JGR Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2025JH000639

Key Points:

- A multimodal network (MMNet) was developed to integrate remote sensing and weather data for soil moisture downscaling
- MMNet achieved RMSE ≈$0.04 m^3/m^3$ in a temporal generalization test, enabling temporal gap filling or extension of in situ measurements
- It captured soil moisture dynamics in spatial and cross-region generalization tests, supporting interpolation in data-scarce areas

Supporting Information:

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

Correspondence to:

Z. Zhang,
zzhang347@wisc.edu

Citation:

Xu, Y., Cai, S., Huang, J., Liu, J., Shang, J., Yang, Z., & Zhang, Z. (2025). A multimodal deep learning approach for soil moisture downscaling using remote sensing and weather data. *Journal of Geophysical Research: Machine Learning and Computation*, *2*, e2025JH000639. https://doi.org/10.1029/2025JH000639

Received 14 FEB 2025
Accepted 12 JUN 2025

Author Contributions:

Conceptualization: Yijia Xu, Jingyi Huang, Zhou Zhang
Data curation: Shuohao Cai
Funding acquisition: Jingyi Huang, Zhou Zhang
Investigation: Yijia Xu, Shuohao Cai
Methodology: Yijia Xu, Shuohao Cai
Supervision: Zhou Zhang
Visualization: Yijia Xu
Writing – original draft: Yijia Xu

A Multimodal Deep Learning Approach for Soil Moisture Downscaling Using Remote Sensing and Weather Data

Yijia Xu[1], Shuohao Cai[2], Jingyi Huang[2], Jiangui Liu[3], Jiali Shang[3], Zhengwei Yang[4], and Zhou Zhang[1]

[1]Department of Biological Systems Engineering, University of Wisconsin-Madison, Madison, WI, USA, [2]Department of Soil Science, University of Wisconsin-Madison, Madison, WI, USA, [3]Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, ON, Canada, [4]Research and Development Division, United States Department of Agriculture, National Agricultural Statistics Service, Washington, DC, USA

Abstract Understanding soil moisture (SM) dynamics is crucial for environmental and agricultural applications. While satellite-based SM products provide extensive coverage, their coarse spatial resolution often fails to capture local SM variability. This study presents a multimodal network (MMNet) that integrates remote sensing and weather data to downscale Soil Moisture Active Passive (SMAP) Level-4 surface SM. We evaluated the performance of MMNet by comparing it with in situ SM observations from the Soil Climate Analysis Network (SCAN) and the United States Climate Reference Network (USCRN) under three scenarios. The results showed that (a) MMNet trained with on-site data provided accurate SM estimates over time in withheld years; (b) MMNet demonstrated spatial transferability, capturing SM dynamics in regions with sparse or no in situ measurements; and (c) the integration of snapshot and time-series data was crucial for maintaining the model's accuracy and generalizability across diverse scenarios. The downscaled SM maps demonstrated its potential for producing high-resolution temporally and spatially continuous SM estimates, which could further support a broad range of environmental and agricultural applications.

Plain Language Summary Soil moisture (SM), or the water content in soil, is essential for understanding various environmental and agricultural issues, including crop growth, vegetation health, wildfire risks, and flooding potential. While NASA's SMAP satellite provides a broad view of SM, its low resolution (9–36 km) misses important local details, limiting its use for applications requiring finer spatial resolution, such as precision agriculture. To generate higher resolution SM estimates, we introduced a deep learning model called MMNet, which builds relationship between in situ SM data and coarse SMAP SM, along with other SM-related features like soil properties, land surface images, and recent weather conditions, generating SM estimates at 100 m resolution. We tested MMNet across diverse scenarios and analyzed how different data types, that is, snapshot and time-series data, contribute to SM estimation. MMNet outperformed existing methods and the original SMAP, effectively capturing SM dynamics and spatial variations in both temporal and spatial extrapolation tests. These results highlight the importance of integrating multiple data types for accurate SM estimation and demonstrate that MMNet offers a valuable alternative to current deep learning models.

1. Introduction

Surface soil moisture (SM) is a critical component of the Earth system and a significant source of atmospheric water (Schmidt et al., 2024). It plays a vital role in applications, including numeric weather prediction, water resource management, flood forecasting, and precision farming (Bauer et al., 2015; Komma et al., 2008; Martínez-Fernández et al., 2016). While in situ SM networks provide accurate point-scale SM measurements, they often lack sufficient spatial coverage over larger areas (Karthikeyan & Mishra, 2021). Satellite remote sensing enables large-scaling continuous SM monitoring, but products like the National Aeronautics and Space Administration (NASA) Soil Moisture Active Passive (SMAP) (Entekhabi et al., 2010) have coarse spatial resolution (>10 km), which is insufficient for field-level applications. Downscaling the coarse-resolution SM products is thus essential to obtain finer-resolution SM estimates (Peng et al., 2017; Xu et al., 2022).

Current downscaling approaches can be categorized into three types: data assimilation, empirical/semiphysical methods, and statistical machine learning (ML) techniques (Kolassa et al., 2018; Wei et al., 2019; Zhao et al., 2022). Data assimilation combines coarse satellite SM data with fine-resolution land surface models (LSMs) to improve spatiotemporal accuracy (Dong et al., 2019), but is constrained by high computational costs

Writing – review & editing:
Jingyi Huang, Jiangui Liu, Jiali Shang, Zhengwei Yang, Zhou Zhang

(Li et al., 2021). Empirical and semiphysical methods exploit relationships between SM and high-resolution land surface parameters (Djamai et al., 2015; Sandholt et al., 2002; Wang et al., 2016), but their reliance on handcrafted indices may under-represent complex SM variations. In contrast, ML models can automatically learn features from diverse data sources, and offer significant advantages for SM downscaling without relying on explicit physical modeling (Zhu et al., 2024). For instance, random forest (RF) and Extreme Gradient Boosting (XGBoost) have been successfully used to downscale SMAP data to 1 km across various regions (Karthikeyan & Mishra, 2021; D. Long et al., 2019). More recently, deep learning (DL) models have demonstrated superior performance due to their scalability and powerful learning capabilities (LeCun et al., 2015). Several DL approaches have been successfully applied for SM downscaling (M. Xu et al., 2022; Zhu et al., 2024). For example, Liu et al. (2022) proposed a multiscale DL approach to enhance the spatial resolution of 36 km SMAP Level-3 products to 9 km over the contiguous United States (CONUS), achieving better accuracy than previous 1 km satellite downscaling efforts. Additionally, Batchu et al. (2023) developed a DL model that integrates inputs of varying resolutions to produce high-resolution (320 m) and accurate global SM estimates.

Despite achieving reasonable accuracy, the aforementioned methods rely on static features or data from the same day the SM is estimated, without adequately capturing the influence of preceding conditions. However, meteorological factors such as precipitation and evapotranspiration from previous days play an important role in soil moisture dynamics (Han et al., 2023). Therefore, some studies have introduced temporal context into SM estimation through two main approaches. First, indices such as the Antecedent Precipitation Evaporation Index (APEI) (Han et al., 2023), derived from past meteorological variables, have been used to account for the influence of prior weather events on current soil moisture. Second, sequential models like recurrent neural networks (RNNs) have been employed to learn temporal dynamics from time-series data for SM estimation (Fang & Shen, 2020; Q. Li et al., 2021; Yu et al., 2021). However, index-based methods do not fully leverage the time-series structure, while sequential models may suffer from over-reliance on limited data sources. These limitations underscore the need for approaches that more effectively integrate diverse SM-related factors for improving SM estimation.

In this study, we developed a multimodal deep learning network (MMNet) for soil moisture estimation. MMNet employs a dual-input architecture to integrate time-invariant geophysical attributes, snapshot satellite observations, and recent weather time-series, enabling comprehensive modeling of factors influencing SM values and dynamics. To develop and evaluate the proposed model, we collected 502,267 daily average SM values from 280 sites across the CONUS between 2016 and 2022, with in situ data from the United States Department of Agriculture's Soil Climate Analysis Network (SCAN, Schaefer et al., 2007) and the National Oceanic and Atmospheric Administration's U.S. Climate Reference Network (USCRN, Bell et al., 2013) as ground truth. The model was assessed in three scenarios (on-site, off-site, and cross-region) to offer comprehensive insights into its spatiotemporal transferability and address the following research questions:

(RQ1) Temporally, are historical in situ SM measurements adequate for developing models that can fill gaps or extend periods of missing data?
(RQ2) Spatially, can the model reliably downscale SM in regions with sparse or nonexistent observational data?
(RQ3) How do different data modalities, that is, time-series and snapshot data, influence the model's transferability across various scenarios?

2. Materials

2.1. Study Area and In Situ SM

As shown in Figure 1, CONUS was selected as the study area, where in situ SM measurements at a 5-cm depth were collected from the International Soil Moisture Network (ISMN), covering 280 stations within the SCAN and USCRN networks between 2016 and 2022. These stations span a variety of landcover types, including barren, crop, developed, forest, grassland, pasture, shrub, and wetland. To ensure data quality, only SM observations with the ISMN flag labeled as "G" (good), "C02" (SM > 0.6 m^3/m^3), "C03" (SM > saturation point derived from Harmonized World Soil Database parameter values), or "C02,C03" (SM > 0.6 m^3/m^3 and SM > saturation point) were retained (Dorigo et al., 2013). As shown in Figure 1b, except for 2019, when elevated SM levels were observed due to the major flooding (NOAA & NIDIS, 2020), the annual average SM exhibited minimal

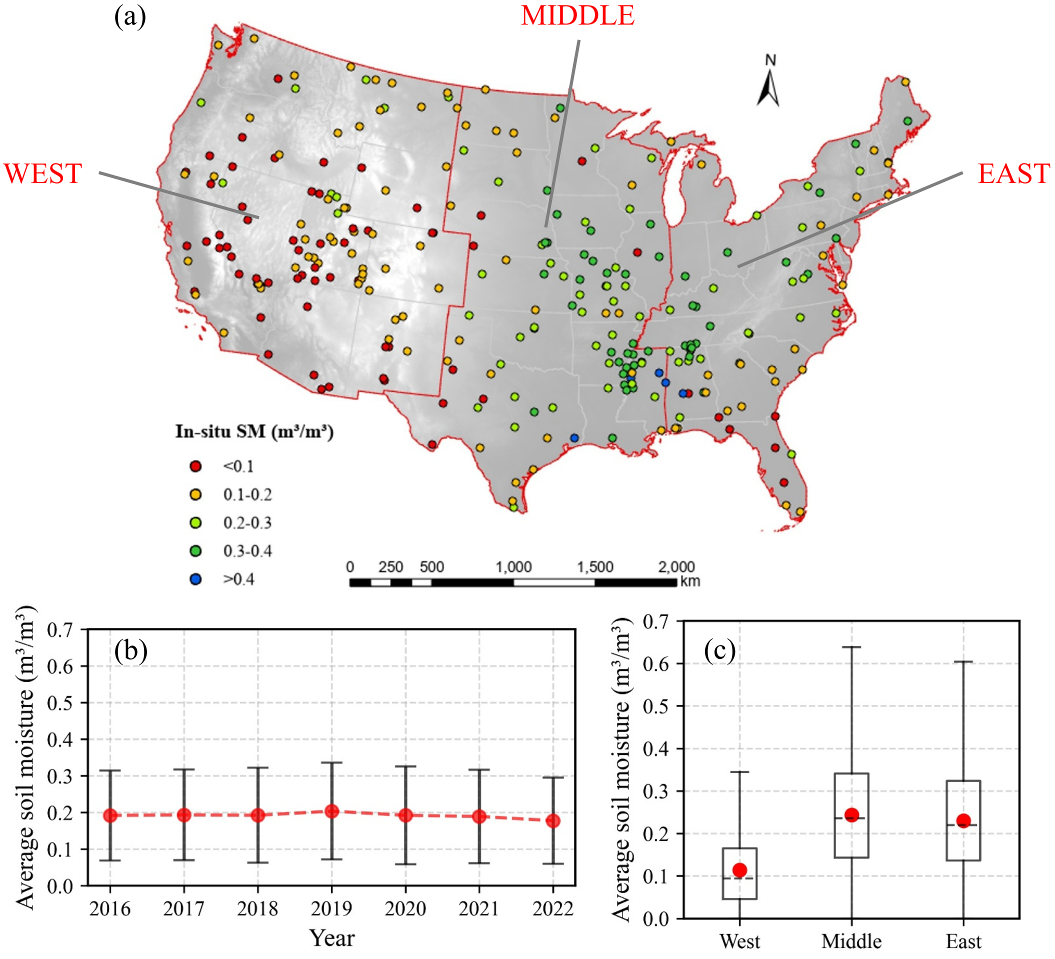

Figure 1. (a) Locations of the SCAN and USCRN stations across the CONUS, divided into three regions (as divided by thick red lines): west, middle, and east. The station colors represent the average SM at each station from 2016 to 2022. The background shows the digital elevation model (DEM). (b) Yearly SM time-series across all stations. Red dots represent the mean SM, and error bars indicate the standard deviation across stations within each year. (c) Boxplot showing the distribution of in situ SM for the west, middle, and east regions. Red dots denote the median values, with the boxes indicating the interquartile range and the whiskers representing the farthest points within 1.5x the interquartile range (IQR) from the box, excluding outliers.

interannual variation, with mean values slightly below 0.2 m^3/m^3. However, the high standard deviation indicates substantial variability across stations and throughout the year. To highlight the spatial differences in SM, we divided CONUS into three regions, west, middle, and east, based on terrain characteristics and SM distribution. The SM distribution for each region is shown in Figure 1c. This division broadly aligns with the geographical and ecological transitions shown in the United States Environmental Protection Agency (EPA) (U.S. EPA, 2015), while we followed a simplified framework to facilitate regional analysis. Specifically, the west region, characterized by high elevations, arid conditions, and complex topography (e.g., Rocky Mountains and Great Basin), typically exhibits low SM levels. The middle region, spanning much of the Great Plains, features relatively flat terrain and fertile soil, leading to generally higher SM levels, with some stations averaging over 0.4 m^3/m^3. The east region exhibits more variability in both terrain and climate. While much of the east exhibits moderate to high SM, the southeast (e.g., Florida and coastal plains) have low SM due to sandy and well-drained soil.

Table 1
Summary of Variables Contained in a Sample of the Proposed Model

Modality	Type (source)	Variables	Spatial resolution	Temporal resolution	Dim
Snapshot data	Coarse SM data (SMAP L4)	surface SM (0–5 cm)	9 km	3h	$\mathbb{R}^1$
	Calculated indices (Landsat 8 and Sentinel-1)	NDVI, EVI, NDWI, LSWI; CR, DPSVIm, and Pol	30 m/10 m	16 d/6 d	$\mathbb{R}^7$
	Terrain features (USGS DEM)	Elevation, Slope, and Aspect	30 m	Static	$\mathbb{R}^3$
	Landcover (NLCD)	Landcover	30 m	Static	$\mathbb{R}^8$
	Soil properties (Polaris)	Clay, Sand content, and Bulk density	30 m	Static	$\mathbb{R}^3$
	Location	Latitude and Longitude	N\A	N\A	$\mathbb{R}^2$
	Time	Day of Year (DoY)	N\A	N\A	$\mathbb{R}^1$
Time-series data	Weather (gridMet)	eto, etr, pr, sph, srad, vpd, and vs	∼4 km	Daily	$\mathbb{R}^{10\times6}$
	Land surface temperature (MODIS)	LST	1 km	Daily	$\mathbb{R}^{10\times1}$

2.2. Data Sets Description

To perform spatial downscaling of SMAP SM data, we integrated a diverse set of multisource auxiliary variables from satellite observations, reanalysis weather products, and static surface data sets (Table 1). The SMAP Level-4 (L4) SM product was selected as the coarse-scale reference due to its high precision (M. Reichle et al., 2022, p. 7). From Landsat 8, optical and shortwave infrared bands were used to calculate four vegetation and water indices: NDVI, enhanced vegetation index (EVI), normalized difference water index (NDWI), and land surface water index (LSWI), to reflect capture vegetation conditions and surface characteristics relevant to surface SM (Xu et al., 2022). Synthetic-aperture radar (SAR) backscatter coefficients from Sentinel-1 were used to compute the cross-polarization ratio (CR), modified dual-polarization SAR vegetation index (DPSVIm), and modified radar vegetation index (Pol) to offer insights into vegetation structure and soil properties. Time-invariant attributes such as terrain, landcover, and soil properties were included to account for spatial heterogeneity and long-term environmental controls. To capture weather-driven impacts on SM dynamics, we incorporated reanalysis weather variables from gridMet (Abatzoglou, 2013a, 2013b), including daily total precipitation (pr), surface downward shortwave radiation (srad), wind velocity at 10 m (vs.), specific humidity (sph), mean vapor pressure deficit (vpd), daily grass reference evapotranspiration (eto), and daily alfalfa reference evapotranspiration (etr). Additionally, daily LST were obtained from Moderate Resolution Imaging Spectroradiometer (MODIS). Following the term in Zhu et al. (2024), we organized the data sets into two modalities: snapshot data and time-series data. Snapshot data represent the conditions on the day SM is to be estimated, and time-series data capture the lagged and accumulative effects of weather factors.

Data from all the aforementioned data sets were extracted for each station, which were preprocessed and downloaded from the Google Earth Engine (GEE) platform (Gorelick et al., 2017). For Sentinel-1, a Gaussian filter was applied to reduce noise in the backscatter signal; for Landsat 8, cloud and cloud shadow pixels were masked using the QA band. All data sets were resampled to a 100-m spatial resolution using nearest neighbor interpolation and standardized to a 1-day temporal resolution using linear interpolation. These choices offer a practical balance between preserving spatial and temporal details and avoiding excessive artifacts introduced by resampling schemes. For each sample, the input features were constructed from the snapshot data of the estimated day and the time-series data of the preceding 10 days. The in situ SM observations were used to train and evaluate the models following protocols in Huang et al. (2020) and O and Orth (2021). In total, 502,267 samples were collected from 2016 to 2022.

3. Methodology

In this section, we describe our proposed MMNet for SM downscaling. We begin with introducing the network architecture, followed by the loss function used to train the model. We then give the details of the experimental setup.

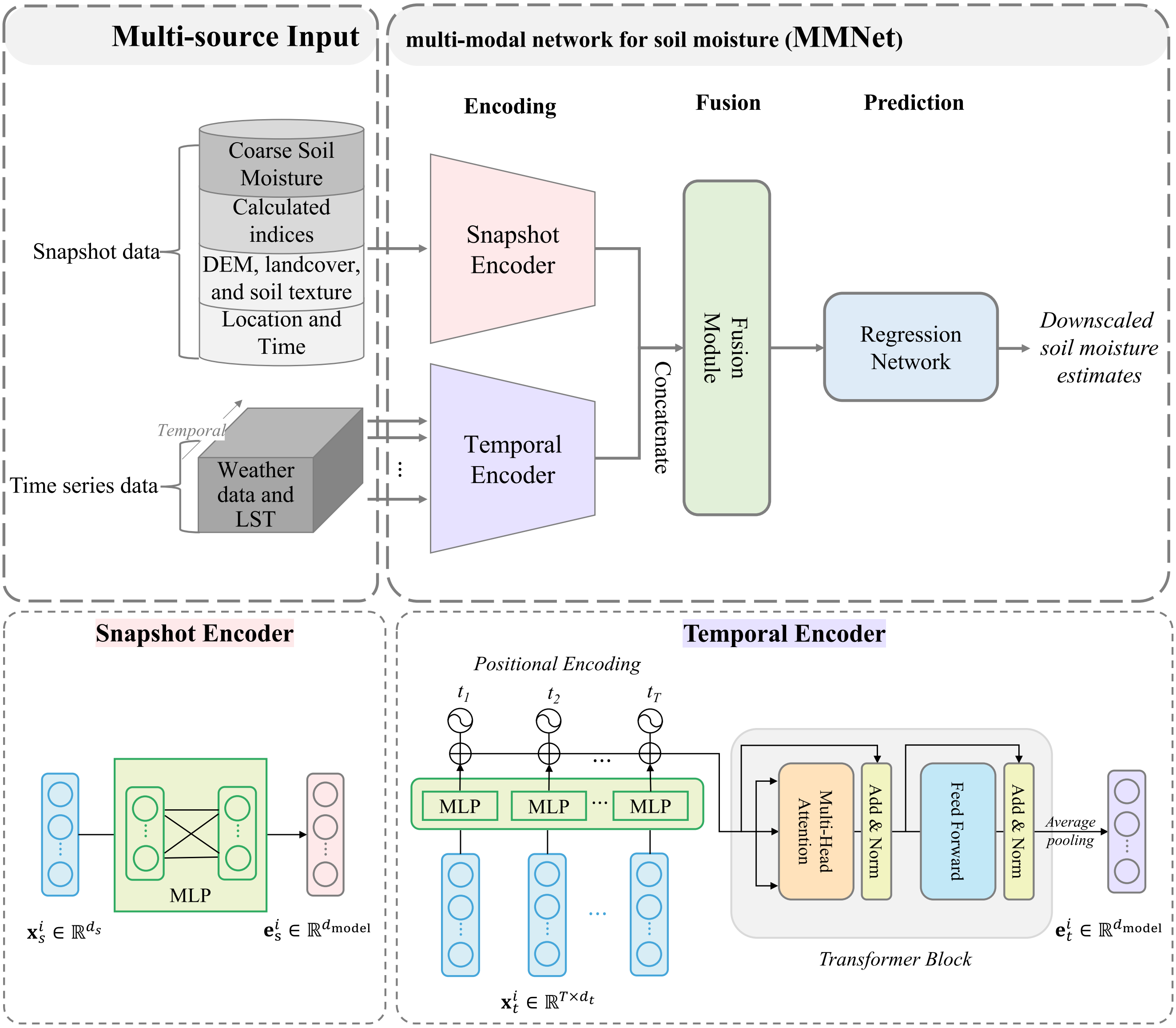

Figure 2. The architecture of the MMNet for SM estimation. Multisource inputs are organized into time-series and snapshot formats, and are processed through two data streams: the snapshot encoder and the temporal encoder. These streams converge in the fusion module, and subsequently flow into the regression network to generate the final downscaled SM estimates. The detailed architectures of both encoders are illustrated in the blocks below.

3.1. Network Architecture

In this study, we developed MMNet, a multimodal model for SM downscaling. As illustrated in Figure 2, MMNet processes input data of two modalities, that is, snapshot and time-series data. The model consists of three main components: (a) encoding—two separate encoders for processing snapshot and time-series data, (b) fusion—a fusion module that combines the representations generated by the encoders, and (c) prediction—a regression network that outputs the downscaled SM estimates. Each component is detailed below.

Encoding. The encoding component consists of two encoders: A snapshot encoder and a temporal encoder. The snapshot encoder processes snapshot data $\mathbf{x}_s^i \in \mathbb{R}^{d_s}$, where $i \in [1, 2, \cdots, N]$; N is the total number of samples, and d_s is the number of features in the snapshot data. It uses a three-layer multilayer perceptron (MLP) to generate the snapshot representation $\mathbf{e}_s^i \in \mathbb{R}^{d_{\text{model}}}$. This encoder captures the relationship between the estimated day's satellite observations and the time-invariant variables related to location-specific patterns, allowing the model to differentiate between SM levels that may have similar satellite observations but differ in geospatial attributes.

The temporal encoder takes input the time-series data $\mathbf{x}_t^i \in \mathbb{R}^{T \times d_t}$, where d_t is the number of features in the time-series data, and T is the temporal length of time-series data. The temporal encoder is built upon the Transformer encoder architecture (Vaswani et al., 2017). It uses a self-attention mechanism to capture the relationships between elements in a temporal sequence, enabling the model to understand how weather patterns evolve over time and affect SM. The time-series data $\mathbf{x}_t^i$ are first passed through a three-layer MLP to project the data into a unified dimension, d_{model}; then, positional information indicating the relative days is incorporated. A multihead self-attention mechanism is then applied to capture the temporal dependencies. This is followed by a position-wise fully connected feedforward network to refine the attention-enhanced features. As in the original work, we also employ residual connections followed by layer normalization to stabilize the learning process. Subsequently, the temporal feature is compressed using average pooling, resulting in $\mathbf{e}_t^i \in \mathbb{R}^{d_{\text{model}}}$.

Fusion. Both the snapshot and time-series features, $\mathbf{e}_s^i$ and $\mathbf{e}_t^i$, are concatenated to form a unified representation, $\mathbf{e}^i \in \mathbb{R}^{2 \times d_{\text{model}}}$. The fusion module then applies a fully connected layer to this concatenated feature, reducing its dimensionality to d_{model} and producing $\mathbf{f}^i \in \mathbb{R}^{d_{\text{model}}}$. The simple yet effective fusion layer enables the model to learn interactions between the snapshot and time-series representations. From the data perspective, snapshot and time-series data jointly influence the SM—for instance, soil properties captured in the snapshot data may influence how recent precipitation affects current SM levels. By this fusion step, the model effectively captures the dependencies between the two data modalities in a data-driven way.

Prediction. The fused representation $\mathbf{f}^i$ is then passed through the regression network to produce the final SM estimation. The regression network consists of a three-layer MLP that progressively refines the fused representation. Each layer in the MLP reduces the feature dimensionality, capturing increasingly abstract representations of the fused snapshot and time-series features. Finally, the output layer of the MLP produces the SM estimates $\hat{y}^i$ at downscaled spatial resolution.

3.2. Loss Function

The loss function for SM estimates is computed using mean square error (MSE), which measures the averaged squared differences between estimated and in situ SM values. To mitigate overfitting, L2 regularization is applied to the model parameters, encouraging the model to maintain simpler patterns that generalize better to unseen data. The model is trained to minimize the following objective:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}^i - y^i\right)^2 + \lambda\|\mathbf{w}\|_2^2 \tag{1}$$

where $\hat{y}^i$ is the downscaled SM estimates for sample i, and y^i is the corresponding in situ SM measurements. In the L2 regularization term, $\|\mathbf{w}\|_2^2$ is the squared L2 norm of the model parameters $\mathbf{w}$, and λ is the weighting parameter that adjusts the trade-off between the MSE loss and the L2 regularization. In our experiments, λ was set to 1×10^{-4}. To optimize the model parameters and minimize the above objective function, we employed the Adam optimization algorithm.

3.3. Experimental Setup

3.3.1. Training Schema

To evaluate the effectiveness of our proposed MMNet, we conducted experiments under three scenarios: on-site, off-site, and cross-region. These scenarios were designed to provide comprehensive insights into the model's generalizability across varying spatial and temporal conditions (RQ 1 and 2). A summary of these three evaluation scenarios is provided in Table 2.

In the on-site scenario, models were trained and tested at the same stations but across different years to assess temporal generalizability. Specifically, models were evaluated using data from four testing years (2019–2022) with training data from the three years preceding each testing year: 2016–2018 for 2019, 2017–2019 for 2020, 2018–2020 for 2021, and 2019–2021 for 2022. Validation data, comprising 10% of the total training data, were used to assess training performance and for model selection. In this scenario, the approximate numbers of training and testing samples were 200,000 and 70,000, respectively.

Table 2
Summary of the Three Evaluation Scenarios

Scenario	Goal	Train/Test split strategy	#Train/#Test
On-site	Temporal generalizability	Same locations across different years	200 k/70 k
Off-site	Spatial generalizability	Random split of stations (5-fold CV)	400 k/100 k
Cross-region	Geographical generalizability	Train on stations from two regions and test on the third	varies by split

For the off-site scenario, models were trained on data from randomly selected stations and tested on the remaining stations, utilizing data from all available years (2017–2022). This scenario aimed to evaluate the model's spatial generalizability. Specifically, fivefold cross-validation was employed, where models were trained on four folds and tested on the remaining fold, with 224 stations used for training and 56 stations for testing. The validation sets were generated similarly to the on-site scenario. The approximate numbers of training and testing samples were 400,000 and 100,000, respectively.

In the cross-region scenario, the training and testing sets are based on geographical regions to assess the model's generalizability across significantly different geographical conditions. Specifically, the CONUS was divided into west, middle, and east regions, as shown in Figure 2. Models were trained on two regions and tested on the third, yielding three experiments: trained on middle and east, tested on west (M, E→W); trained on west and east, tested on middle (W, E→M); and trained on west and middle, tested on east (W, M→E). The sample sizes for the three regions were approximately 200,000, 170,000, and 120,000 for west, middle, and east, respectively. As a result, different combinations of training and testing regions led to varying training and testing sizes. Due to the substantial differences in climate and physiographic conditions across these regions, this cross-region scenario represents the most challenging transfer setting.

3.3.2. Comparisons

Although several ML/DL methods have been proposed for SM estimation or downscaling (Batchu et al., 2023; Xu et al., 2022; Zhu et al., 2024), a direct comparison with them was infeasible due to different experimental settings (Zhu et al., 2024). In this study, to validate the influence of different data modalities on model transferability (RQ3), we compared MMNet with two DL models: MLP and Transformer. Both models utilize the same regression network as MMNet but differ in their encoder structures. Specifically, MLP processes snapshot data, while the Transformer processes time-series data, allowing us to assess how different data modalities impact model performance across different scenarios.

- MLP: This model focuses solely on the snapshot encoder component of MMNet. It utilized all data types listed in Table 1 but only retained weather and LST data from the estimated day rather than incorporating time-series data. This approach represents a typical ML method for SM estimation, where models are often developed using static and snapshot data from the day of estimation.
- Transformer: This model, on the other hand, exclusively employs the temporal encoder. In this setup, the SMAP L4 data set used for downscaling is extended into a time-series and combined with other time-series data used in MMNet. Specifically, the input of the Transformer includes SMAP L4, weather, and LST data from the 10 days preceding the estimated day. This approach aims to assess the influence of time-series data on the transferability of DL models for SM estimation.

Additionally, in Section 5.1, we compared MMNet with MLP$_{\text{full}}$ and Transformer$_{\text{full}}$, both of which leveraged multimodal data but employed different processing strategies. Specifically, MLP$_{\text{full}}$ concatenated snapshot data with flattened time-series data, while Transformer$_{\text{full}}$ appended snapshot data to each time step of the time-series data. This comparison allows us to evaluate the effectiveness of different model architectures in integrating multimodal data.

3.3.3. Evaluation Metrics

To evaluate the performances of the models, we selected four metrics: The Pearson correlation coefficient (R), root mean square error (RMSE), bias, and unbiased RMSE (ubRMSE). Following the practices of Karthikeyan and Mishra (2021) and Zhu et al. (2024), we calculated these metrics for individual stations, referred to as station-wise performance, and analyzed the overall statistics and distribution of the metrics. R assesses how well the model captures temporal trends at a given station by measuring the correlation between estimated and in situ SM values. Root mean square error reflects the overall magnitude of error between estimates and actual SM measurements. Bias quantifies any systematic overestimation or underestimation by the model, providing insight into consistent deviations at each station. Lastly, ubRMSE evaluates the model's precision after accounting for bias, which is particularly useful when we prioritize capturing SM dynamic patterns over absolute values. Together, these metrics provide a comprehensive evaluation of model performance and are calculated as follows:

$$\mathrm{R} = \frac{\sum_{i=1}^{n}\left(y^i - \overline{y}\right)\left(\hat{y}^i - \overline{\hat{y}}\right)}{\sqrt{\sum_{i=1}^{n}\left(y^i - \overline{y}\right)^2 \sum_{i=1}^{n}\left(\hat{y}^i - \overline{\hat{y}}\right)^2}} \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}^i - y^i\right)^2} \tag{3}$$

$$\mathrm{bias} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}^i - y^i\right) \tag{4}$$

$$\mathrm{ubRMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left(\hat{y}^i - \overline{\hat{y}}\right) - \left(y^i - \overline{y}\right)\right)^2} \tag{5}$$

where $\overline{y}$ is the mean of the in situ SM measurements y^i, $\overline{\hat{y}}$ is the mean of the estimated SM $\hat{y}^i$, and n is the number of samples at a station.

3.3.4. Implementation Details

In our experiments, the snapshot data have a feature dimension of $d_s = 25$, and the time-series data have $d_t = 7$ and $T = 10$, representing 10-day sequences of 7 variables. The snapshot encoder is a three-layer MLP with hidden sizes of 32, 64, and 128, projecting the 25-dimensional snapshot input into a 128-dimensional feature space. The time-series encoder first applies a shared MLP with hidden sizes of 32, 64, and 128 to each time step independently, followed by a single-layer Transformer encoder with $d_{\mathrm{model}} = 128$, 16 attention heads, and a feedforward dimension of 256 to capture temporal dependencies. The fusion layer is a fully connected layer that maps the concatenated features $\mathbf{e}^i \in \mathbb{R}^{256}$ to 128 dimensions. The regression network is a three-layer MLP with hidden dimensions of 64, 32, and 1, producing the final SM estimate. Batch normalization and ReLU activations are applied between layers to improve training stability. In the training process, the model is trained for 100 epochs using the Adam optimizer (learning rate = 0.01 and weight decay = 5e−4), with a batch size of 512. These hyperparameters were tuned on the validation set for the 2019 on-site experiment and applied the same configuration across all experiments (as done in Nyborg et al. (2022).

All comparison models were with identical model components and training settings for fair comparison as our model. The MLP models follow the same architecture as the snapshot encoder followed by the regression network. The base MLP has an input dimension of 32, while $\mathrm{MLP}_{\mathrm{full}}$ flattens the 10-day time-series and concatenates it with snapshot data, resulting in an input dimension of 95. The Transformer models adopt the same temporal encoder and the same regression network as in our model. The input to the base Transformer has a shape of 10×8, while $\mathrm{Transformer}_{\mathrm{full}}$ repeatedly appends snapshot features to each time step, yielding an input shape of 10×32. All models were implemented in PyTorch and trained on a single NVIDIA RTX A5000 GPU.

4. Results

4.1. Results for the On-Site Scenario

The scatterplots of the predicted SM versus measured SM for three models on the validation and testing sets in the on-site scenario are presented in Figure 3. Since the validation sets are dependent on the training sets, they reflect the models' training performance. The results indicated that MMNet consistently outperformed the other models in both validation and testing phases, demonstrating a strong correlation between predicted and measured SM values, with R-values of 0.9703 and 0.9198, respectively. The MLP model also showed effective performance in

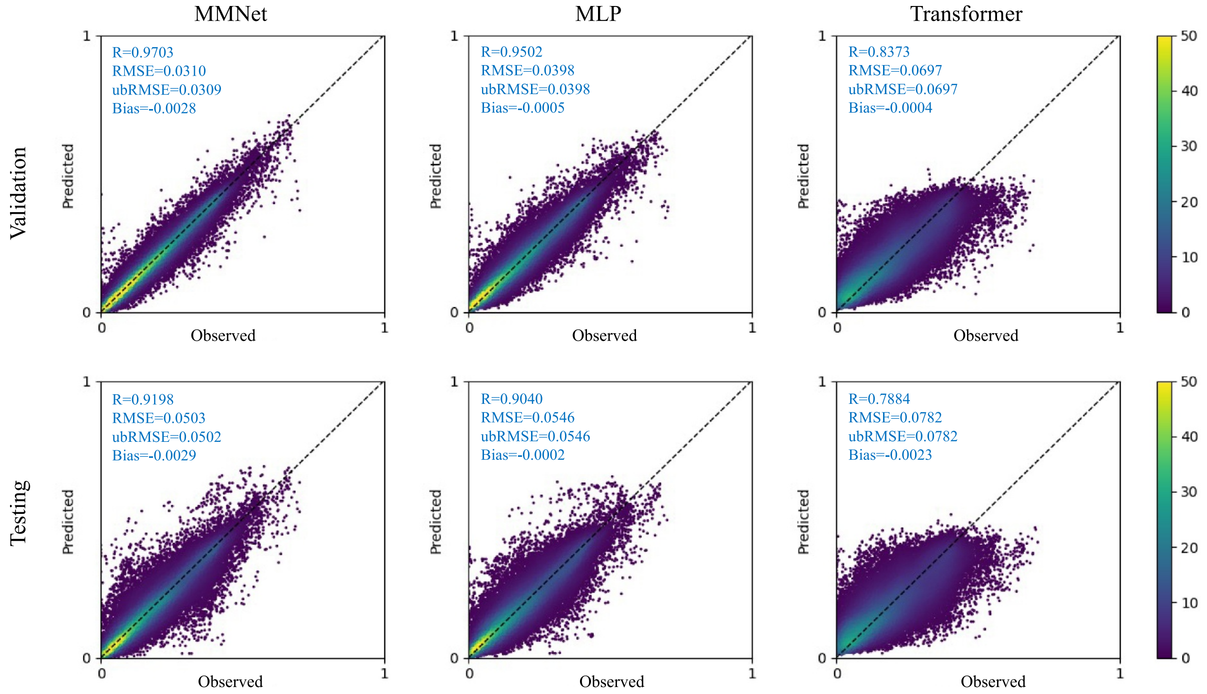

Figure 3. Scatter plots of predicted SM versus measured SM for three models on the validation and testing sets. Each plot shows 50,000 random samples from the four testing years (2019–2022), with color intensity representing the density of data points. The metrics shown in this figure were calculated using all samples.

this on-site scenario, though slightly lower than MMNet. The RMSE values for MMNet and MLP on the testing set were 0.0502 and 0.0546 m^3/m^3, respectively, both surpassing the general SAR retrieval target of 0.06 m^3/m^3 (Zhu et al., 2024). The robust performance of MMNet and MLP suggested that when historical in situ SM measurements are available, employing diverse data types in a DL model could produce reliable SM estimates, making it a viable approach for filling gaps or extending SM time-series. In contrast, the Transformer model exhibited worse performance, with its R-value dropping to 0.7884 during testing. The scatterplots for the Transformer showed a noticeable truncation, with its predicted SM range being narrower than the observed range. This indicated that the snapshot data, such as soil properties, terrain attributes, and remote sensing observations, were crucial for accurate SM estimation. Relying solely on SMAP L4 data and weather time-series was proven insufficient. These findings highlight the importance of integrating various data types to provide rich information and enable robust SM estimation.

We then presented the station-wise performance of the three models across the testing years from 2019 to 2022. The median values of the metrics for each testing year are presented in Table 3. To provide a more comprehensive view, spatial distributions of the metrics for each station are shown in Figure S1 of Supporting Information S1, and boxplots of these metrics for the three models and the SMAP baseline are provided in Figure S2 in Supporting

Table 3
Median Station-Wise Performance Metrics for the Three Models in the On-Site Scenario

	MMNet				MLP				Transformer			
	R	RMSE	bias	ubRMSE	R	RMSE	bias	ubRMSE	R	RMSE	bias	ubRMSE
2019	**0.8594**	**0.0387**	−0.0055	**0.0340**	0.7657	0.0475	−0.0040	0.0422	0.7355	0.0661	**−0.0038**	0.0490
2020	**0.8546**	**0.0374**	−0.0028	**0.0307**	0.7786	0.0420	**0.0003**	0.0374	0.7340	0.0612	−0.0008	0.0473
2021	**0.8342**	**0.0395**	−0.0048	**0.0329**	0.7014	0.0454	**−0.0026**	0.0388	0.7011	0.0628	−0.0059	0.0460
2022	**0.8358**	**0.0423**	**−0.0008**	**0.0357**	0.7508	0.0477	0.0017	0.0401	0.7066	0.0627	0.0027	0.0499
Average	**0.8460**	**0.0395**	−0.0035	**0.0333**	0.7491	0.0457	**−0.0012**	0.0396	0.7193	0.0632	−0.0020	0.0481

Note. The best-performing ones for each year and the average results are highlighted in bold.

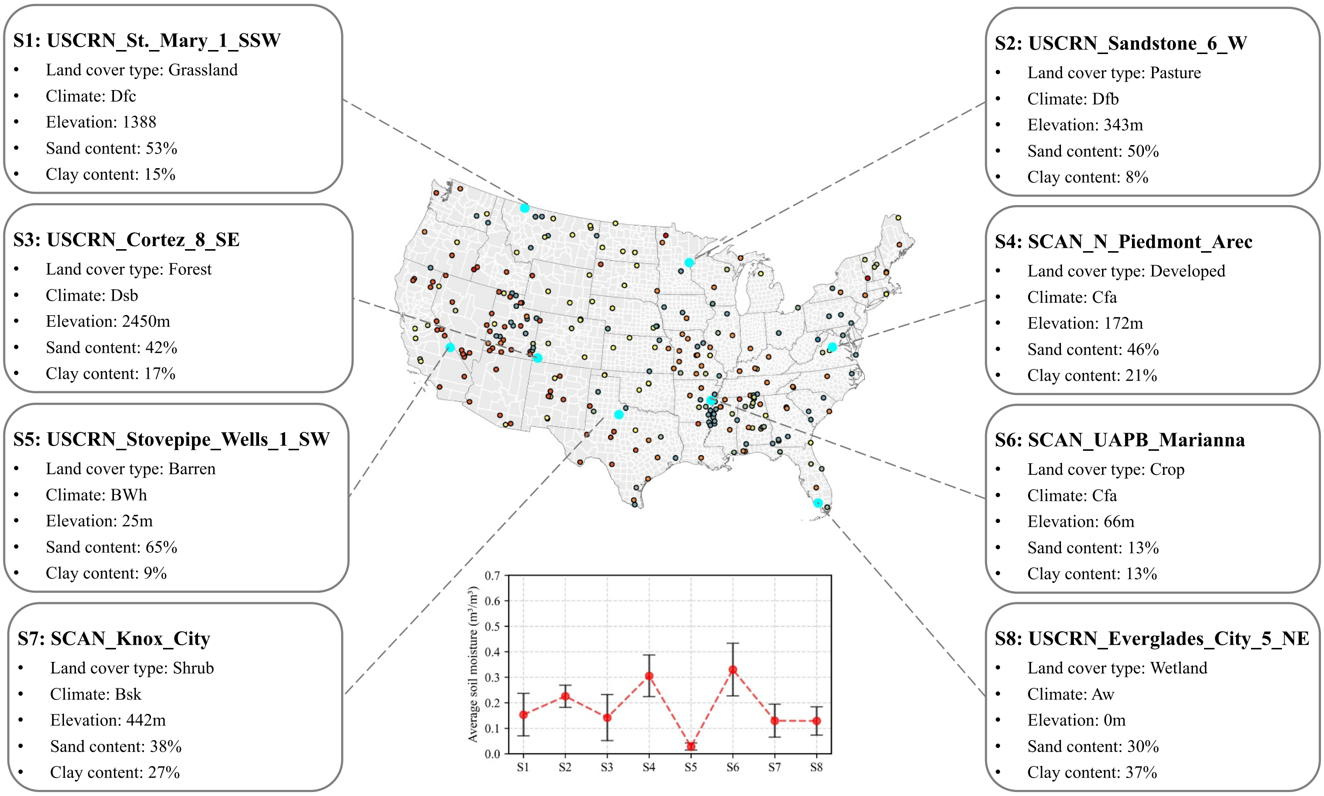

Figure 4. Overview of the selected stations, each characterized by their landcover type, climate, elevation, and soil properties. The mean and standard deviation SM values for each station from 2016 to 2022 are also indicated in this figure.

Information S1. Overall, MMNet consistently outperformed the other models across all years and locations. Notably, it remained relatively stable performance over the years, with median R-values consistently above 0.83 and RMSE values around 0.04 m^3/m^3. In contrast, although the MLP model achieved similar station-wise performance to MMNet in Figure S1 in Supporting Information S1, its performance exhibited more variation over the years, as shown in Table 3. The R-value for MLP ranged from 0.7014 to 0.7786, indicating that MLP was more sensitive to year-to-year variability. The Transformer model performed the worst, with RMSE values consistently exceeding 0.06 m^3/m^3 each year (Table 3) and large biases at several stations (Figure S1 in Supporting Information S1). This outcome reinforces the finding that time-series data alone are insufficient for accurate SM estimation, highlighting the importance of integrating snapshot data in modeling. In this scenario, all methods achieved satisfactory ubRMSE values across stations, with median values below 0.06 m^3/m^3 across years. This demonstrates that DL models can effectively capture SM dynamics when historical data are available. Furthermore, as shown in Figure S2 in Supporting Information S1, MMNet achieved the highest R-values, the lowest RMSE and ubRMSE, and nearly zero bias. This highlights the effectiveness of incorporating auxiliary data sets for substantial improvement brought to SMAP SM through MMNet's ability to integrate multimodal data.

To further illustrate the performances of three different models, we selected eight stations from both networks, each representing diverse landcover, climate, geographic, and soil conditions, for visual comparison of how well each model captures the temporal dynamics of SM across varying environmental conditions. Figure 4 displays the geolocations and general information of these stations (denoted as S1 to S8). A detailed description of the climate and soil characteristics of the selected stations can be seen in the Supporting Information S1.

Figure 5 illustrates the SM time-series from the three models—MMNet, MLP, and Transformer—alongside SMAP, compared against in situ SM measurements. The detailed numerical performances for each station are provided in Table S1 in Supporting Information S1. Across all stations, MMNet consistently showed strong temporal alignment with in situ SM, maintaining RMSE values below 0.06 m^3/m^3 and responding to precipitation events. More detailed station-wise analysis is provided in Text S2 of Supporting Information S1.

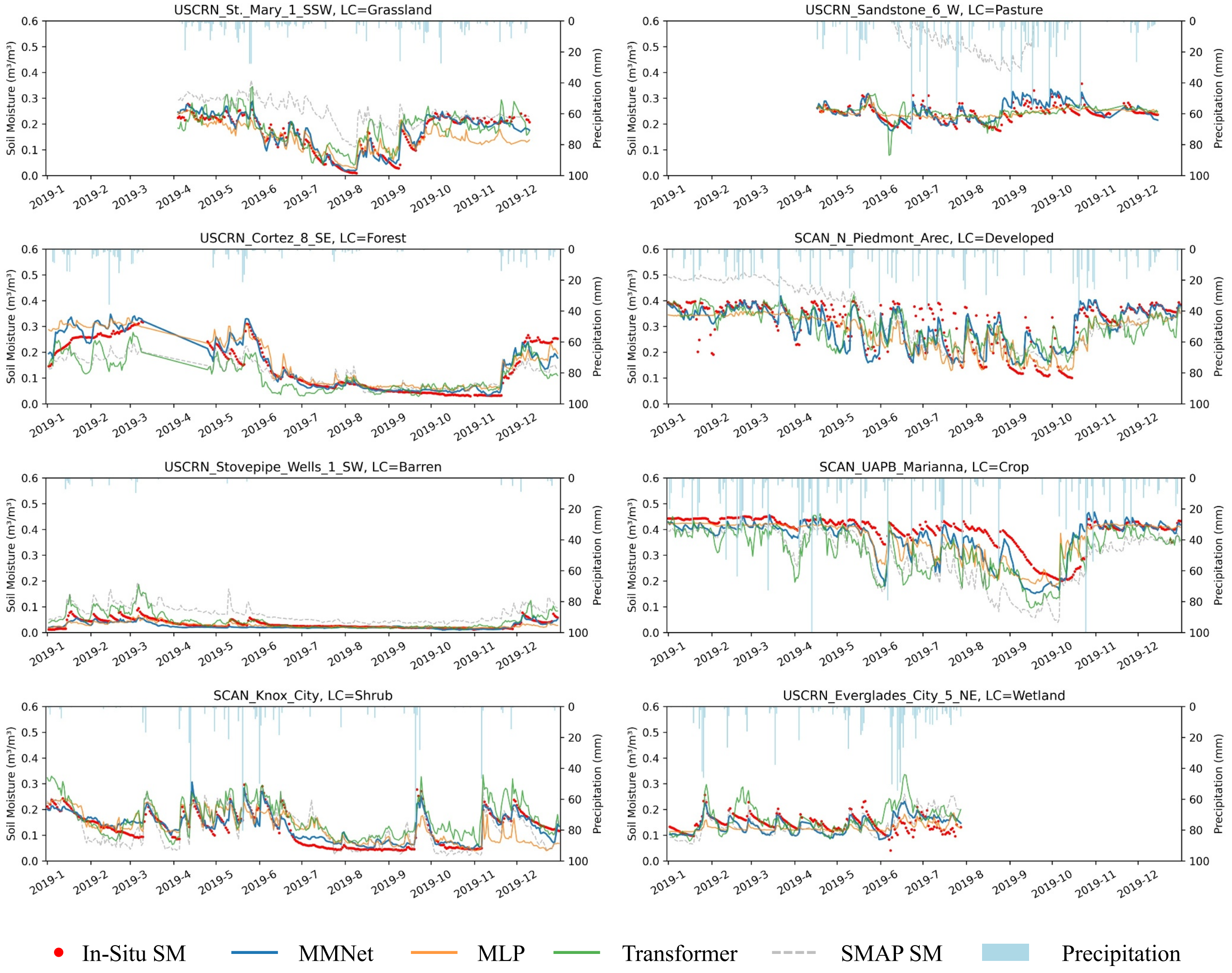

Figure 5. Time-series SM in the on-site scenario at selected stations. Each plot includes in situ, MMNet, MLP, Transformer, and SMAP L4 SM data, with daily precipitation displayed on the inverted *y*-axis.

Furthermore, since MMNet estimates showed good agreement with in situ observations across diverse conditions, we suggest that it has the potential to generate gap-filled or extended SM time-series. To illustrate this capability, Figure S3 in Supporting Information S1 shows the complete SM time-series at four stations from Figure 5 with missing observations. These results demonstrate that, by leveraging in situ observations for training, MMNet can produce temporally continuous SM estimates during periods without ground observation, enabling long-term monitoring and analysis of SM dynamics.

In summary, within the on-site scenario, the DL models demonstrated their effectiveness in SM estimation, showing significant improvements over SMAP. This underscores that on-site SM measurements are crucial for developing robust models to estimate SM for periods of absent observational data. Among the models, MMNet consistently exhibited the most accurate and robust performances, adeptly capturing SM dynamics across various conditions. In contrast, the MLP model, which relied solely on snapshot data, was less responsive to SM variations. Meanwhile, the Transformer model often aligned closely with SMAP values or overreacted to weather patterns.

4.2. Results for the Off-Site Scenario

To assess the spatial generalizability of the models, we performed off-site experiments. Results for each fold in the fivefold cross-validation are shown in Table 4. Similar to the on-site scenario, MMNet consistently achieved

Table 4
Median Station-Wise Performance Metrics for the Three Models in the Off-Site Scenario

	MMNet				MLP				Transformer			
	R	RMSE	bias	ubRMSE	R	RMSE	bias	ubRMSE	R	RMSE	bias	ubRMSE
Fold 1	**0.7659**	**0.0715**	−0.0095	**0.0489**	0.6900	0.0818	**−0.0053**	0.0572	0.7129	0.0742	−0.0188	0.0610
Fold 2	**0.7851**	**0.0659**	**0.0031**	**0.0446**	0.6770	0.0814	0.0115	0.0527	0.7164	0.0747	0.0033	0.0542
Fold 3	**0.7729**	0.0644	**0.0021**	**0.0438**	0.6704	0.0672	0.0030	0.0504	0.7260	**0.0643**	−0.0058	0.0516
Fold 4	**0.7997**	**0.0653**	**0.0008**	**0.0452**	0.7097	0.0738	−0.0098	0.0510	0.7440	0.0714	−0.0032	0.0565
Fold 5	0.7816	0.0682	−0.0035	**0.0458**	**0.7833**	**0.0466**	0.0056	0.0442	0.7093	0.0648	**−0.0020**	0.0535
Overall	**0.7816**	**0.0672**	**−0.0004**	**0.0456**	0.7043	0.0717	0.0035	0.0501	0.7240	0.0697	−0.0047	0.0545

Note. The results for each fold are computed using models trained on the other four folds. "Overall" represents the statistics across all stations. The best-performing ones for each fold and the overall results are highlighted in bold.

the best overall performance and demonstrated higher transferability, with a ubRMSE of 0.0456 m^3/m^3 across all stations (Overall in Table 4). The Transformer model showed comparable transferability to that of MLP, generally achieving higher R-values and more stable RMSEs, but worse ubRMSE values. This could be attributed to Transformer's focus on learning temporal patterns from time-series data, which enhanced its ability to generalize to unseen stations. However, the lack of additional variables led it to overreact to certain patterns, such as precipitation events, causing SM to show exaggerated increases following rainfall while potentially overlooking the stabilization effects due to soil properties, resulting in less precise estimates. Conversely, MLP's reliance on snapshot data, while effective at capturing station-specific SM responses at training stations, hindered its transferability to unseen stations and diverse environments. MMNet, by integrating both time-series and snapshot data, managed to combine high accuracy in SM estimation with improved spatial transferability.

To further illustrate the model's performance details, we presented spatial distribution maps of R, bias, and ubRMSE in Figure S4 of Supporting Information S1, along with performance metric distributions in Figure S5 of Supporting Information S1. Compared with the on-site scenario, all models experienced a performance decrease due to more challenging off-site settings. A key observation is the worsening bias for both MMNet and MLP. The bias maps in Figure S4 of Supporting Information S1 showed more extreme red and blue. Additionally, the bias distributions for MMNet and MLP in Figure S5 of Supporting Information S1 were broader than those in Figure S2 of Supporting Information S1, indicating higher variability and larger absolute bias values across stations. This suggests that these models are more prone to systematic overestimation or underestimation of SM when applied to previously unseen stations, highlighting ongoing challenges for DL models in accurately estimating SM at new locations. However, excluding the bias component of the error, the ubRMSE for most stations remained low, as depicted by the deep blue to light blue gradient across stations in Figure S4 of Supporting Information S1. Furthermore, Figure S5 of Supporting Information S1 demonstrated that the distributions of R and ubRMSE for MMNet outperformed those of SMAP, confirming its advantage in maintaining consistency with in situ SM dynamics in this off-site scenario.

We further compared the SM time-series of the three models at selected stations, as shown in Figure 6, and provided detailed numerical performance metrics for each station in Table S2 of Supporting Information S1. Overall, MMNet estimates generally aligned with the in situ SM temporal variability, maintaining ubRMSE values below 0.06 m^3/m^3 across all selected stations. In contrast, MLP and Transformer showed performance degradation at several stations compared to the on-site scenario (Figure 5). The high bias issues were also evident in the time-series plot, while MMNet demonstrated better consistency with the in situ temporal dynamics. Specific examples and detailed analyses are provided in detail in Text S3 of Supporting Information S1.

In summary, the off-site experiments demonstrated that snapshot data provided critical SM-related information to enhance estimation accuracy, while time-series data enabled models to capture temporal trends, ensuring stable estimation and improved transferability. By integrating both data modalities, MMNet leveraged their strengths, achieving superior performance to the original SMAP. Despite some systematic bias, MMNet effectively captured SM dynamics consistent with in situ measurements, making it a viable option for generating spatially

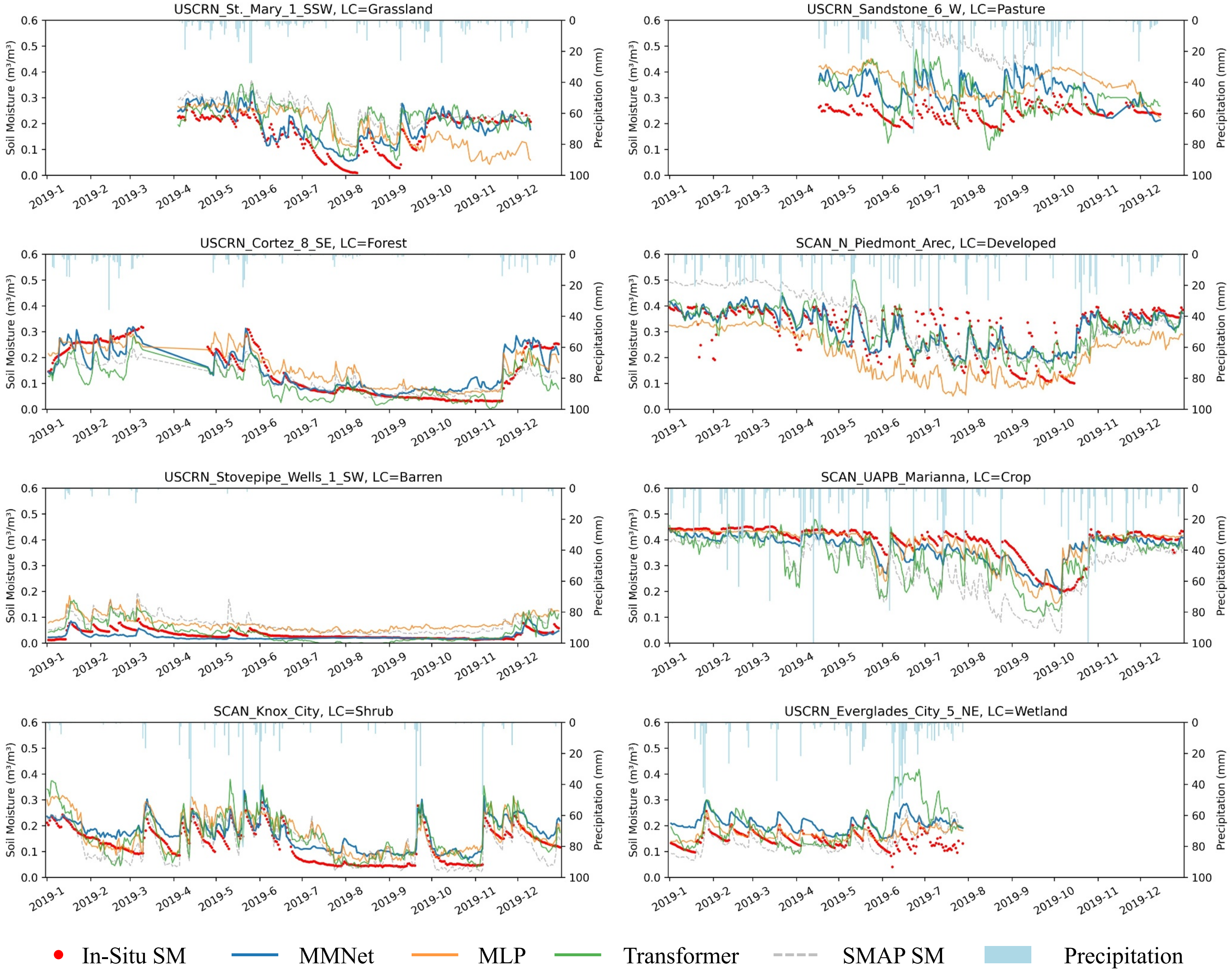

Figure 6. Time-series SM in the off-site scenario at selected stations. Each plot includes in situ, MMNet, MLP, Transformer, and SMAP L4 SM data, with daily precipitation displayed on the inverted *y*-axis.

continuous SM estimates in regions with sparse in situ stations, particularly when the focus is on capturing temporal variability rather than absolute values.

4.3. Results for the Cross-Region Scenario

In the off-site experiments, although the training and testing stations were independent, they were often located close to one another, suggesting that the region of interest had some sparse station coverage. However, scenarios where no stations are available within the region of interest are also worth exploring. Toward this end, we conducted cross-region experiments, where the models were trained on data from one or more regions and then tested in a distinct region. As shown in Figure 1, the three regions exhibit significant differences in both geospatial environments and SM levels. In ML, models trained on labeled data from one domain often perform poorly when applied to a different domain, a challenge known as "domain shift" (Ji et al., 2024; Ma et al., 2024). In this context, while high bias is expected, the primary goal is to assess whether the models can still capture reliable SM dynamics.

The median performance metrics for all regions in the cross-region experiments are summarized in Table 5. Overall, MMNet demonstrated the best performance, with median R-values of 0.7268 and ubRMSE of 0.0527 m^3/m^3, highlighting its ability to capture SM dynamics in this cross-region scenario. The performance relationships between MLP and the Transformer were consistent with those observed in the off-site experiments,

Table 5
Median Station-Wise Performance Metrics for the Three Models in the Cross-Region Scenario

	MMNet				MLP				Transformer			
	R	RMSE	bias	ubRMSE	R	RMSE	bias	ubRMSE	R	RMSE	bias	ubRMSE
M, E→W	**0.7295**	**0.0654**	**−0.0004**	**0.0472**	0.6330	0.0726	−0.0153	0.0535	0.6842	0.0663	0.0151	0.0527
W, E→M	**0.7347**	**0.0815**	−0.0197	**0.0604**	0.6742	0.0831	**0.0057**	0.0606	0.6889	0.0912	−0.0419	0.0659
W, M→E	**0.7124**	0.0831	**−0.0092**	**0.0532**	0.5631	0.0991	−0.0117	0.0572	0.6629	**0.0815**	0.0342	0.0561
Overall	**0.7268**	**0.0759**	−0.0105	**0.0527**	0.6456	0.0805	−0.0063	0.0559	0.6819	0.0764	**−0.0006**	0.0570

Note. "Overall" represents the statistics across all stations. The best-performing ones for each experiment and the overall results are highlighted in bold.

where the Transformer generally achieved higher R-values and lower RMSE, but with slightly worse ubRMSE. This suggested that time-series data enhanced model transferability, while snapshot data provided the necessary information to ensure an accurate SM dynamics range. Spatially, as shown in Figure S6 of Supporting Information S1, MMNet maintained acceptable R-values across CONUS, with most stations showing R-values greater than 0.7, demonstrating its adaptability to diverse geospatial conditions and its ability to preserve temporal correlations between model estimates and in situ measurements. However, the station-wise bias maps revealed that all three models tended to systematically overestimate or underestimate in specific regions. For example, the models underestimated SM (dark blue) in areas with higher SM levels, such as Texas (middle) and the Mississippi River Basin (middle and east), while overestimating (dark red) in drier regions like the western coast (west) and Florida (east). These patterns underscore the difficulty DL models faced in capturing the true magnitude of SM levels, particularly under extreme conditions. Additionally, as shown by the metrics distributions in Figure S7 of Supporting Information S1, both MLP and the Transformer slightly underperformed as compared with the original SMAP. This further confirms that significant discrepancies between training and testing data can degrade the accuracy of DL models in SM estimation. However, MMNet achieved slight improvements in both R and ubRMSE, suggesting that it can deliver comparable or better SM dynamic estimates while performing spatial downscaling.

The cross-region SM time-series plots are shown in Figure 7, with numerical performances detailed in Table S3 of Supporting Information S1. Overall, DL models struggle to make accurate estimates, exhibiting even larger bias than in the off-site scenario. Nevertheless, MMNet was able to effectively capture SM dynamics across diverse environments, while MLP and the Transformer encountered performance issues at certain stations. A detailed analysis of model performance for each domain is provided in Text S4 of Supporting Information S1.

In summary, the cross-region experiments revealed that the domain shift challenges hindered the ability of DL models trained in one region to provide reliable SM estimates in another. In other words, DL models struggled to produce accurate SM estimates for regions without in situ station coverage, particularly under extreme or region-specific conditions. This aligns with the limitation of data-driven ML models, which often struggle with out-of-distribution (OoD) generalization (Liu et al., 2023). In this context, MMNet demonstrated relative robustness with lower ubRMSE values and improved temporal dynamics than the original SMAP. This suggests that MMNet is an effective method for SM estimation over large scales (e.g., CONUS), achieving better estimates of SM dynamics and performing spatial downscaling compared to SMAP, even in the absence of observational data.

4.4. Comparison With SMAP and SMAP-HB

To comprehensively evaluate the performance of MMNet, we compared it against the SMAP L4 and the SMAP-HydroBlocks (SMAP-HB) products. Table 6 summarizes the median station-wise metrics computed using all available data during testing years per station. MMNet consistently outperformed or matched the SMAP L4 product across all scenarios. Even in the most challenging cross-region setting, MMNet achieved a higher R (0.7268) and lower ubRMSE (0.0527 m^3/m^3) than SMAP (R = 0.7075, ubRMSE = 0.0535 m^3/m^3), although with a larger bias. Compared to SMAP-HB, MMNet's performance in the off-site scenario is competitive, achieving an R of 0.7816 and ubRMSE of 0.0456 m^3/m^3, better than those of SMAP-HB (R = 0.7159 and ubRMSE = 0.0516 m^3/m^3). These results demonstrate the promise of MMNet for generating spatially and temporally continuous SM products from sparse in situ training data. To further evaluate the impact of

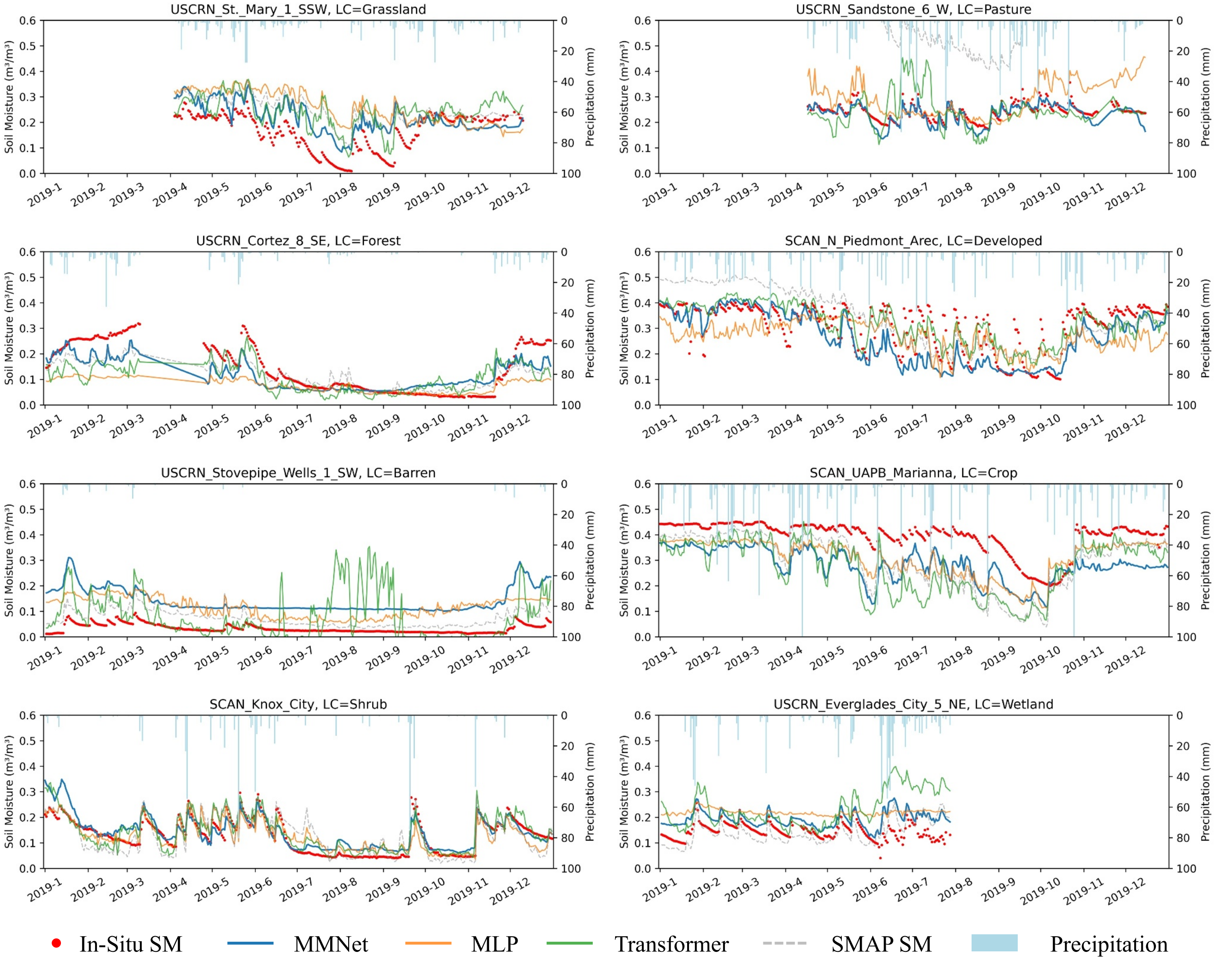

Figure 7. Time-series SM in the off-site scenario at selected stations. Each plot includes in situ, MMNet, MLP, Transformer, and SMAP L4 SM data, with daily precipitation displayed on the inverted *y*-axis.

environmental factors on model performance, Figure S8 in Supporting Information S1 presents the off-site scenario results grouped by soil texture, elevation, landcover type, and annual precipitation, in comparison with SMAP and SMAP-HB. This analysis provides additional insights into MMNet's performance across diverse landscape conditions.

5. Discussion

5.1. Advantages of Multimodal Data Fusion and MMNet

The MMNet utilizes a classic dual-input architecture (Ngiam et al., 2011), processing current snapshot data (including static features) and antecedent weather time-series for SM downscaling. Most existing DL methods for SM estimation rely on single-modality inputs: either current snapshot data only (represented by MLP in Section 3.3.2) (Abowarda et al., 2021; Batchu et al., 2023; Wei et al., 2019) or antecedent time-series data only (represented by Transformer in Section 3.3.2) (Q. Li, Zhu, et al., 2022). Some approaches using sequential models attempt to integrate snapshot features by concatenating them to each time step of the time-series features (Anshuman & Eldho, 2022; Liu et al., 2022), which are computationally less efficient. Other

Table 6
Median Station-Wise Performance Metrics for the Three Scenarios, Compared With Soil Moisture Active Passive and SMAP-HB

	R	RMSE	bias	ubRMSE
On-site	0.8216	0.0425	−0.0037	0.0395
Off-site	0.7816	0.0672	−0.0004	0.0456
Cross-region	0.7268	0.0759	−0.0105	0.0527
SMAP	0.7075	0.0741	0.0054	0.0535
SMAP-HB	0.7159	0.0673	−0.0066	0.0516

Table 7
Median Station-Wise Performance Metrics for the MMNet, MLP_{full}, and $Transformer_{full}$ in Three Scenarios

		R	RMSE	bias	ubRMSE
On-site	MMNet	**0.8216**	**0.0425**	−0.0037	**0.0395**
	MLP_{full}	0.7742	0.0461	**0.0008**	0.0430
	$\text{Transformer}_{\text{full}}$	0.7600	0.0485	−0.0010	0.0451
Off-site	MMNet	**0.7816**	0.0672	**−0.0004**	**0.0456**
	MLP_{full}	0.7685	**0.0671**	**−0.0004**	0.0461
	$\text{Transformer}_{\text{full}}$	0.7496	0.0702	−0.0029	0.0497
Cross-region	MMNet	**0.7268**	**0.0759**	−0.0105	**0.0527**
	MLP_{full}	0.7192	0.0761	**0.0065**	0.0536
	$\text{Transformer}_{\text{full}}$	0.6649	0.0845	0.0095	0.0550

Note. The results follow the calculation in Section 4. The best-performing ones for each scenario are highlighted in bold.

methods flatten time-series data for MLP processing (O & Orth, 2021), which may lose temporal patterns that are critical for SM dynamics. While Zhu et al. (2024) utilized a TempCNN block (Cui et al., 2016) to process time-series data, the time-series features span the past year with an 8-day interval to accommodate the coarse temporal resolution of NDVI. However, for SM estimation, where rapid responses to meteorological factors—especially recent precipitation and evapotranspiration—are critical, using time-series with coarse temporal resolution and a long historical window may limit the model's ability to accurately capture SM dynamics, as past events have a diminishing influence on current SM (Han et al., 2023). In contrast, MMNet is more refined in model architecture and integration of physical knowledge: Its dual-input architecture ensures that both data modalities are efficiently utilized, and incorporating recent, high-temporal resolution weather time-series enhances the model's responsiveness to rapid environmental changes.

The advantages of integrating multimodal data were assessed by comparing MMNet with the MLP and the Transformer, without a direct comparison with the abovementioned methods due to different input features, data sources, and metrics calculation. We compare the three models through comprehensive experimental settings in three scenarios: on-site (Section 4.1), off-site (Section 4.2), and cross-region (Section 4.3). Generally, MMNet outperforms the single-modality MLP and the Transformer in both model accuracy and temporal-spatial transferability with in situ data as the reference. Notably, in the on-site scenario, MMNet achieved biases close to 0 (Figure S2 in Supporting Information S1) and a median RMSE $< 0.05\ \text{m}^3/\text{m}^3$ (Table 3), demonstrating its ability to produce reliable SM estimates for gap filling or extending SM time-series when historical in situ SM measurements are available. In off-site and cross-region scenarios, differences in environmental conditions and soil properties across sites led to a decline in all models' performance, exhibiting large biases (Figures S5 and S7 in Supporting Information S1). Nevertheless, MMNet more effectively captured SM dynamics compared to MLP, Transformer, and original SMAP product (Figures 6 and 7). Therefore, if the primary goal is to capture temporal variability, MMNet remains a viable option for generating spatially continuous SM estimates with improved dynamics and performing spatial downscaling in regions with sparse or no in situ stations.

To further evaluate the effectiveness of MMNet's dual-input architecture in leveraging multimodal data, we compared MMNet with MLP_{full} and $\text{Transformer}_{\text{full}}$ under the three scenarios. Table 7 reports the median performance of each model across three scenarios. Overall, MMNet generally outperformed the other models, followed by MLP_{full}. Given the comparable input variables among these models, we attributed this advantage to MMNet's data fusion strategy. Its dual-head structure separately learns snapshot and time-series representations before fusing them. These representations are then integrated via a fully connected fusion layer, allowing the model to capture their interactions. This enables the model to effectively balance and integrate complementary data modalities, resulting in improved SM estimation. In contrast, MLP_{full} mixed flattened time-series data with snapshot data, which not only weakened temporal relationship information but also reduced the model's focus on snapshot features, leading to worse performance than MMNet, particularly in the on-site scenario. $\text{Transformer}_{\text{full}}$, despite using more temporal information (i.e., SMAP SM time-series), showed the least satisfactory performance. This may be due to its repeated input of snapshot data, which not only improves SM estimation but also limits generalization. In comparison, MMNet neither overemphasizes snapshot data nor overlooks it, with the support of temporal information, achieving robust performance in all scenarios. Although MMNet's advantage in off-site and cross-region scenarios is constrained by spatial transferability, expanding the number of training stations could help alleviate this issue. Additionally, unlike MLP_{full}, which requires a fixed input length, MMNet's temporal encoder can handle variable-length time-series, offering greater flexibility to different data settings.

Furthermore, comparing Table 7 with Tables 3, 4, and 5, MLP_{full} showed significant improvements over MLP, highlighting the importance of incorporating antecedent meteorological conditions on SM estimation. Similarly, $\text{Transformer}_{\text{full}}$ outperformed the Transformer in on-site and off-site scenarios, demonstrating the contribution of snapshot data to SM estimation accuracy. However, in the cross-region scenario, $\text{Transformer}_{\text{full}}$ performed worse than the Transformer, suggesting that under severe domain shift, the way snapshot data is integrated may hinder rather than enhance SM estimation accuracy. Overall, these results demonstrated the necessity of leveraging

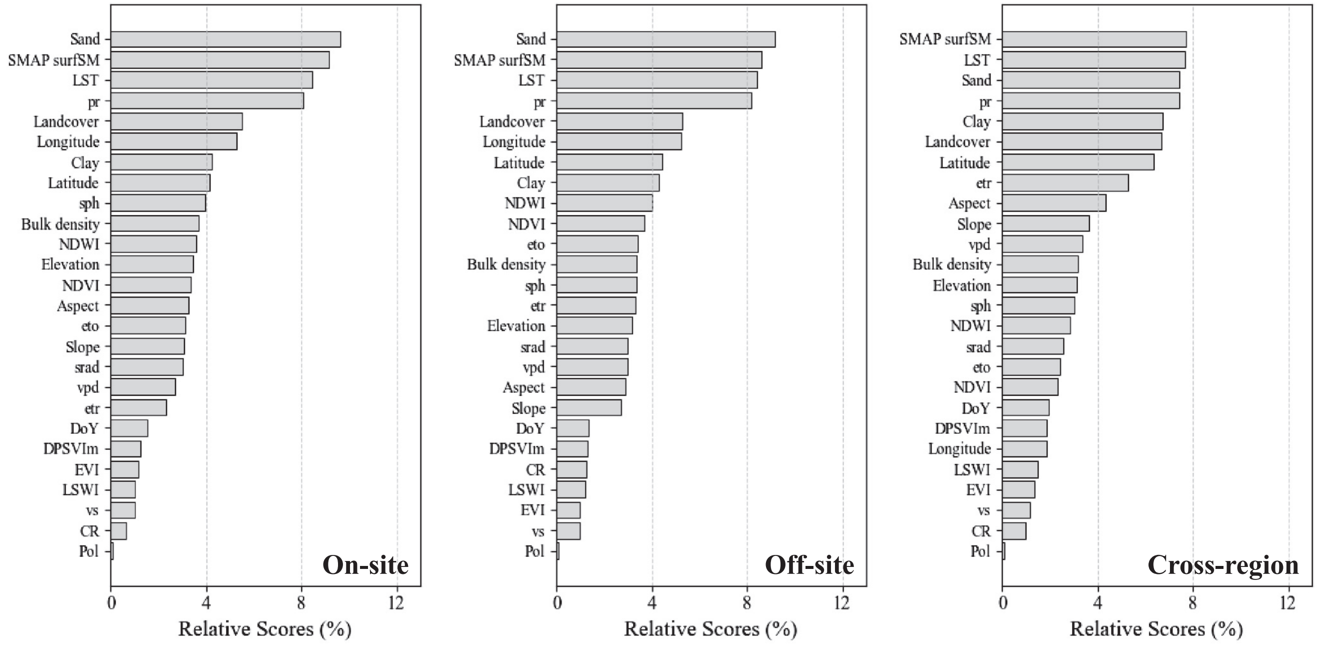

Figure 8. The relative importance of input variables across three experiments (rescaled to percentages summing to 100%). Please see the abbreviations in Section 2.2.

multimodal data. MMNet, by effectively integrating both modalities, achieved satisfactory performance in all scenarios. Future studies should focus on refining snapshot data and its utilization to improve the model's spatial generalizability.

Finally, we emphasize that MMNet is proposed as a protocol model architecture that utilizes a dual-input structure to integrate recent meteorological time-series and estimated day snapshot data for SM downscaling. While the data sets in Table 1 were used in our experiments, MMNet is not restricted to these specific inputs or data sets. Instead, it provides a flexible framework that allows users to incorporate any accessible and relevant input features. The feature importance analysis in Section 5.2 offers further insights into the input data selection for future studies.

5.2. Feature Importance Analysis

We analyzed the feature importance in the proposed MMNet using SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017). The feature importance measured by SHAP GradientExplainer was calculated as the absolute values of expected gradients to quantify the contribution of input variables to the model output. Figure 8 shows the average importance scores of each input variable across three experiments.

Soil Moisture Active Passive surface SM, as the baseline for downscaling, consistently ranked first or second in all experiments. Soil properties, including sand, clay content, and bulk density, also showed notable importance, aligning with the findings in Gaur and Mohanty (2016). Sand content, ranking in the top three across all scenarios, showed the highest impacts among the soil properties. This is due to its influences on soil water retention capacity, which directly affects surface SM dynamics (Han et al., 2023). LST and precipitation were the two most important time-series variables. LST was closely linked to surface SM by influencing evaporation and transpiration processes (Cammalleri & Vogt, 2015; Tian et al., 2023). Meanwhile, precipitation directly drives SM inputs (Li et al., 2016; Mondal & Mishra, 2024). Compared to the findings of Karthikeyan and Mishra (2021), where LST and precipitation had the lowest importance, our model enhanced their influence by incorporating LST and precipitation as time-series, better reflecting the physical processes involved. NDWI and NDVI were the most important calculated indices, capturing vegetation's influence on surface SM while providing fine spatial details for downscaling (Mohite et al., 2022; Nadeem et al., 2023). It is worth noting that station-specific static features (e.g., soil properties, terrestrial attributes, and locations) were important to the model. As suggested by Rao et al. (2022), the number and location of the samples affect the contribution of static features to the model.

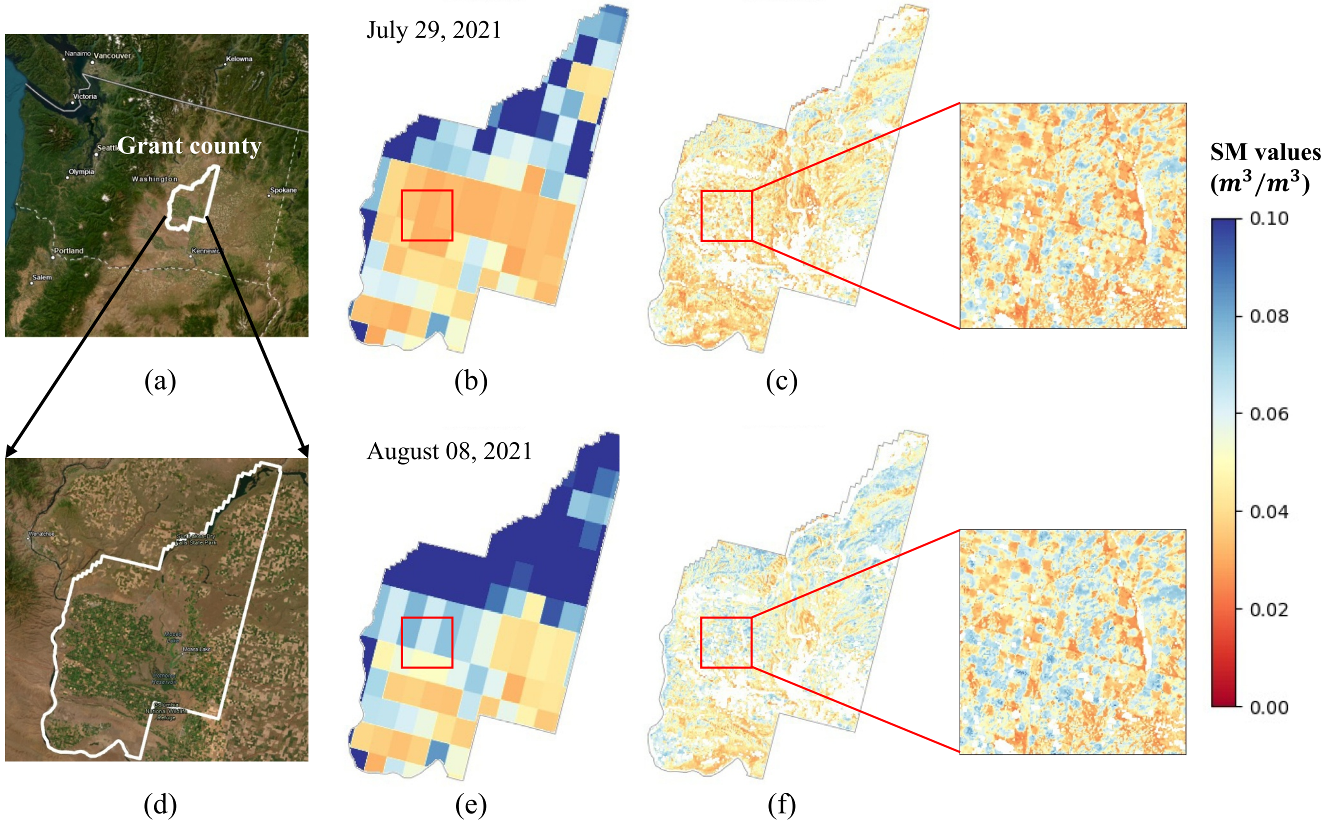

Figure 9. Comparison of SMAP L4 data (9 km) and the MMNet downscaled results. (a) Optical image of the Washington state and location of Grant County, (b) SMAP L4 data (9 km), and (c) MMNet downscaled results (100 m) with a zoomed in view on 19 July 2021 [dry day], (d) optical image zoomed into Grant County, (e) SMAP L4 data (9 km), and (f) MMNet downscaled results (100 m) with a zoomed in view on 08 August 2021 [wet day].

Therefore, a large and widely distributed data set would be crucial for learning the underlying patterns regarding the important static features, improving model's generalizability. Additionally, Sentinel-1-derived indices showed limited importance across all experiments. The feature importance analysis helps understand the physical processes underlying SM dynamics and could provide insight into feature set design in future studies, simplifying data preparation and improving computational efficiency.

5.3. Spatial Variability of Downscaled SM

The proposed MMNet downscales the coarse-resolution SMAP L4 SM product by incorporating diverse high-resolution auxiliary data, such as soil properties and remote sensing images. To assess its ability to capture fine-scale SM spatial variability, we generated two 100-m surface SM maps at Grant County, WA (Figure 9). Grant County is characterized by a semiarid climate with minimal annual precipitation and dry soil conditions (Consulting, 2018). Its landcover is dominated by cropland, followed by grassland and shrubland (Wickham et al., 2021). Figure 9 showed a comparison between optical images, SMAP L4 SM product, and the downscaled results with a zoomed in view of the area of mixed cropland and grassland on 29 July 2021, and 08 August 2021, respectively, representing conditions before and after several precipitation events. The maps were generated for the eight landcover types described in Section 2.1 using a model trained on all available in situ data.

Nine showed that the spatial distribution of the downscaled SM data generally aligns with the original SMAP, exhibiting an overall decrease trend from the northeast to the southwest. After the precipitation events in early August, both maps on August 8 showed an increase in high-value areas compared to July 29, demonstrating the model's ability to capture the rainfall impacts on surface SM. Moreover, the downscaled maps provided more detailed spatial information. In areas with a mix of cropland and grassland (highlighted by red boxes), the 9 km SMAP SM maps fail to capture the spatial heterogeneity, while the downscaled results reflect the variations in landcover types and surface topography, as shown in the zoomed in windows. This demonstrated the potential of the downscaling method to enhance precise agricultural water management at the field scale. Additionally, the SM ranges of the downscaled results slightly differ from the original SMAP due to the integration of auxiliary data. Similar gaps were observed in the time-series result in Figures 5, 6, and 7. While real SM ranges in this area may not be directly measurable, the ability to capture SM dynamics and spatial heterogeneity is more valued in the context of downscaling.

5.4. Limitations and Future Directions

Overall, MMNet holds promise as an effective approach for generating high-resolution, spatially, and temporally continuous SM estimates based on satellite observations and auxiliary data sets. As demonstrated in this study, it achieves better or comparable accuracy to SMAP and SMAP-HB. However, several limitations remain, particularly regarding model generalization.

First, the performance of MMNet decreased in off-site and cross-region settings, particularly in regions with extreme or unique environmental conditions. This highlights the need for a more diverse and representative training set in deep learning methods for SM estimation. Expanding training data coverage or incorporating more in situ stations, particularly in underrepresented regions, can help improve the model generalizability and reduce systematic bias (Entekhabi et al., 2010; Karthikeyan & Mishra, 2021).

Second, while the current input features were included due to their correlation with SM, our feature importance analysis (Section 5.2) suggests that not all features contribute significantly to the final predictions. For instance, Sentinel-1-related features showed limited impact on model performance (Figure 8). Moreover, as suggested by Meyer et al. (2019), geolocation features (e.g., latitude and longitude) can lead to overfitting and constrain the model's spatial generalizability. Removing such variables could enhance model robustness in off-site or cross-region scenarios. Additionally, the SMAP L4 product's assimilation processes may introduce uncertainties associated with vegetation density (Reichle et al., 2019, p. 4). Incorporating additional variables, such as leaf area index (LAI) and vegetation optical depth (VOD) (Konings et al., 2016; W. Li, Zhu, et al., 2022), could be further explored to mitigate the impact of these uncertainties.

Third, as domain shift remains a critical challenge in cross-region scenarios (Section 4.3), exploring transfer learning or domain adaptation strategies (Ma et al., 2024) could strengthen model generalization across domains and improve its applicability in diverse environments. For example, Zhu et al. (2025) have proposed a multiscale domain adaptation method to enhance high-resolution SM retrieval in data-scarce regions. Emerging domain adaptation strategies, such as adversarial training (Ganin et al., 2017), feature alignment (Long et al., 2015), and meta-learning (Tseng et al., 2021) also offer promising directions for developing more transferable SM estimation models.

Finally, given the current limitations in ML-based methods in SM estimation, integrating uncertainty quantification (e.g., Monte Carlo dropout) can help assess the predictive confidence of the model (Gal & Ghahramani, n.d.). This strategy would enhance model interpretability and reliability, providing valuable guidance for downstream decision-making in practical hydrology and agriculture applications (Fang et al., 2020).

6. Conclusion

In this study, we introduced MMNet, a multimodal DL model that integrates remote sensing and weather data to downscale SMAP surface SM. By leveraging complementary information from snapshot and time-series data modalities, MMNet consistently outperformed baseline models in terms of accuracy and transferability. The key findings include the following: (a) MMNet produced reliable SM estimates in the on-site scenario, showing that historical in situ SM measurements can effectively be used to build models to temporally gap fill or extend SM time-series (RQ1). (b) Despite systematic biases, MMNet effectively captured SM dynamics in spatial transfer settings (off-site and cross-region scenarios), enabling SM downscaling for regions with sparse or no observational data (RQ2). (c) Snapshot data enhanced estimation accuracy, and time-series data contributed to stable estimation and improved transferability, underlining the importance of integrating both data modalities for reliable SM estimates (RQ3). These results demonstrate the potential of MMNet to support high-resolution SM monitoring, which is crucial for applications such as precision agriculture, drought assessment, and hydrological modeling.

Disclaimer

The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or U.S. government determination or policy.

Conflict of Interest

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

Data Availability Statement

The MMNet code relevant to this work can be accessed at (Xu, 2025). The in situ data from SCAN and USCRN can be accessed from the International Soil Moisture Network (ISMN) (Dorigo et al., 2021). All used input data sets can be accessed through the Google Earth Engine (GEE) platform (Gorelick et al., 2017). The SMAP Level-4 (L4) Soil Moisture Product can be accessed from R. Reichle et al. (2018). Landsat 8 data can be accessed from Earth Resources Observation and Science (EROS) Center (2020). Sentinel-1 data can be accessed from European Space Agency (2021). Weather variables from gridMET can be accessed from Abatzoglou (2013). MODIS Land Surface Temperature (LST) data can be accessed from Wan et al. (2021). Terrain attributes can be accessed from U.S. Geological Survey (2018). National Land Cover Data set (NLCD) can be accessed from Dewitz (2020, 2021, 2023). Soil properties from POLARIS can be accessed from Chaney et al. (2019).

Acknowledgments
This work was supported by the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) Agriculture and Food Research Initiative: Foundational Program (Grant 2023-67021-40007), the USDA NIFA Agriculture and Food Research Initiative Foundational Program (Award 2022-67021-36468), and the USDA NIFA Hatch Project (Accession 7005141). Jingyi Huang acknowledges the support from the Marvin T. Beatty Professorship in Soil Science.

References

Abatzoglou, J. T. (2013a). Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, *33*(1), 121–131. https://doi.org/10.1002/joc.3413

Abatzoglou, J. T. (2013b). Gridded surface meteorological data (gridMET) [Dataset]. *University of Idaho*. https://www.climatologylab.org/gridmet.html

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

Anshuman, A., & Eldho, T. I. (2022). Entity aware sequence to sequence learning using LSTMs for estimation of groundwater contamination release history and transport parameters. *Journal of Hydrology*, *608*, 127662. https://doi.org/10.1016/j.jhydrol.2022.127662

Batchu, V., Nearing, G., & Gulshan, V. (2023). A deep learning data fusion model using sentinel-1/2, SoilGrids, SMAP, and GLDAS for soil moisture retrieval. *Journal of Hydrometeorology*, *24*(10), 1789–1823. https://doi.org/10.1175/JHM-D-22-0118.1

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Bell, J. E., Palecki, M. A., Baker, C. B., Collins, W. G., Lawrimore, J. H., Leeper, R. D., et al. (2013). U.S. Climate reference network soil moisture and temperature observations. *Journal of Hydrometeorology*, *14*(3), 977–988. https://doi.org/10.1175/JHM-D-12-0146.1

Cammalleri, C., & Vogt, J. (2015). On the role of land surface temperature as proxy of soil moisture status for drought monitoring in europe. *Remote Sensing*, *7*(12), 16849–16864. Article 12. https://doi.org/10.3390/rs71215857

Chaney, N. W., Minasny, B., Herman, J. D., Nauman, T. W., Brungard, C. W., Morgan, C. L. S., et al. (2019). POLARIS soil properties: 30-m probabilistic maps of soil properties over the contiguous United States. *Water Resources Research*, *55*(4), 2916–2938. https://doi.org/10.1029/2018WR022797

Consulting, W. B. (2018). Grant county comprehensive plan.

Cui, Z., Chen, W., & Chen, Y. (2016). Multi-scale convolutional neural Networks for time series classification (arXiv:1603.06995). *arXiv*. https://doi.org/10.48550/arXiv.1603.06995

Dewitz, J. (2020). National Land Cover Database (NLCD) 2016 products (ver. 3.0, November 2023) [Dataset]. https://doi.org/10.5066/P96HHBIE

Dewitz, J. (2021). National Land Cover Database (NLCD) 2019 products [Dataset]. https://doi.org/10.5066/P9KZCM54

Dewitz, J. (2023). National Land Cover Database (NLCD) 2021 products [Dataset]. https://doi.org/10.5066/P9JZ7AO3

Djamai, N., Magagi, R., Goita, K., Merlin, O., Kerr, Y., & Walker, A. (2015). Disaggregation of SMOS soil moisture over the Canadian Prairies. *Remote Sensing of Environment*, *170*, 255–268. https://doi.org/10.1016/j.rse.2015.09.013

Dong, J., Crow, W., Reichle, R., Liu, Q., Lei, F., & Cosh, M. H. (2019). A global assessment of added value in the SMAP level 4 soil moisture product relative to its baseline land surface model. *Geophysical Research Letters*, *46*(12), 6604–6613. https://doi.org/10.1029/2019GL083398

Dorigo, W., Himmelbauer, I., Aberer, D., Schremmer, L., Petrakovic, I., Zappa, L., et al. (2021). The international soil moisture network: Serving earth system science for over a decade. *Hydrology and Earth System Sciences*, *25*(11), 5749–5804. https://doi.org/10.5194/hess-25-5749-2021

Dorigo, W. a., Xaver, A., Vreugdenhil, M., Gruber, A., Hegyiová, 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), vzj2012.0097-21. https://doi.org/10.2136/vzj2012.0097

Earth Resources Observation and Science (EROS) Center. (2020). Landsat 8-9 operational land imager/thermal infrared sensor level-1, collection 2 [Dataset]. *U.S. Geological Survey*. https://doi.org/10.5066/P975CC9B

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* (Vol. 98(5), 704–716). https://doi.org/10.1109/JPROC.2010.2043918

European Space Agency. (2021). Sentinel-1 synthetic aperture radar Ground Range Detected (GRD) imagery [Dataset]. *Google Earth Engine / Copernicus*. Retrieved from https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S1_GRD

Fang, K., Kifer, D., Lawson, K., & Shen, C. (2020). Evaluating the potential and challenges of an uncertainty quantification method for long short-term memory models for soil moisture predictions. *Water Resources Research*, *56*(12), e2020WR028095. https://doi.org/10.1029/2020WR028095

Fang, K., & Shen, C. (2020). Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration Kernel. *Journal of Hydrometeorology*, *21*(3), 399–413. https://doi.org/10.1175/JHM-D-19-0169.1

Gal, Y., & Ghahramani, Z. (n.d.). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning (Vol. *10*).

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2017). Domain-adversarial training of neural networks. In G. Csurka (Ed.), *Domain adaptation in computer vision applications* (pp. 189–209). Springer International Publishing. Retrieved from http://link.springer.com/10.1007/978-3-319-58347-1_10

Gaur, N., & Mohanty, B. P. (2016). Land-surface controls on near-surface soil moisture dynamics: Traversing remote sensing footprints. *Water Resources Research*, *52*(8), 6365–6385. https://doi.org/10.1002/2015WR018095

Geological Survey, U. S. (2018). 3D elevation Program (3DEP) digital elevation models [Dataset]. https://www.usgs.gov/3d-elevation-program

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

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. Article 1. https://doi.org/10.1038/s41597-023-02011-7

Huang, J., Desai, A. R., Zhu, J., Hartemink, A. E., Stoy, P. C., Loheide, S. P., et al. (2020). Retrieving heterogeneous surface soil moisture at 100 m across the globe via fusion of remote sensing and land surface parameters. *Frontiers in Water*, *2*. https://doi.org/10.3389/frwa.2020.578367

Ji, Y., Sun, W., Wang, Y., Lv, Z., Yang, G., Zhan, Y., & Li, C. (2024). Domain adaptive and interactive differential attention network for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–16. https://doi.org/10.1109/TGRS.2024.3382116

Karthikeyan, L., & Mishra, A. K. (2021). Multi-layer high-resolution soil moisture estimation using machine learning over the United States. *Remote Sensing of Environment*, *266*, 112706. https://doi.org/10.1016/j.rse.2021.112706

Kolassa, J., Reichle, R. H., Liu, Q., Alemohammad, S. H., Gentine, P., Aida, K., et al. (2018). Estimating surface soil moisture from SMAP observations using a Neural Network technique. *Remote Sensing of Environment*, *204*, 43–59. https://doi.org/10.1016/j.rse.2017.10.045

Komma, J., Blöschl, G., & Reszler, C. (2008). Soil moisture updating by Ensemble Kalman Filtering in real-time flood forecasting. *Journal of Hydrology*, *357*(3), 228–242. https://doi.org/10.1016/j.jhydrol.2008.05.020

Konings, A. G., Piles, M., Rötzer, K., McColl, K. A., Chan, S. K., & Entekhabi, D. (2016). Vegetation optical depth and scattering albedo retrieval using time series of dual-polarized L-band radiometer observations. *Remote Sensing of Environment*, *172*, 178–189. https://doi.org/10.1016/j.rse.2015.11.009

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. Article 7553. https://doi.org/10.1038/nature14539

Li, B., Wang, L., Kaseke, K. F., Li, L., & Seely, M. K. (2016). The impact of rainfall on soil moisture dynamics in a foggy desert. *PLoS One*, *11*(10), e0164982. https://doi.org/10.1371/journal.pone.0164982

Li, Q., Wang, Z., Shangguan, W., Li, L., Yao, Y., & Yu, F. (2021). Improved daily SMAP satellite soil moisture prediction over China using deep learning model with transfer learning. *Journal of Hydrology*, *600*, 126698. https://doi.org/10.1016/j.jhydrol.2021.126698

Li, Q., Zhu, Y., Shangguan, W., Wang, X., Li, L., & Yu, F. (2022). An attention-aware LSTM model for soil moisture and soil temperature prediction. *Geoderma*, *409*, 115651. https://doi.org/10.1016/j.geoderma.2021.115651

Li, W., Migliavacca, M., Forkel, M., Denissen, J. M. C., Reichstein, M., Yang, H., et al. (2022). Widespread increasing vegetation sensitivity to soil moisture. *Nature Communications*, *13*(1), 3959. https://doi.org/10.1038/s41467-022-31667-9

Liu, J., Rahmani, F., Lawson, K., & Shen, C. (2022). A multiscale deep learning model for soil moisture integrating satellite and in situ data. *Geophysical Research Letters*, *49*(7), e2021GL096847. https://doi.org/10.1029/2021GL096847

Liu, J., Shen, Z., He, Y., Zhang, X., Xu, R., Yu, H., & Cui, P. (2023). Towards out-of-distribution generalization: A Survey (arXiv:2108.13624). *arXiv*. http://arxiv.org/abs/2108.13624

Long, D., Bai, L., Yan, L., Zhang, C., Yang, W., Lei, H., et al. (2019). Generation of spatially complete and daily continuous surface soil moisture of high spatial resolution. *Remote Sensing of Environment*, *233*, 111364. https://doi.org/10.1016/j.rse.2019.111364

Long, M., Cao, Y., Wang, J., & Jordan, M. I. (2015). *Learning transferable features with deep adaptation networks* (Vol. 1, pp. 97–105). Scopus.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv*. arXiv:1705.07874). https://doi.org/10.48550/arXiv.1705.07874

Ma, Y., Chen, S., Ermon, S., & Lobell, D. B. (2024). Transfer learning in environmental remote sensing. *Remote Sensing of Environment*, *301*, 113924. https://doi.org/10.1016/j.rse.2023.113924

Martínez-Fernández, J., González-Zamora, A., Sánchez, N., Gumuzzio, A., & Herrero-Jiménez, C. M. (2016). Satellite soil moisture for agricultural drought monitoring: Assessment of the SMOS derived Soil Water Deficit Index. *Remote Sensing of Environment*, *177*, 277–286. https://doi.org/10.1016/j.rse.2016.02.064

Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications—Moving from data reproduction to spatial prediction. *Ecological Modelling*, *411*, 108815. https://doi.org/10.1016/j.ecolmodel.2019.108815

Mohite, J. D., Sawant, S. A., Pandit, A., & Pappula, S. (2022). Spatial downscaling of SMAP soil moisture using the MODIS and SRTM observations. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLIII-B3–2022*, 933–938. XXIV ISPRS Congress “Imaging today, foreseeing tomorrow”, Commission III - 2022 edition, 6–11 June 2022, Nice, France. https://doi.org/10.5194/isprs-archives-XLIII-B3-2022-933-2022

Mondal, S., & Mishra, A. (2024). Quantifying the precipitation, evapotranspiration, and soil moisture network’s interaction over global land surface hydrological cycle. *Water Resources Research*, *60*(2), e2023WR034861. https://doi.org/10.1029/2023WR034861

Nadeem, A. A., Zha, Y., Shi, L., Ali, S., Wang, X., Zafar, Z., et al. (2023). Spatial downscaling and gap-filling of SMAP soil moisture to high resolution using MODIS surface variables and machine learning approaches over ShanDian River Basin, China. *Remote Sensing*, *15*(3), 812. https://doi.org/10.3390/rs15030812

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th international conference on international conference on machine learning* (pp. 689–696).

NOAA and NIDIS. (2020). Regional climate quarterly midwest-September 2019. *National Integrated Drought Information System*. Retrieved from https://www.drought.gov/sites/default/files/2020-10/Midwest%20Summer%202019.pdf

Nyborg, J., Pelletier, C., Lefèvre, S., & Assent, I. (2022). TimeMatch: Unsupervised cross-region adaptation by temporal shift estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, *188*, 301–313. https://doi.org/10.1016/j.isprsjprs.2022.04.018

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

Peng, J., Loew, A., Merlin, O., & Verhoest, N. E. C. (2017). A review of spatial downscaling of satellite remotely sensed soil moisture. *Reviews of Geophysics*, *55*(2), 341–366. https://doi.org/10.1002/2016RG000543

Rao, P., Wang, Y., Wang, F., Liu, Y., Wang, X., & Wang, Z. (2022). Daily soil moisture mapping at 1km resolution based on SMAP data for desertification areas in northern China. *Earth System Science Data*, *14*(7), 3053–3073. https://doi.org/10.5194/essd-14-3053-2022

Reichle, M. R., De Lannoy, G., Koster, R. D., Crow, W. T., Kimball, J. S., & Liu, Q. (2022). *SMAP L4 global 3-hourly 9 km EASE-grid surface and root zone soil moisture analysis update, version 7*. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/LWJ6TF5SZRG3

Reichle, R., De Lannoy, G., Koster, R., Crow, W., Kimball, J., & Liu, Q. (2018). SMAP L4 global 3-hourly 9 km EASE-grid surface and root zone soil moisture analysis update, version 4 [Dataset]. *NASA National Snow and Ice Data Center Distributed Active Archive Center*. https://doi.org/10.5067/60HB8VIP2T8W

Reichle, R. H., Liu, Q., Koster, R. D., Crow, W. T., De Lannoy, G. J. M., Kimball, J. S., et al. (2019). Version 4 of the SMAP level-4 soil moisture algorithm and data product. *Journal of Advances in Modeling Earth Systems*, *11*(10), 3106–3130. https://doi.org/10.1029/2019MS001729

Sandholt, I., Rasmussen, K., & Andersen, J. (2002). A simple interpretation of the surface temperature/vegetation index space for assessment of surface moisture status. *Remote Sensing of Environment*, *79*(2), 213–224. https://doi.org/10.1016/S0034-4257(01)00274-7

Schaefer, G. L., Cosh, M. H., & Jackson, T. J. (2007). The USDA natural resources conservation service Soil Climate Analysis Network (SCAN). *Journal of Atmospheric and Oceanic Technology*, *24*(12), 2073–2077. https://doi.org/10.1175/2007JTECHA930.1

Schmidt, T., Schrön, M., Li, Z., Francke, T., Zacharias, S., Hildebrandt, A., & Peng, J. (2024). Comprehensive quality assessment of satellite- and model-based soil moisture products against the COSMOS network in Germany. *Remote Sensing of Environment*, *301*, 113930. https://doi.org/10.1016/j.rse.2023.113930

Tian, J., Lu, H., Yang, K., Qin, J., Zhao, L., Jiang, Y., et al. (2023). Improving surface soil moisture estimation through assimilating satellite land surface temperature with a linear SM-LST relationship. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *16*, 7777–7790. https://doi.org/10.1109/JSTARS.2023.3305888

Tseng, G., Kerner, H., Nakalembe, C., & Becker-Reshef, I. (2021). Learning to predict crop type from heterogeneous sparse labels using meta-learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1111–1120). https://doi.org/10.1109/CVPRW53098.2021.00122

U.S. EPA. (2015). Level III and IV Ecoregions of the continental United States [data and tools]. Retrieved from https://www.epa.gov/eco-research/level-iii-and-iv-ecoregions-continental-united-states

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wan, Z., Hook, S., & Hulley, G. (2021). MODIS/Terra land surface temperature/emissivity daily L3 global 1km SIN grid V061 [Dataset]. *NASA EOSDIS Land Processes Distributed Active Archive Center*. https://doi.org/10.5067/MODIS/MOD11A1.061

Wang, Z., Yan, W., & Oates, T. (2016). Time series Classification from Scratch with deep neural networks: A strong baseline (arXiv:1611.06455). *arXiv*. http://arxiv.org/abs/1611.06455

Wei, Z., Meng, Y., Zhang, W., Peng, J., & Meng, L. (2019). Downscaling SMAP soil moisture estimation with gradient boosting decision tree regression over the Tibetan Plateau. *Remote Sensing of Environment*, *225*, 30–44. https://doi.org/10.1016/j.rse.2019.02.022

Wickham, J., Stehman, S. V., Sorenson, D. G., Gass, L., & Dewitz, J. A. (2021). Thematic accuracy assessment of the NLCD 2016 land cover for the conterminous United States. *Remote Sensing of Environment*, *257*, 112357. https://doi.org/10.1016/j.rse.2021.112357

Xu, M., Yao, N., Yang, H., Xu, J., Hu, A., Gustavo Goncalves de Goncalves, L., & Liu, G. (2022). Downscaling SMAP soil moisture using a wide and deep learning method over the Continental United States. *Journal of Hydrology*, *609*, 127784. https://doi.org/10.1016/j.jhydrol.2022.127784

Xu, Y. (2025). MMNet: Multi-Modal network for soil moisture downscaling [Software]. *Computer Software*. https://doi.org/10.5281/zenodo.15319934

Yu, J., Zhang, X., Xu, L., Dong, J., & Zhangzhong, L. (2021). A hybrid CNN-GRU model for predicting soil moisture in maize root zone. *Agricultural Water Management*, *245*, 106649. https://doi.org/10.1016/j.agwat.2020.106649

Zhao, H., Li, J., Yuan, Q., Lin, L., Yue, L., & Xu, H. (2022). Downscaling of soil moisture products using deep learning: Comparison and analysis on Tibetan Plateau. *Journal of Hydrology*, *607*, 127570. https://doi.org/10.1016/j.jhydrol.2022.127570

Zhu, L., Cai, Q., Jin, J., Yuan, S., Shen, X., & Walker, J. P. (2025). Multi-Scale domain adaptation for high-resolution soil moisture retrieval from synthetic aperture radar in data-scarce regions. *Journal of Hydrology*, *657*, 133073. https://doi.org/10.1016/j.jhydrol.2025.133073

Zhu, L., Dai, J., Liu, Y., Yuan, S., Qin, T., & Walker, J. P. (2024). A cross-resolution transfer learning approach for soil moisture retrieval from Sentinel-1 using limited training samples. *Remote Sensing of Environment*, *301*, 113944. https://doi.org/10.1016/j.rse.2023.113944