Prinn, R. G., McGuire, A. D., Steudler, P. A., et al. (2004). Methane fluxes between terrestrial ecosystems and the atmosphere at northern high latitudes during the past century: A retrospective analysis with a process-based biogeochemistry model. *Global Biogeochemical Cycles*, *18*(3). https://doi.org/10.1029/2004gb002239