JGR Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2025JH001039

Weak Physics-Guided Multi-Agent Learning for Surface to Subsurface Moisture Estimation Across Diverse Climate and Soil Conditions

Abhilash Singh[1], **Vidhi Singh**[2], **and Kumar Gaurav**[2]

[1]School of Mathematics, Faculty of Engineering and Physical Sciences, University of Leeds, Leeds, UK, [2]Department of Earth and Environmental Sciences, Fluvial Geomorphology and Remote Sensing Laboratory, Indian Institute of Science Education and Research Bhopal, Bhopal, India

Key Points:

- Multi-agent diffusion model estimates subsurface soil moisture via weak physics constraints, needing no site-specific soil parameters
- Model tested on 20 global ISMN sites (hourly) and three Zambian sites (10-min) with up to 12 years of data
- Results show high accuracy, robust uncertainty estimates, and scalability across climates and depths

Supporting Information:

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

Correspondence to:

A. Singh and K. Gaurav,
a.singh4@leeds.ac.uk;
abhilash.iiserb@gmail.com;
kgaurav@iiserb.ac.in

Citation:

Singh, A., Singh, V., & Gaurav, K. (2026). Weak physics-guided Multi-Agent learning for surface to subsurface Moisture estimation across diverse climate and Soil conditions. *Journal of Geophysical Research: Machine Learning and Computation*, *3*, e2025JH001039. https://doi.org/10.1029/2025JH001039

Received 26 SEP 2025
Accepted 3 JAN 2026

Author Contributions:

Conceptualization: Abhilash Singh, Kumar Gaurav
Data curation: Vidhi Singh
Formal analysis: Vidhi Singh
Funding acquisition: Kumar Gaurav
Investigation: Abhilash Singh, Vidhi Singh, Kumar Gaurav
Methodology: Abhilash Singh, Vidhi Singh, Kumar Gaurav
Software: Abhilash Singh
Supervision: Kumar Gaurav
Validation: Abhilash Singh

Abstract Estimating subsurface soil moisture remains challenging due to limited in situ observations and the complexity of soil water dynamics. Although surface soil moisture can be retrieved from satellites with high accuracy, deeper layers are not directly observable. Traditional physics-based models that predict subsurface soil moisture require site-specific hydraulic properties of the soils. This limits their large-scale applicability. Alternative data-driven machine learning models for subsurface soil moisture estimation generally lack physical interpretability. To address the limitation of physical and machine learning models, we propose a weakly physics-constrained, multi-agent diffusion model for subsurface soil moisture estimation. The model employs lightweight physical regularization (flux smoothness and feasible-range constraints) that guide predictions without enforcing strict parameterization, while a multi-agent structure allows specialization across dry, intermediate, and wet soil regimes. This framework balances predictive flexibility with hydrological consistency and provides uncertainty quantification through stochastic diffusion sampling. The model is evaluated using globally distributed in situ data sets from 20 different sites within the International Soil Moisture Network (ISMN) and from three sites in Zambia, Africa. Soil moisture observations from ISMN are available at hourly intervals, while measurements from the Zambian stations are recorded every 10-min. The results show a strong agreement between the modeled and observed soil moisture at multiple depths (10, 20, and 40 cm), with median values of R^2 exceeding 0.91 and nRMSE of 0.37 at 10 cm and remaining robust at deeper layers, although performance decreases with depth as expected. The model outperforms several benchmark machine learning algorithms, particularly at greater depths, and exhibits stability under stochastic initialization and input perturbations.

Plain Language Summary Subsurface soil moisture plays a critical role in agriculture, water resources, and climate processes, but is difficult to measure directly. Satellites can estimate surface moisture accurately, but deeper soil layers, where plants access water and where long-term water storage occurs, are much harder to measure. This study proposes a new machine learning model that estimates subsurface soil moisture using information from surface measurements. Our approach combines the flexibility of data-driven learning with simple physics-based rules to ensure realistic results, without requiring detailed soil properties that are often unavailable. The model is also designed with multiple components that each specialize in different soil wetness conditions, making it more reliable across diverse climates. We test it on global soil moisture records from 20 sites with hourly data and three additional sites in Zambia with high-frequency (10-min) data. The model shows strong agreement across all sites and depths, providing accurate subsurface soil moisture. This work demonstrates how combining physics and machine learning can deliver scalable and practical solutions for monitoring water resources in regions where in situ measurements are sparse.

1. Introduction

Soil moisture plays an important role in the hydrological cycle, directly affecting the interaction between land and atmosphere, influencing vegetation health and driving agricultural productivity (Li et al., 2024; Rahmati et al., 2024). It strongly influences the development of droughts and heatwaves (Malik & Mishra, 2025; Tripathy et al., 2025). Although satellites (such as SMAP, SMOS, Sentinel, and recently launched NISAR) can estimate surface soil moisture, measuring moisture in deeper soil layers remains challenging (Agyeman et al., 2025; Ahmad et al., 2022; A. Singh & Gaurav, 2024; Dinesh et al., 2024; Jing et al., 2024; Lawston et al., 2017; A. Singh

Visualization: Abhilash Singh, Kumar Gaurav
Writing – original draft: Abhilash Singh, Vidhi Singh, Kumar Gaurav
Writing – review & editing: Abhilash Singh, Kumar Gaurav

et al., 2025). To measure subsurface soil moisture, sensors must be installed at different depths beneath the ground, which is expensive and difficult to install and manage, particularly in remote locations. To address this challenge, several studies estimate deeper soil moisture directly from surface soil moisture observations (Tian et al., 2019; A. Singh, Singh, & Gaurav, 2025b).

Traditionally, researchers have estimated subsurface soil moisture by using various empirical, semi-empirical, and physics-based hydrological models that simulate unsaturated flow (Draper & Reichle, 2015; Verma & Nema, 2022; Zhang et al., 2024). These models rely on site-specific soil properties and hydraulic parameters. However, such detailed parameters (e.g., soil porosity, permeability, retention curves, etc.) are either unavailable or not accessible at global scale. As a result, purely physics-based approaches can be difficult to apply at regional or global scales.

In recent years, machine learning has emerged as a powerful alternative for predicting subsurface moisture (A. Singh, Singh, & Gaurav, 2025a; V. Singh et al., 2025). These data-driven models learn the governing relationships between surface observations and subsurface conditions. They have shown good performance even without detailed information about the soil parameters (Liu et al., 2024). However, purely data-driven methods are often criticized for their poor generalization capabilities (Rudin, 2019). A model trained on one region or set of conditions might struggle when faced with unseen climates, soil types, or land-cover conditions.

To address this, physics-informed machine learning (PIML) frameworks have gained popularity (Chavoshi et al., 2024). These hybrid approaches incorporate the governing physical laws (e.g., Richards' equation) into the learning process of the models. PIML models usually require careful calibration for each site, meaning the model needs accurate local soil parameters to work well. This dependence on local data (saturated hydraulic conductivity or specific soil retention properties) can limit their applicability at large scales. In practice, the need for site-specific tuning can offset the benefits of physics guidance, because those parameters might not be available everywhere or transferable from one site to another.

This study proposes a new framework that avoids these limitations: a weakly physics-constrained, multi-agent diffusion model for subsurface soil moisture estimation. Instead of enforcing strict site-specific physical parameters, we introduce lightweight physical constraints as gentle guidance. For example, our model is encouraged to maintain smooth changes in water flux and keep soil moisture values within feasible ranges, but these are soft constraints rather than hard rules. This design preserves the flexibility and learning capacity of a data-driven model while still respects the key physical principles of hydrology. Furthermore, the model uses a multi-agent structure, meaning different agents (or components of the model) specialize in different soil moisture regimes: dry, intermediate, and wet conditions. This specialization makes the model more robust across a variety of conditions, as each agent becomes an expert in its regime. Further, by combining a stochastic diffusion modeling approach with these weak physical regularizations, we maintain a balance between accuracy, stability across different sites, and physical consistency in the predictions. To evaluate this framework, we test it on globally distributed in situ soil moisture data sets (hourly) from 20 International Soil Moisture Network (ISMN) and three additional high-frequency (10 min) sites in Zambia. The data time series at different stations vary from 7 months to 12 years of records. This diverse evaluation setup ensures that the proposed model is assessed across a broad range of climates, soil textures, and temporal scales. To the best of our knowledge, no prior study has integrated a multi-agent architecture, diffusion modeling, and weak physical constraints (either individually or in combination) for subsurface soil moisture estimation. The key research questions this study addresses are as follows;

- Can adding weak physical constraints to a data-driven model improve the subsurface soil moisture predictions as compared to purely data-driven models?
- How robust is the proposed model across different climates, soil textures, and land-cover types as represented in the ISMN data set?
- To what extent does the multi-agent framework enhance predictive performance and uncertainty quantification across various depths (10, 20, and 40 cm) and temporal resolutions?

This study contributes to developing scalable yet physically consistent machine learning approaches for subsurface hydrology. We demonstrate that it is possible to achieve high predictive accuracy without heavy reliance on site-specific parameters, making these models more useful in different environments.

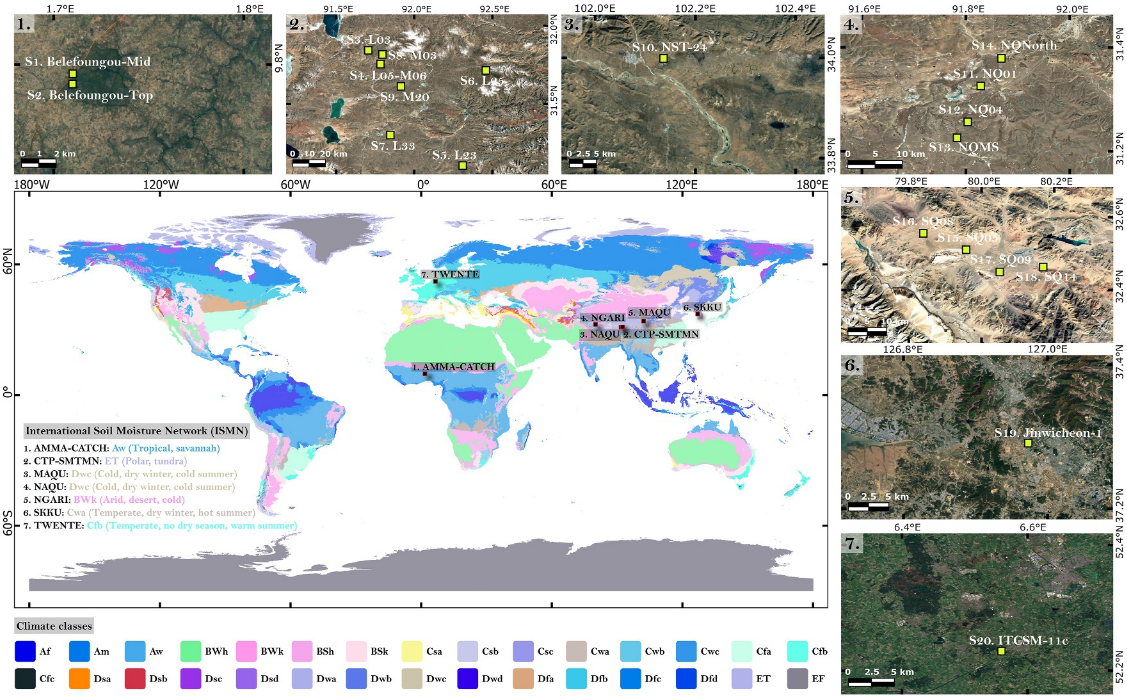


Figure 1. Geographical representation of International Soil Moisture Network (ISMN) locations classified by Köpen Geiger climate classes (1–7). The main map illustrates the worldwide distribution, while the top and right panels present zoomed-in satellite images showing the spatial configuration of stations within the respective networks (S1–S20), indicated by green squares (Credit © Google Maps Platform).

2. Material and Methods

2.1. In Situ Observations

We acquired soil moisture time series data from the ISMN repository, assimilating seven monitoring networks distributed across diverse hydroclimatic regions (Dorigo et al., 2011, 2021; Gruber et al., 2013). These include AMMA-CATCH in Benin, West Africa; CTP-SMTMN, MAQU, NAQU, and NGARI in China; SKKU in South Korea; and TWENTE in the Netherlands (Dente et al., 2012; Lebel et al., 2009; Nguyen et al., 2017; Su et al., 2011; van der Velde et al., 2023; Yang et al., 2013). From these networks, we selected 20 stations: two from AMMA-CATCH (S1: Belefoungou-Mid, S2: Belefoungou-Top), seven from CTP-SMTMN (S3: L03, S4: L05-M06, S5: L23, S6: L25, S7: L33, S8: M03, S9: M20), one from MAQU (S10: NST-24), four each from NAQU (S11: NQ01, S12: NQ04, S13: NQMS, S14: NQNorth) and NGARI (S15: SQ05, S16: SQ08, S17: SQ09, S18: SQ14), one from SKKU (S19: Jinwicheon-1), and one from TWENTE (S20: ITCSM-11c). These sites represent six distinct Köppen Geiger climate regimes spanning all major climate groups that is, tropical (Aw), dry (BWk), temperate (Cwa, Cfb), continental (Dwc), and polar (ET) (Beck et al., 2023) as shown in Figure 1. There exists land cover heterogeneity across these sites which comprise of tree cover, cropland, grassland, barren terrain, arid/semi-arid transitional biomes, and sparsely vegetated surfaces. Soil properties also vary substantially, with clay ranging from 2.2% to 23%, sand from 35% to 88%, silt from 8% to 44%, and organic carbon from 0.6% to 5.8%. Measurements recorded at 5, 10, 20, and 40 cm soil depths, each with an hourly resolution are integrated in this study for all the sites. Their observation period extends from as short as 7 months to as long as 12 years. Soil moisture probes like TDR (Time Domain Reflectometry) and FDR (Frequency Domain Reflectometry) types are deployed at these sites where only AMMA-CATCH stations use TDR sensors (CS616) while other stations record values from FDR sensors like ECH2O EC-TM, EC-TM, and 5TM (details in Table 1). To further evaluate the model under higher-frequency observations, we employed in situ soil moisture data sampled at 10-min intervals from three stations in Zambia: Bbondo, Kalipululira, and Margaret. These records, obtained through the Zambia Meteorological Department (ZMD), provide continuous coverage over 4–5 years. Kalipululira is located within a hot semi-arid steppe climate (BSh), whereas Bbondo and Margaret fall under a temperate, dry-winter, hot-summer regime (Cwa).

Table 1
Overview of ISMN Station Attributes for 20 Sites (S1–S20), Including Their Geographic Location, Land Cover Classification, Soil Moisture Sensor Type, Climatic Regime, Data Availability Period, and Soil Properties

Attribute	S1. Belefoungou-Mid	S2. Belefoungou-Top	S3. L03	S4. L05-M06	S5. L23	S6. L25	S7. L33	S8. M03	S9. M20	S10. NST-24
Network	AMMA-CATCH		CTP-SMTMN							MAQU
Country	Benin		China							China
Land cover	Tree cover		Grassland							Cropland
Sensor	CS616	CS616	EC-TM	5TM	EC-TM	EC-TM	EC-TM	5TM	5TM	ECH20/ EC-TM
Climate	Aw	Aw	ET	ET	ET	ET	Dwc	ET	ET	Dwc
Data start	2006-01-01	2006-03-07	2010-08-01	2010-08-01	2010-08-01	2010-08-01	2010-08-01	2011-07-01	2011-07-04	2014-10-21
Data end	2017-11-10	2017-12-31	2016-09-19	2016-09-19	2016-09-19	2013-01-17	2016-07-01	2016-09-19	2016-09-12	2019-06-01
Clay (%)	22	22	6	6	22	17	10	22	6	23
Sand (%)	49	49	56	56	39	37	74	39	56	38
Silt (%)	29	29	38	38	39	46	16	39	38	39
Organic carbon (%)	1.01	1.01	1.41	1.41	1.65	3.14	0.70	1.65	1.41	1.95

Attribute	S11. NQ01	S12. NQ04	S13. NQMS	S14. NQNorth	S15. SQ05	S16. SQ08	S17. SQ09	S18. SQ14	S19. Jinwicheon-1	S20. ITCSM-11c
Network	NAQU				NGARI				SKKU	TWENTE
Country	China				China				South Korea	Netherlands
Land cover	Grassland				Bare areas	Desert steppe	Sparse vegetation (tree)	Desert steppe	Sparse vegetation (tree)	Grassland
Sensor	5TM	5TM	5TM	5TM	5TM	5TM	5TM	5TM	5TM	5TM
Climate	Dwc	Dwc	Dwc	Dwc	BWk	BWk	BWk	BWk	Cwa	Cfb
Data start	2016-08-15	2016-08-16	2010-07-28	2011-08-22	2010-07-22	2010-07-22	2010-07-23	2010-07-23	2014-05-08	2015-10-09
Data end	2019-09-12	2019-09-12	2019-09-12	2019-09-11	2014-05-21	2019-09-09	2014-08-05	2019-09-09	2014-11-19	2016-06-29
Clay (%)	2.3	5.06	23	7.03	19	2.56	21	4.45	23	4
Sand (%)	88.7	76.04	35	76.98	37	87.56	36	86.92	37	86
Silt (%)	8.99	18.9	42	15.99	44	9.88	43	8.64	40	10
Organic carbon (%)	0.63	5.54	1.26	5.82	2.13	1.08	0.65	0.72	1.07	1.97

Note. Each site provides soil moisture records at hourly resolution for depths of 5, 10, 20, and 40 cm. The Köpen Geiger climate categories for the sites are: S1–S2 (Tropical, savannah), S3–S9 (Polar, tundra), S10–S14 (Cold, dry winter, cold summer), S15–S18 (Arid, desert, cold), S19 (Temperate, dry winter, hot summer), and S20 (Temperate, no dry season, warm summer).

Overall, these data sets, characterized by high temporal resolution, multi-depth measurements, and contrasting climate settings and soil type, enables systematic investigation of soil moisture variability from the surface to the subsurface. It provides a robust foundation for evaluating model performance under a wide range of climatic gradients and ecological conditions facilitating a reliable assessment of the model's ability to generalize across regions.

2.2. Model Development

We modeled subsurface soil moisture at chronological time t as shown in Equation 1:

$$\mathbf{y}_t = [\mathrm{SM}_{10,t},\ \mathrm{SM}_{20,t},\ \mathrm{SM}_{40,t}]^\top \in \mathbb{R}^3, \tag{1}$$

conditioned on surface moisture at 5 cm and a seasonal descriptor. For each station, depth-specific files (5, 10, 20, 40 cm) are parsed and merged on a common timestamp; rows with any missing values are removed to avoid implicit imputation. The merged series is split chronologically so that the earliest 70% of rows form the training set and the remaining 30% form the test set, which prevents temporal leakage. Two independent min-max transforms are fitted on the training set only. For the 5 cm driver, $x_t = \mathrm{SM}_{5,t}$, we define the normalized value according to Equation 2:

$$\tilde{x}_t = \frac{x_t - x_{\min}}{x_{\max} - x_{\min}} \in [0,1], \tag{2}$$

and for the three-dimensional target we apply a component-wise transform $\tilde{\mathbf{y}}_t = \mathrm{MM}_y(\mathbf{y}_t) \in [0,1]^3$ using per-column training minima and ranges. Seasonality is encoded with the month-based trigonometric pair, yielding the conditioning vector as shown in Equation 3:

$$\mathbf{c}_t = [\tilde{x}_t,\ \sin(2\pi m_t/12),\ \cos(2\pi m_t/12)]^\top \in \mathbb{R}^3, \tag{3}$$

where $m_t \in \{1, \dots, 12\}$ is the calendar month. All learning and sampling are performed in the normalized target space.

To accommodate regime-dependent vadose-zone behavior, we train three specialist agents (Dry, Mid, Wet) by partitioning the training set with respect to $\tilde{x}_t$ at the (0.33, 0.66) quantiles. Writing these breakpoints as $q_{0.33}$ and $q_{0.66}$, the Dry subset uses samples with $\tilde{x}_t \le q_{0.33}$, the Mid subset uses $q_{0.33} < \tilde{x}_t < q_{0.66}$, and the Wet subset uses $\tilde{x}_t \ge q_{0.66}$. If any subset becomes too small to support stable mini-batches, that agent falls back to the full training set to keep optimization well-posed without discarding specialization. This design encourages each agent to learn a mapping tailored to its moisture regime rather than forcing a single network to compromise across disparate behaviors (Amirkhani & Barshooi, 2022).

Each agent implements a conditional diffusion model in the normalized target space (Ho et al., 2020). The forward (noising) process follows a denoising diffusion probabilistic model (DDPM)-style schedule with T_{diff} discrete steps and a linear variance schedule β_τ. Defining $\alpha_\tau = 1 - \beta_\tau$ and the cumulative product $\bar{\alpha}_\tau = \prod_{k=1}^{\tau} \alpha_k$, a clean target $\mathbf{x}_0 = \tilde{\mathbf{y}}_t$ (constructed from sample at chronological index t) is corrupted at diffusion step τ according to Equation 4:

$$\mathbf{x}_\tau = \sqrt{\bar{\alpha}_\tau}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_\tau}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_3). \tag{4}$$

The reverse model $\boldsymbol{\epsilon}_\theta(\mathbf{x}_\tau, \tau, \mathbf{c}_t)$ is a small multilayer perceptron that predicts the noise given the corrupted state, a scalar diffusion-step embedding $\tau/1000$, and the conditioning vector for the chronological sample. Training samples a random $\tau \in \{0, \dots, T_{\mathrm{diff}} - 1\}$ for each mini-batch, synthesizes $\mathbf{x}_\tau$ by injecting Gaussian noise, and minimizes the mean-squared error between the predicted and true noise, as shown in Equation 5:

$$\mathcal{L}_{\mathrm{data}} = \mathbb{E}\big[\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_\tau, \tau, \mathbf{c}_t) - \boldsymbol{\epsilon}\|_2^2\big]. \tag{5}$$

We train all agents with the Adam optimizer using a constant learning rate (lr) and mini-batches of moderate size. To stabilize optimization of the diffusion objective, we apply gradient-norm clipping at a fixed threshold. Training proceeds for a fixed number of epochs (E) with the same optimization settings across agents to ensure comparability; the learning rate, batch size, and clipping threshold were selected from preliminary trials to balance stability and wall-clock time. The model parameter values are listed in Table 2, the pseudocode is presented in Algorithm 1, and the complete methodological flow is illustrated in Figure 2.

Table 2
Model, Training, and Inference Hyperparameters Used in This Study

Symbol/Name	Value
Train/Test split	70%/30% (chronological)
Regime thresholds	$q_{0.33}$, $q_{0.66}$ (on train $\tilde{x}_t$)
Denoiser architecture	MLP: 2 × 128 ReLU + linear head $\left(\mathbb{R}^3 \to \mathbb{R}^3\right)$
Optimizer	Adam (lr 10^{-4}, grad clip 1.0)
Epochs/Batch size	400/128
Diffusion steps T_{diff}	400
Variance schedule β_τ	Linear $\left[10^{-4}, 2 \times 10^{-2}\right]$;
Time embedding	$\tau/1000$ (scalar)
x_0 recon. ε	10^{-8}
Physics weight λ_{flux}	0.2
Physics weight λ_{bounds}	0.05
Gating $\mathbf{w}_t$	normalized $[w_{\text{dry}}, w_{\text{mid}}, w_{\text{wet}}]$, $\delta = 10^{-6}$
Sampling policy	Top-1: $a^\star(t) = \text{argmax}_a w_a(t)$
Monte-Carlo mean samples M	5
Ensemble size N	$\{5, 10, \ldots, 100\}$
Consensus corrector steps S	30
Corrector step size	0.05
Random seed	0–29

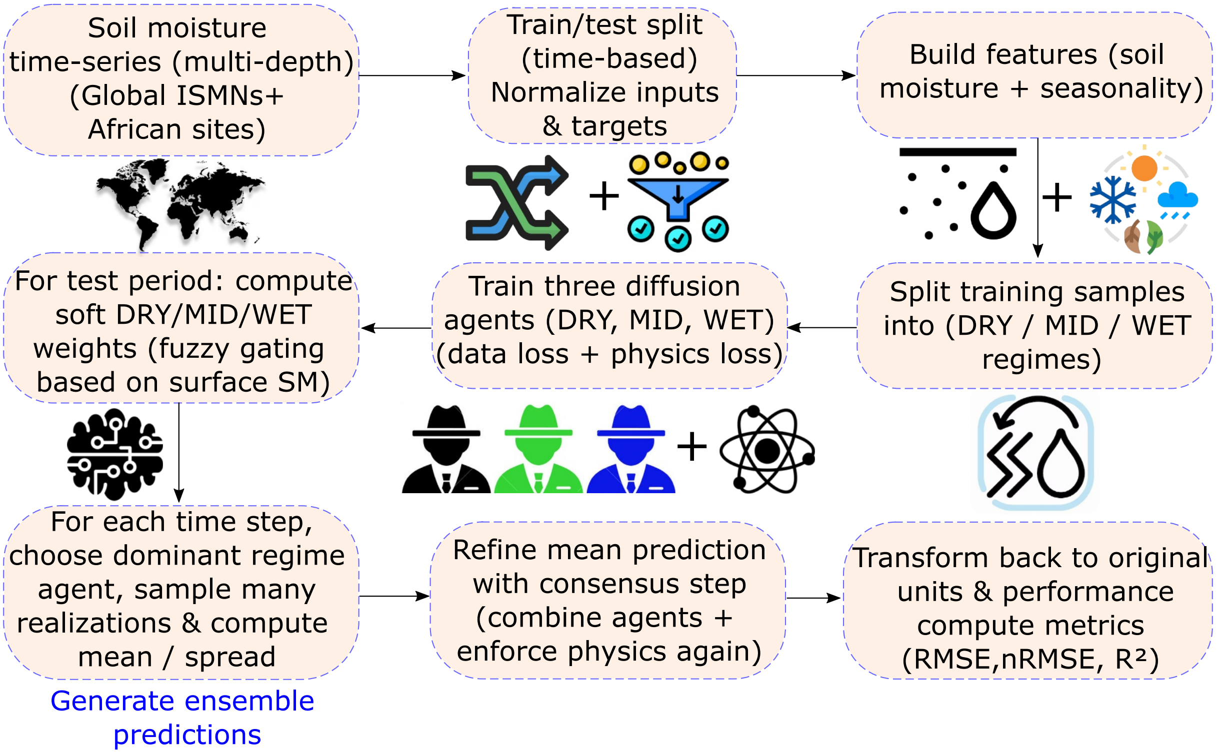


Figure 2. Flowchart showing the multi-agent diffusion pipeline: data loading, preprocessing, regime-based agent training, fuzzy gating, ensemble prediction, physics-guided refinement, and final evaluation with uncertainty outputs.

Algorithm 1. Pseudo code for multi-agent physics-informed diffusion model

```
Require: Soil moisture at {5,10,20,40} cm per station.
Merge & split. Merge on timestamp; drop rows with any missing values. Split: train = earliest 70%, test = remaining 30%.
Fit scalers. Fit $\mathrm{MM}_x$ on train $x_t$; compute $\tilde{x}_t$ for train/test. Fit $\mathrm{MM}_y$ per depth on train $\mathbf{y}_t$; compute $\tilde{\mathbf{y}}_t$.
Conditioning. Build $\mathbf{c}_t = [\tilde{x}_t, \sin(2\pi m_t/12), \cos(2\pi m_t/12)]^\top$.
Regimes. Compute train quantiles $q_{0.33}, q_{0.66}$ of $\tilde{x}_t$; define Dry/Mid/Wet subsets (fallback to full train if a subset is too small).
Schedule. Construct linear $\beta_\tau$; set $\alpha_\tau = 1 - \beta_\tau$; $\bar{\alpha}_\tau = \mathrm{cumprod}(\alpha_\tau)$.
Train each agent $a \in \{\mathrm{dry}, \mathrm{mid}, \mathrm{wet}\}$.
for epoch = 1 to $E$ do
  for minibatch $\{(\tilde{\mathbf{y}}_t, \mathbf{c}_t)\}$ from agent's regime data do
    Sample $\tau \in \{0, \ldots, T_{\mathrm{diff}} - 1\}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_3)$
    Noising: $\mathbf{x}_\tau \leftarrow \sqrt{\bar{\alpha}_\tau}\,\tilde{\mathbf{y}}_t + \sqrt{1 - \bar{\alpha}_\tau}\,\boldsymbol{\epsilon}$
    Denoise: $\hat{\boldsymbol{\epsilon}} \leftarrow \boldsymbol{\epsilon}_\theta(\mathbf{x}_\tau, \tau, \mathbf{c}_t)$
    Reconstruct: $\hat{\mathbf{x}}_0 \leftarrow \left(\mathbf{x}_\tau - \sqrt{1 - \bar{\alpha}_\tau}\,\hat{\boldsymbol{\epsilon}}\right)/\left(\sqrt{\bar{\alpha}_\tau} + \varepsilon\right)$
    Losses: $\mathcal{L}_{\mathrm{data}} = \|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}\|_2^2$; $\mathcal{L}_{\mathrm{flux}}$; $\mathcal{L}_{\mathrm{bounds}}$
    Total: $\mathcal{L} = \mathcal{L}_{\mathrm{data}} + \lambda_{\mathrm{flux}}\mathcal{L}_{\mathrm{flux}} + \lambda_{\mathrm{bounds}}\mathcal{L}_{\mathrm{bounds}}$; update $\theta_a$ (Adam, clip 1.0)
  end for
end for
Inference (for $t = 1 : T_{\mathrm{eval}}$).
Compute gating weights $\mathbf{w}_t$ via triangular memberships; set $a^\star(t) = \mathrm{argmax}_a w_a(t)$.
For each agent $a$: estimate mean $\boldsymbol{\mu}_{a,t} = \frac{1}{M}\sum_{m=1}^{M}\textsc{SampleDDPM}(\theta_a, \mathbf{c}_t)$; mix $\boldsymbol{\mu}_{\mathrm{mix},t} = \sum_a w_a(t)\,\boldsymbol{\mu}_{a,t}$.
Draw $N$ samples only from $a^\star(t)$: $\left\{\tilde{\mathbf{y}}_t^{(n)}\right\}_{n=1}^{N} \leftarrow \textsc{SampleDDPM}(\theta_a, \mathbf{c}_t)$; compute $\bar{\mathbf{x}}_t$ and std/min/max (normalized).
Consensus corrector: initialize $\mathbf{z} \leftarrow \bar{\mathbf{x}}_t$; for $s = 1 : S$ take an Adam step on
    $J(\mathbf{z}) = \|\mathbf{z} - \boldsymbol{\mu}_{\mathrm{mix},t}\|_2^2 + \lambda_{\mathrm{flux}}\mathcal{L}_{\mathrm{flux}}(\mathbf{z}) + \lambda_{\mathrm{bounds}}\mathcal{L}_{\mathrm{bounds}}(\mathbf{z})$;
set normalized mean $\mathbf{z}_t^\star \leftarrow \mathbf{z}$.
Back-transform: mean $\mathrm{MM}_y^{-1}(\mathbf{z}_t^\star)$; min/max via $\mathrm{MM}_y^{-1}$; std by $s_{\mathrm{phys}} = \tilde{s}/\mathrm{scale}$ (element-wise).
Compute metrics (per depth $d \in \{10, 20, 40\}$). ($\mathrm{RMSE}_d$, $\mathrm{nRMSE}_d$, and $R_d^2$).
Outputs. Time-indexed predictions (physical units) with uncertainty; evaluation metrics.
```

Physics enters as auxiliary regularization on the denoised estimate at each training step. Given the noisy input $\mathbf{x}_\tau$ and the network's prediction $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\theta(\mathbf{x}_\tau, \tau, \mathbf{c}_t)$, we reconstruct the clean signal via the standard estimator, as shown in Equation 6:

$$\hat{\mathbf{x}}_0 = \frac{\mathbf{x}_\tau - \sqrt{1 - \bar{\alpha}_\tau}\,\hat{\boldsymbol{\epsilon}}}{\sqrt{\bar{\alpha}_\tau} + \varepsilon}, \tag{6}$$

with a small $\varepsilon > 0$ for numerical safety. Two priors are applied to $\hat{\mathbf{x}}_0$. The first is a flux-consistency penalty across depths, which implements a discrete second-difference regularizer. Writing $\hat{x}_{10}, \hat{x}_{20}, \hat{x}_{40}$ for the components of $\hat{\mathbf{x}}_0$, we define first differences $g_1 = \hat{x}_{20} - \hat{x}_{10}$ and $g_2 = \hat{x}_{40} - \hat{x}_{20}$. The penalty $\mathcal{L}_{\mathrm{flux}}$ as shown in Equation 7:

$$\mathcal{L}_{\mathrm{flux}} = \mathbb{E}\left[(g_2 - g_1)^2\right], \tag{7}$$

discourages abrupt curvature between adjacent layers and is a lightweight surrogate for a constant-diffusivity Fickian profile over the three discretized depths (Wu, 2003). The second is a feasible-range barrier that softly keeps predictions inside $[0,1]^3$ in normalized units by penalizing excursions with the softplus function $\text{sp}(u) = \log(1 + e^u)$, as defined in Equation 8:

$$\mathcal{L}_{\text{bounds}} = \mathbb{E}[\text{sp}(0 - \hat{\mathbf{x}}_0) + \text{sp}(\hat{\mathbf{x}}_0 - 1)]. \tag{8}$$

The agent's total objective, which combines data and physics terms, is defined in Equation 9:

$$\mathcal{L} = \mathcal{L}_{\text{data}} + \lambda_{\text{flux}}\mathcal{L}_{\text{flux}} + \lambda_{\text{bounds}}\mathcal{L}_{\text{bounds}}, \tag{9}$$

with physics-regularization weights set at modest magnitudes so that they act as auxiliary guidance rather than overwhelming the data term. We selected these coefficients through preliminary sweeps to ensure that the physics-driven gradients remained consistently smaller than the likelihood gradients, thereby encouraging physically coherent profiles while preserving data fidelity. To formalize this selection, we conducted an ablation study over $\lambda_{\text{flux}} \in \{0.05, 0.1, 0.2, 0.4\}$ and $\lambda_{\text{bounds}} \in \{0.01, 0.03, 0.05, 0.1\}$ across multiple stations and depths. Moderate values provided the best balance: larger weights produced overly smoothed vertical profiles, whereas smaller weights made the priors effectively inactive. Based on these results, we fixed $\lambda_{\text{flux}} = 0.2$ and $\lambda_{\text{bounds}} = 0.05$ for all experiments. Using global (non site-specific) weights further avoids overfitting and preserves the domain-agnostic character of the proposed multi-agent diffusion framework.

At inference the three agents are coupled by a fuzzy gating mechanism that assigns time-varying regime weights from the test-time driver $\tilde{x}_t$. Let $x_{\min}^{\text{test}}$ and $x_{\max}^{\text{test}}$ denote the minimum and maximum of $\tilde{x}_t$ on the test set, and let $\delta > 0$ be a small constant to avoid division by zero. Triangular memberships are defined by Equations 10 and 11:

$$w_{\text{dry}}(t) = \text{clip}\left(\frac{q_{0.33} - \tilde{x}_t}{q_{0.33} - x_{\min}^{\text{test}} + \delta}, 0, 1\right), \qquad w_{\text{wet}}(t) = \text{clip}\left(\frac{\tilde{x}_t - q_{0.66}}{x_{\max}^{\text{test}} - q_{0.66} + \delta}, 0, 1\right), \tag{10}$$

and

$$w_{\text{mid}}(t) = 1 - \max\{w_{\text{dry}}(t), w_{\text{wet}}(t)\}, \qquad \mathbf{w}_t = \frac{[\, w_{\text{dry}}(t), w_{\text{mid}}(t), w_{\text{wet}}(t)\,]}{\sum_a w_a(t) + \delta}, \tag{11}$$

where $\text{clip}(u, 0, 1) = \min\{\max\{u, 0\}, 1\}$. To keep computation close to a single model, we perform "top-1" sampling: at each chronological time t we draw diffusion samples only from the agent $a^\star(t) = \text{argmax}_a w_a(t)$. Nevertheless, the final mean is allowed to negotiate among agents. Before forming the consensus, each agent's expected output at time t is approximated with a small Monte-Carlo average $\boldsymbol{\mu}_{a,t} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{x}_{a,t}^{(m)}$ with M samples per agent. For ensemble size N, we first compute the ensemble mean $\bar{\mathbf{x}}_t$ from N samples drawn under the top-1 policy (all in normalized space). We then refine $\bar{\mathbf{x}}_t$ by solving a three-variable optimization initialized at $\bar{\mathbf{x}}_t$, as formulated in Equation 12:

$$\mathbf{z}_t^\star = \underset{\mathbf{z} \in \mathbb{R}^3}{\text{argmin}} \left\| \mathbf{z} - \sum_a w_a(t)\boldsymbol{\mu}_{a,t} \right\|_2^2 + \lambda_{\text{flux}}\mathcal{L}_{\text{flux}}(\mathbf{z}) + \lambda_{\text{bounds}}\mathcal{L}_{\text{bounds}}(\mathbf{z}), \tag{12}$$

using a small number of Adam steps (here, 30 at step size 0.05). This "consensus corrector" is negligible in cost but enforces agreement with the fuzzy-weighted mixture of agent means while re-imposing diffusive vertical structure and feasible bounds. The corrected mean $\mathbf{z}_t^\star$ becomes the reported mean in normalized units.

Uncertainty is quantified through ensembles. For each $N \in \{5, 10, \ldots, 100\}$ the sampler draws N trajectories per timestamp under the top-1 gating and records the mean, standard deviation, minimum, and maximum over the N samples. The inverse mapping to physical units is exact for affine min–max transforms: means, minima, and maxima are mapped by MM_y^{-1} component-wise. For dispersion, let $\tilde{s}$ denotes the normalized standard deviation

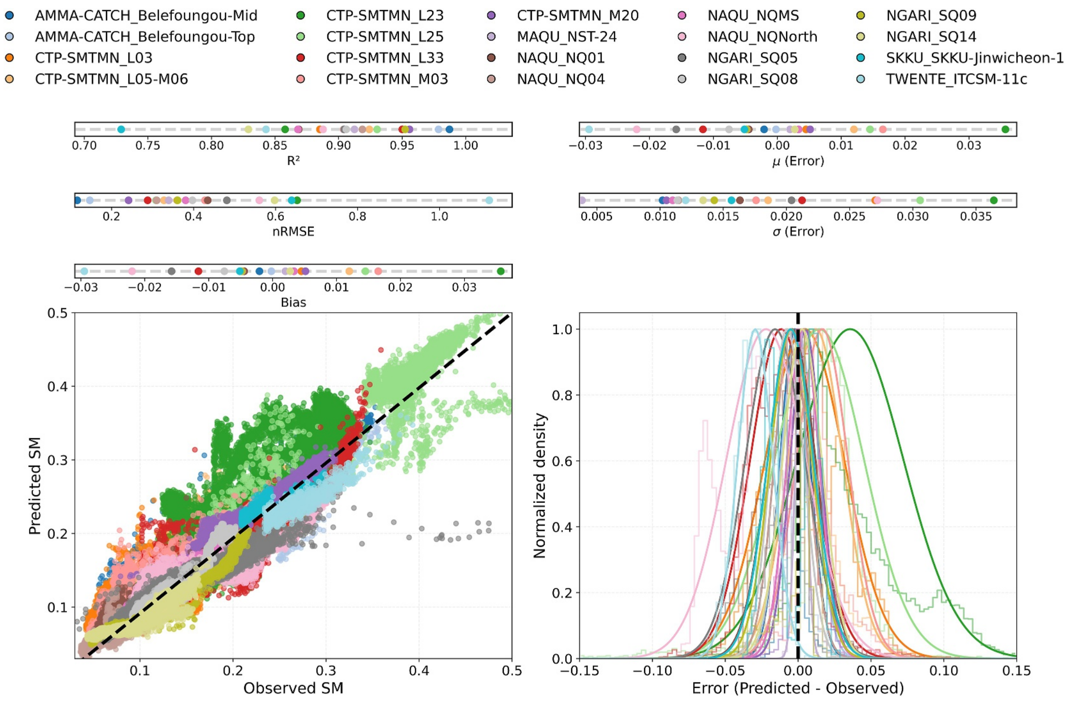


Figure 3. Cross-site evaluation of 10-cm soil-moisture predictions from the $N = 100$ physics-informed multi-agent ensemble model. The left panel plots observed against predicted ensemble means for all stations (colors mark stations). The dashed black 1:1 line indicates perfect agreement. Insets summarize per-station performance using the coefficient of determination, normalized RMSE, and mean bias. The right panel shows the error histogram analysis at each station. Step curves show the distribution of timestamp-wise prediction errors, peak-normalized to enable cross-station comparison. Solid curves show normal-distribution fits, and the insets report the fitted mean (systematic bias) and standard deviation (random spread).

of a component, and let scale $= 1/(y_{\max} - y_{\min})$ be the corresponding scaler factor from training. Then, we apply Equation 13 element-wise to obtain physically scaled uncertainties:

$$s_{\text{phys}} = \frac{\tilde{s}}{\text{scale}} = \tilde{s}\,(y_{\max} - y_{\min}). \tag{13}$$

Evaluation on the held-out temporal segment uses root-mean-squared error computed per depth $d \in \{10, 20, 40\}$ using Equation 14:

$$\text{RMSE}_d = \sqrt{\frac{1}{T_{\text{eval}}} \sum_{t=1}^{T_{\text{eval}}} \left(\hat{y}_{d,t} - y_{d,t}\right)^2}, \tag{14}$$

normalized RMSE, $\text{nRMSE}_d = \text{RMSE}_d/\sigma_{\text{obs},d}$ using the observed test-set standard deviation $\sigma_{\text{obs},d}$ (with a guard for $\sigma_{\text{obs},d} = 0$), and the coefficient of determination using Equation 15:

$$R_d^2 = 1 - \frac{\sum_{t=1}^{T_{\text{eval}}} \left(\hat{y}_{d,t} - y_{d,t}\right)^2}{\sum_{t=1}^{T_{\text{eval}}} \left(y_{d,t} - \bar{y}_d\right)^2}, \qquad \bar{y}_d = \frac{1}{T_{\text{eval}}} \sum_{t=1}^{T_{\text{eval}}} y_{d,t}. \tag{15}$$

3. Results and Discussion

3.1. Model Performance

We plot the linear fit between predicted and observed moisture (left panel of Figure 3). The data points are clustered close to the 1:1 line, indicating a good model fit. The model performs well in estimating 10 cm soil moisture across the 20 validation sites, with coefficient of determination (R^2) values ranging from 0.73 to 0.99.

Most stations show an R^2 around 0.9 (median $R^2 \approx 0.91$), meaning that a large fraction of the variance in observed soil moisture is captured. The errors are generally low in magnitude, with normalized root-mean-square errors (nRMSE) ranging from 0.116 at the best site to 1.121 at the worst site, with most stations between 0.3 and 0.5 (median $\approx$ 0.37). Model bias is minimal, with the mean prediction error (Bias) staying within ± 0.035 volumetric water content units for all stations, indicating no significant over- or under-prediction. This is reflected in the error histogram (right panel of Figure 3), where the residuals for most stations are tightly clustered around zero. At many sites, the error standard deviation is around 0.01–0.02 (in volumetric fraction), further emphasizing the model's accuracy. A time series for several sites at different depths is shown in Figure S1.

The best performance is observed at two tropical savannah sites in Benin (part of the AMMA-CATCH network). Specifically, the Belefoungou-Mid and Belefoungou-Top stations (which have an Aw climate, tropical wet-dry savannah) show very high agreement between model predictions and actual observations. Both stations have R^2 values close to 0.98–0.99, meaning the model's predictions are almost perfectly aligned with the real data. The nRMSE is also very low (around 0.12–0.15), and the bias is nearly zero (<0.3% in absolute terms). This suggests the model accurately captures the soil moisture dynamics in the tropical savannah. The data points for these stations are clustered along the 1:1 line, indicating highly precise predictions. This strong performance is likely due to the clear seasonal cycle and relatively consistent soil properties at these Benin sites, which the model seems to handle very well.

In contrast, the weakest model performance is seen at one temperate and one monsoonal site. The most challenging site is TWENTE_ITCSM-11c in the Netherlands (Cfb climate, temperate maritime). Here, the nRMSE is the highest, around 1.12, meaning that the prediction error is slightly greater than the average soil moisture value. This is reflected in a wider error distribution and greater scatter in the results. Although the correlation remains fair $\left(R^2 \approx 0.84\right)$, it is lower than at most other sites, indicating that some of the moisture variability is not captured. A moderate bias is also present, with the model tending to underestimate soil moisture (bias ≈ -0.03 in volumetric fraction), which appears as a shift in the error histogram below zero. Another relatively weak result comes from the SKKU Jinwicheon-1 site in South Korea (Cwa climate, temperate with dry winters). This location shows the lowest correlation $\left(R^2 \approx 0.73\right)$ among all stations, indicating weaker agreement between predictions and observations. Its nRMSE is about 0.64, which is relatively high compared to the network average. The scatterplot for this site shows a wider spread around the 1:1 line, consistent with larger and more variable errors possibly due to complex topography or strong variability in monsoonal rainfall.

Despite these challenges, the model still maintains moderate accuracy, with small bias (around 0.5% underestimation) even at Jinwicheon-1. Overall, while these two temperate-climate sites one maritime (Netherlands) and one monsoonal (Korea) show the lowest performance, the model still provides reasonably good estimates. The weaker results may be due to factors such as limited training data, local soil characteristics (e.g., sandy soils (86%) at Twente with fast drainage), or highly variable climate conditions that are harder to model.

A clear climate-related trend is observed: stations in tropical and high-latitude/cold climates generally show better model performance than those in certain temperate regions. The two Benin sites (Aw climate) are examples where tropical savannah environments lead to very high predictability of soil moisture. The seasonal rainfall patterns and moderate soil texture (around 22% clay, 49% sand) likely produce soil moisture variations that the model tracks effectively. Similarly, many high-altitude cold-climate stations on the Tibetan Plateau (China) show R^2 values between 0.85 and 0.95, with low error. For instance, stations in the CTP-SMTMN and NAQU networks in alpine climates (ET-polar tundra, and Dwc-cold dry-winter) show R^2 values around 0.87–0.95 and nRMSE values between 0.30 and 0.43, indicating strong performance. These sites (e.g., Naqu and Maqu) experience long frozen winters and short summers, and the model's success suggests it accurately captures the seasonal thaw and moisture dynamics. The error histograms for these cold-climate stations show tight residual distributions, indicating consistent model behavior under freeze-thaw conditions.

Arid desert climates (BWk) present a more mixed result. The four NGARI stations in northwestern China (cold desert steppe) showed high correlations (R^2 around 0.83–0.95), but the nRMSE values (0.36–0.60) were higher. The SQ14 site (BWk, sparse vegetation) has R^2 around 0.83 and nRMSE around 0.60, making it one of the lower-performing sites. These desert areas have sandy soils (up to ~87% sand at SQ14) and low moisture content, which amplify relative errors. Even small absolute errors become large fractions of the average moisture, raising the nRMSE. Nevertheless, the model remains largely unbiased in arid environments (e.g., SQ14 bias is about

Table 3
Summary of Model Behavior at Higher Depths (20 and 40 cm): Deeper Layers Generally Exhibit Reduced Agreement and Broader Residual Spreads Relative to 10 cm, Consistent With Damped and Lagged Hydrologic Responses

		R^2			nRMSE			Bias		
Site	ISMN station	10 cm	20 cm	40 cm	10 cm	20 cm	40 cm	10 cm	20 cm	40 cm
S1	AMMA-CATCH_Belefoungou-Mid	0.987	0.957	0.913	0.116	0.263	0.300	−0.002	0.014	0.003
S2	AMMA-CATCH_Belefoungou-Top	0.979	0.937	0.888	0.147	0.283	0.372	0.000	0.008	−0.012
S3	CTP-SMTMN_L03	0.885	0.847	0.917	0.360	0.651	0.340	0.004	0.026	−0.009
S4	CTP-SMTMN_L05-M06	0.924	0.938	0.929	0.328	0.305	0.430	0.012	−0.004	0.019
S5	CTP-SMTMN_L23	0.858	0.929	0.935	0.653	0.319	0.325	0.036	0.010	−0.009
S6	CTP-SMTMN_L25	0.930	0.922	0.844	0.308	0.317	0.483	0.015	−0.007	−0.016
S7	CTP-SMTMN_L33	0.950	0.893	0.884	0.288	0.470	0.359	−0.012	−0.019	−0.003
S8	CTP-SMTMN_M03	0.904	0.800	0.623	0.428	0.567	0.632	0.017	0.011	0.000
S9	CTP-SMTMN_M20	0.956	0.935	0.912	0.241	0.292	0.362	0.005	−0.005	−0.001
S10	MAQU_NST-24	0.912	0.903	0.839	0.339	0.976	0.544	0.002	0.010	0.004
S11	NAQU_NQ01	0.869	0.842	0.626	0.434	0.567	0.634	−0.004	−0.012	−0.004
S12	NAQU_NQ04	0.919	0.835	0.791	0.310	0.415	0.548	−0.005	−0.001	0.008
S13	NAQU_NQMS	0.868	0.738	0.733	0.380	0.519	0.558	0.003	−0.002	−0.001
S14	NAQU_NQNorth	0.888	0.911	0.873	0.560	3.829	0.564	−0.022	0.110	−0.015
S15	NGARI_SQ05	0.904	0.870	0.617	0.481	0.399	0.625	−0.016	−0.003	−0.006
S16	NGARI_SQ08	0.906	0.827	0.473	0.397	0.907	0.930	−0.008	−0.009	−0.006
S17	NGARI_SQ09	0.952	0.926	0.719	0.361	0.385	0.679	−0.005	0.000	−0.001
S18	NGARI_SQ14	0.829	0.821	0.763	0.598	1.761	0.592	0.003	−0.036	0.003
S19	SKKU-Jinwicheon-1	0.729	0.735	0.630	0.639	2.768	1.153	−0.005	0.041	0.026
S20	TWENTE_ITCSM-11c	0.843	0.791	0.626	1.121	0.987	2.365	−0.029	−0.022	−0.019

+0.003), and for other NGARI stations, such as SQ09 (which also has sandy soil), the model's performance was very good $\left(R^2 \sim 0.95\right)$, with moderate error. This suggests that the model can capture moisture fluctuations in arid regions fairly well, although the limited moisture and rapid drainage in sandy soils make perfect predictions difficult.

Temperate climates, on the other hand, show relatively lower performance, as seen in the Netherlands (Cfb) and South Korea (Cwa). The Cfb climate in the Netherlands, characterized by frequent rainfall and mild summer, might result in moderate and persistent soil moisture, but the model's underestimation indicates possible calibration issues or unmodeled factors (such as high groundwater or land management practices). The Cwa climate in South Korea, with dry winters and heavy summer rains, likely presents challenges in capturing the timing and intensity of monsoon rainfall, indicating the model may require more localized tuning. Soil texture and organic content also vary across these climates: the coarse soil at the Dutch site (around 4% clay and 86% sand) likely causes rapid moisture fluctuations, making predictions more difficult. In contrast, the Korean site, with a more balanced soil texture (23% clay, 40% silt), had a very short observation period (only one summer), which may not fully capture the site's climate variability.

3.2. Performance at Higher Depth

As depth increases from 10 to 40 cm, the model's performance metrics clearly show a decline in the performance (Table 3). Specifically, the coefficient of determination drops significantly from $R^2 \approx 0.91$ at 10 cm to around 0.80 at 40 cm, while the normalized RMSE roughly doubles, increasing from about 0.37 to 0.55. These changes suggest that the model explains less variance and incurs larger relative errors at deeper soil layers. In contrast, the bias remains very small at all depths (on the order of 10^{-3} in volumetric moisture fraction, with no consistent increase), indicating that there is no systematic over- or under-prediction with depth. The drop in R^2 and the rise in

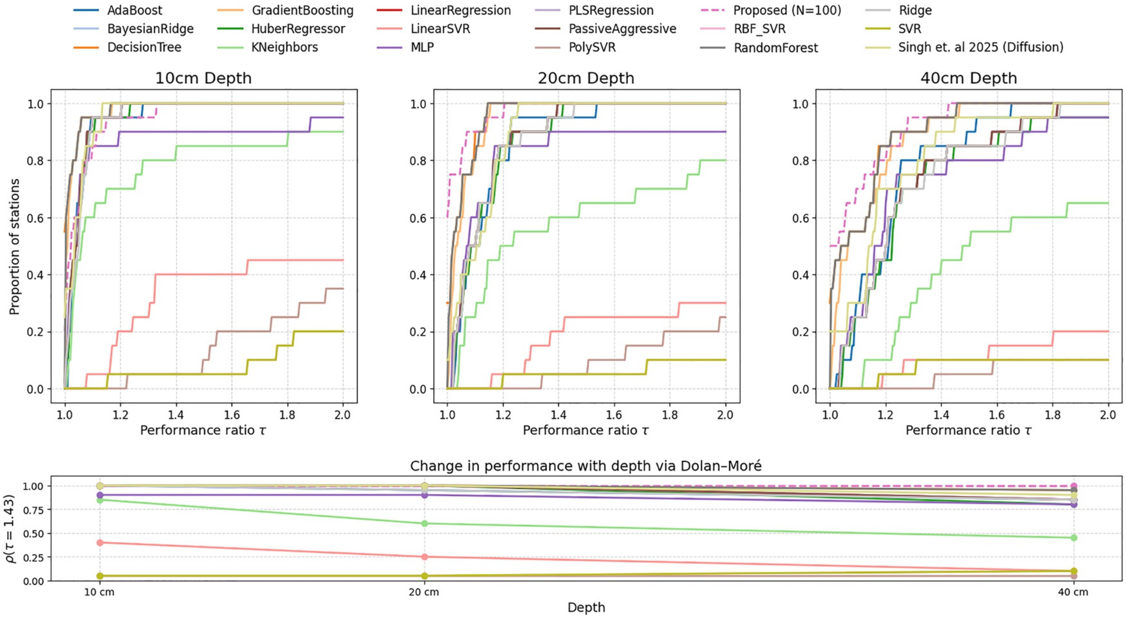


Figure 4. Performance comparison of multiple algorithms across depths using Dolan–Moré profiles. Top row: For each depth (10, 20, 40 cm), the Dolan–Moré performance profile shows the proportion of stations for which each model achieves a performance ratio $\leq \tau$, where τ is computed relative to the best-performing model at each station (larger values indicate better performance). Bottom row: Variation in $\rho(\tau)$ with depth, where $\rho(\tau)$ is the proportion of stations below a fixed performance ratio threshold ($\tau = 1.43$) for each model.

nRMSE suggest that the model's performance decreases with increasing depth, likely because the dynamics of deeper soil moisture are more buffered and exhibit less variability. Deeper layers smooth out short-term fluctuations, which reduces the signal-to-noise ratio for the model. As a result, even small prediction errors become more significant relative to the lower variability at deeper depths, and time-lagged processes (such as delayed percolation and upward capillary flux) become more difficult for the model to capture. Therefore, the model performs best at 10 cm, and degrades progressively at 20 and 40 cm, which is consistent with the understanding that deeper soil moisture signals are harder to predict due to their reduced and delayed variability.

3.3. Comparison With the Benchmark Algorithms

We employ Dolan-Moré performance profiles to compare the proposed method against benchmark models (Figure 4). A performance profile plots the fraction $\rho_s(\tau)$ of test problems on which algorithm s attains performance within a factor τ of the best solver. Formally, for a set of solvers S and test problems P, let $t_{p,s}$ be a performance measure (e.g., R^2) of solver s on problem p. Define the performance ratio as shown in Equation 16:

$$r_{p,s} = \frac{t_{p,s}}{\min_{s' \in S} t_{p,s'}}. \tag{16}$$

Then the performance profile $\rho_s(\tau)$ is calculated using Equation 17:

$$\rho_s(\tau) = \frac{1}{|P|} \left|\{p \in P: \ r_{p,s} \leq \tau\}\right|, \qquad \tau \geq 1, \tag{17}$$

that is, the fraction of problems p for which solver s is within a factor τ of the best attainable performance on p. In particular, $\rho_s(1)$ is the fraction of problems where s is *best*, while $\rho_s(\tau)$ for larger τ reflects robustness (the ability to achieve near-best results across P). We constructed performance profiles separately for each depth (10, 20, 40 cm) by taking all test instances at a given depth as the problem set P and computing $\rho_s(\tau)$ for every algorithm, with special attention to $\tau = 1.43$ (within 43% of the best) for the deepest depth.

At the shallow depth of 10 cm, most baseline algorithms perform very well. For example, AdaBoost, Decision Tree, Gradient Boosting, and several other models achieve $\rho(1.3) \approx 1.0$, meaning they are essentially always within 30% of the best solver on 10 cm tasks. The Proposed method attains $\rho(1.3) = 0.95$ at 10 cm, slightly below the top tier (it is within 30% of the best solution on 95% of the shallow-depth problems). This indicates that in the easiest scenario (10 cm), a few existing models were marginally more often the fastest or most accurate, though the proposed approach still solved the vast majority of cases near-optimally. At the intermediate depth of 20 cm, the proposed algorithm's performance profile improves to $\rho(1.2) = 1.0$, meaning it matched or came within 20% of the best solver on every 20 cm test problem. In other words, our method solved 100% of the medium-depth cases essentially as well as the optimal model. Many competing methods begin to show declines in performance by 20 cm. For instance, AdaBoost drops to $\rho(1.2) = 0.8$, and k-Nearest Neighbors (KNN) plummets to 0.5, indicating that KNN is within 20% of the best on only 50% of the 20 cm cases. Several robust linear models (Bayesian Ridge, Huber, Passive-Aggressive) and neural nets (MLP) maintain decent performance at 20 cm ($\rho(1.2) = 0.80$–0.85 range), but they still do not reach the proposed method's perfect score. This also includes the recently proposed diffusion model by A. Singh, Singh, & Gaurav (2025b), which demonstrates robust performance at 10 cm, but its performance starts deteriorating as the depth increases ($\rho(1.2) = 0.80$). These results demonstrate that as the problem difficulty increases (from 10 to 20 cm), the proposed algorithm adapts well, overtaking most benchmarks to become one of the top-performing approach at $\tau = 1.2$ for 20 cm along with Decision Tree, Gradient Boosting and Random Forest. The performance gap becomes most pronounced at the deepest depth of 40 cm. In this challenging scenario, our algorithm retains a perfect $\rho(1.43) = 1.0$ with no other model achieves a 100% success fraction at 40 cm at this threshold. The next-best performers (Decision Tree, Gradient Boosting, and Random Forest) reach only $\rho(1.43) = 0.95$ at 40 cm, meaning they fail to produce near-best results on 5% of the test cases that the proposed method still handles effectively. Most other baseline algorithms degrade substantially by 40 cm. For example, Bayesian Ridge and Huber Regressor drop to $\rho(1.43) = 0.80$ (only 80% of the problems within 43% of best), and KNN falls further to 0.45. Several support vector regression variants essentially break down at this depth: Linear SVR yields $\rho(1.43) = 0.10$ and the RBF-kernel SVR only 0.10, while the polynomial SVR solves 5% of the deep cases within 43% of the best ($\rho(1.43) = 0.05$). In summary, the proposed model outperforms all benchmarks at 40 cm by a wide margin. Its $\rho(1.43) = 1.0$ indicates that on every deep test instance, our method's error remains within 43% of the best-observed error (and often it is the best), whereas all other algorithms struggle on a significant fraction of these difficult problems.

The observed trend, where many baselines perform similarly at 10 cm but diverge sharply at deeper layers, can be explained by the strengths of the proposed framework. The Dry/Mid/Wet regime specialization allows each agent to learn depth-dependent and moisture-state-dependent dynamics rather than forcing a single model to represent all regimes simultaneously, which becomes important as the vertical gradients weaken and noise increases with depth. In parallel, the weak physics constraints provide gentle but meaningful guidance that prevents unrealistic curvature and stabilizes predictions in zones where observational information becomes sparse or ambiguous. Together, these components help the model maintain physical coherence and robustness as the problem becomes harder, enabling it to outperform generic machine-learning baselines at 40 cm where purely data-driven methods tend to deteriorate.

3.4. Stochastic Robustness Analysis

Figure 5 presents a stochastic robustness analysis of the model's performance across all test sites and soil depths (10, 20, and 40 cm). For each site, a violin plot shows the distribution of the metrics (nRMSE and R^2) over 30 runs with different random seeds. The solid line within each violin denotes the site's median performance, while the dashed horizontal line indicates the global median across all sites for that depth. Additionally, a strip chart above each panel plots each site's median as a colored dot on the same axis (with the panel-wise global median marked by a vertical dashed line), allowing a quick visual comparison of site medians against the overall median. The model's accuracy is highest at the shallowest soil depth and declines with increasing depth. At 10 cm, the global median R^2 across all sites and seeds is about 0.91, and the median nRMSE is around 0.37. By 20 cm these values shift to a median $R^2 \approx 0.88$ and nRMSE ≈ 0.49. The drop in performance is most pronounced at 40 cm, where the median R^2 falls to ~0.80 and median nRMSE rises to ~0.55. This trend suggests that the model captures surface soil moisture dynamics better than deeper-layer moisture. Deeper soil moisture is generally more

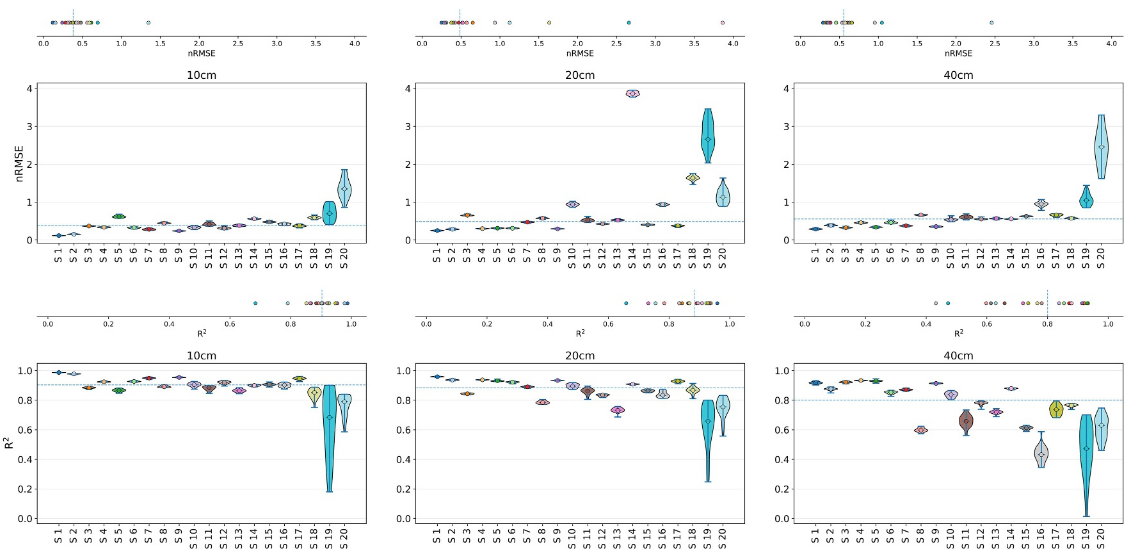


Figure 5. Stochastic robustness analysis of model performance across sites and depths. Columns denote target depth (10, 20, 40 cm) and rows denote metrics (R^2 and nRMSE). For each site, violins summarize the distribution over stochastic seeds (random initializations); the central line is the site median, and the dashed horizontal line is the panel-wise global median. Above each panel, a strip inset plots site medians as colored dots on the metric axis, with a vertical dashed line for the global median.

challenging to predict, likely due to the integrative effects of longer-term processes and subsurface heterogeneity, so a reduction in R^2 and an increase in error with depth is expected.

Despite high overall accuracy at 10 cm, performance varied across sites. Most sites achieved R^2 values above 0.85 at 10 cm, and several sites had exceptionally high accuracy. For example, the AMMA-CATCH Belefoungou Mid site (a West African site, S1) attained a median $R^2 \approx 0.99$ at 10 cm with a correspondingly small nRMSE (≈0.12), essentially a near-perfect prediction across all seed runs. Many other sites in favorable conditions (e.g., ample training data and strong seasonal signals) also showed very high and consistent performance at 10 cm. On the other hand, a few sites proved difficult even at the surface layer. The SKKU Jinwicheon-1 site (a Korean site, S19) had the lowest median R^2 (∼0.68) at 10 cm and a relatively high median nRMSE (∼0.70). Its violin plot is notably wider than others, across the 30 seeds, R^2 for this site ranged from as high as ∼0.90 down to essentially 0 (as low as 0.18 in one run). In other words, depending on the random initialization, the model's success on the Jinwicheon-1 data varied dramatically. Such a broad distribution suggests that the model training at this site is not consistently converging to a good solution, possibly due to the limited data (less than 1 year) that make the optimization landscape challenging. Aside from this outlier, however, most 10 cm site violins are narrow, indicating that performance is robust to different random seeds, the central (median) values might differ by site, but the variance due to initialization is generally small.

At 20 cm depth, the overall performance declines slightly from 10 cm, and most sites follow the same pattern (slightly lower R^2, higher error). The spread of performance distributions (violin widths) for most sites remains fairly tight, implying the model's robustness to random seed persists at 20 cm. However, this depth also reveals some extreme site-specific anomalies. The most striking case is the NAQU_NQNorth site (S14). This site's median R^2 at 20 cm is actually very high (∼0.91, indicating the model captures temporal fluctuations well), but its nRMSE is by far the worst of any site (median ∼3.87, with the distribution ranging roughly 3.77–3.96). In the Figure 5, NAQU_NQNorth's violin at 20 cm is extremely high on the nRMSE axis, standing out above all others. This combination of high R^2 and large nRMSE suggests that the relative errors are high despite correctly tracking the direction of change. This is possibly due to the fact that the observed soil moisture at this site has an unusually small dynamic range (e.g., nearly constant moisture). In such cases, even a small absolute prediction error can translate to a very large normalized error, inflating nRMSE while R^2 remains high (since R^2 is insensitive to bias and reflects correlation). Apart from NAQU, the SKKU Jinwicheon-1 site again shows notably high error at 20 cm (second-highest nRMSE with a median ∼2.66) and a lower-than-average R^2 (∼0.66), consistent with the difficulties observed at 10 cm. These outliers highlight that certain site conditions (or data issues) can significantly affect model performance and its stability across runs.

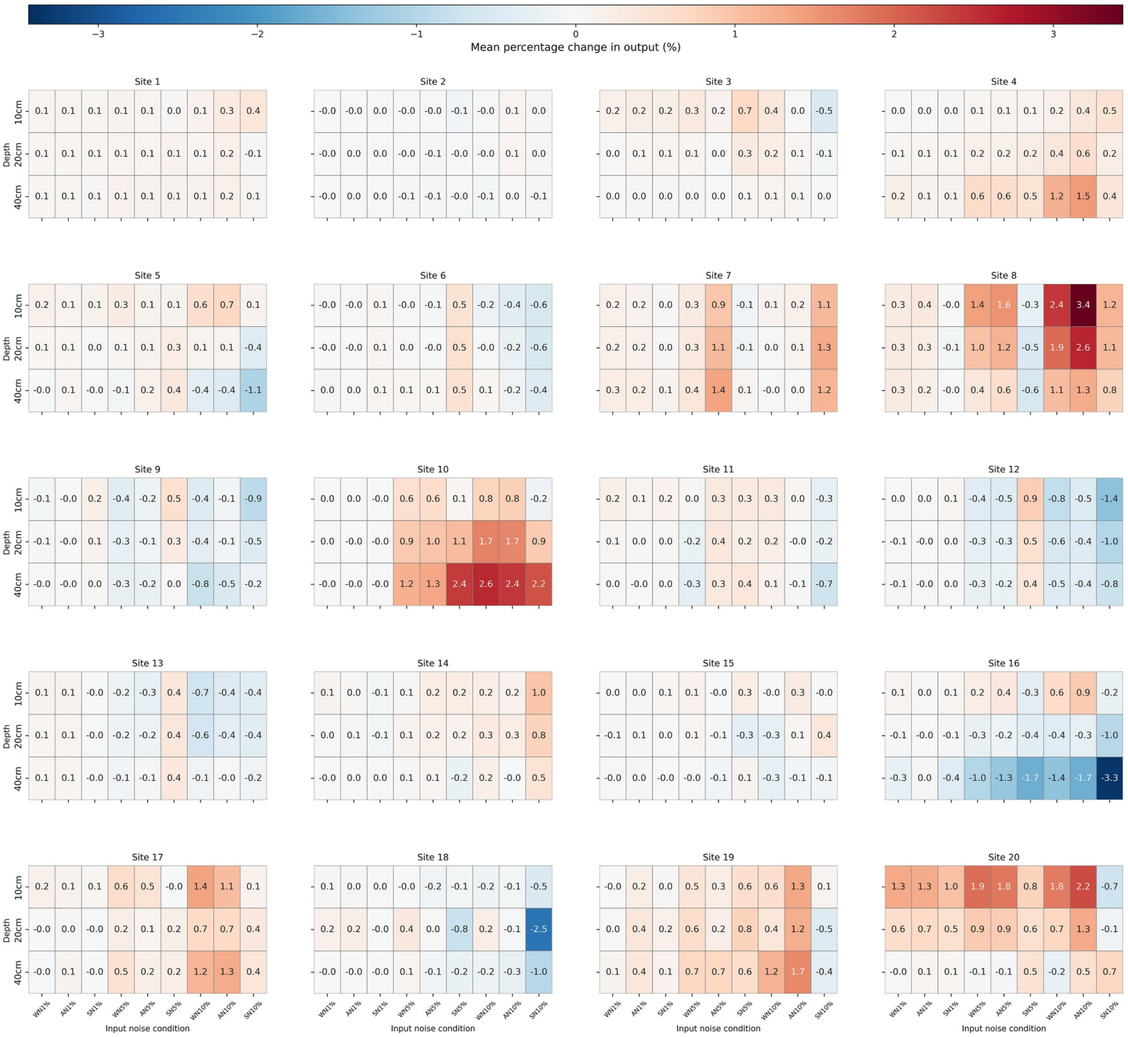
The disparity between sites becomes more pronounced at 40 cm, and overall error levels are higher. Still, many sites retain reasonably good performance. For instance, the best-performing site at 40 cm (Belefoungou Mid) has a median R^2 around 0.92 with nRMSE ~0.29, only a modest drop from its 10 cm performance, indicating the model can generalize well to deeper soil moisture at that location. In contrast, some sites see substantial degradation. NGARI SQ08 (located in an arid, high-altitude region, S14) exhibits the lowest median R^2 (~0.43) at 40 cm, suggesting the model struggles to explain much of the variance in deep moisture. The SKKU Jinwicheon-1 also remains among the poorest performers (median $R^2 \approx 0.47$) and continues to show a wide performance spread, in fact, in one of the 40 cm seed runs the model had essentially no skill ($R^2 \approx 0.01$). Another noteworthy case is the TWENTE_ITCSM-11c site (S20). At 40 cm its median nRMSE is ~2.46, which is extremely high compared to other sites (most of which have 40 cm nRMSE below 0.7). This indicates a deep-layer prediction error at TWENTE that is much larger relative to the variability in soil moisture, possibly pointing to some form of model mismatch or data peculiarity (e.g., a shallow water table or measurement error at 40 cm). Despite these few cases, it's important to note that most sites at 40 cm still achieve median R^2 values in the 0.6–0.9 range and nRMSE in the 0.4–0.6 range, demonstrating that the model maintains a decent level of predictive power even for deeper soil moisture.

3.5. Uncertainty Analysis

To evaluate robustness, we performed uncertainty analysis by perturbing the 5 cm input with three representative error processes: (a) white Gaussian noise (zero-mean, independent across time) to mimic electronic jitter; (b) autoregressive AR(1) noise with coefficient (≈ 0.8) to emulate short-term, temporally persistent bias; and (c) structured low-frequency drift modeled as a single-cycle sinusoid with random phase to represent slow calibration/temperature-induced offsets. Each process was tested at ±1%, ±5%, and ±10% amplitude, spanning typical to conservative sensor-error magnitudes. This design captures high-frequency uncorrelated noise, correlated biases with memory, and systematic drift, thereby reflecting realistic field uncertainties. The resulting percent-change heatmaps summarize the average impact at 10, 20, and 40 cm for every site and noise scenario, enabling a direct assessment of how surface-level uncertainty propagates to subsurface predictions (Figure 6).

In the first set of columns (±1% noise for each type), nearly all sites show almost no change. A 1% perturbation in the top-layer moisture is essentially within the noise floor of the model it causes on the order of 0.05%–0.1% or less change in the deeper moisture predictions on average, which is negligible. This indicates the model's predictions are stable under very small sensor errors. At ±5% noise levels, a slight increase in effect is visible, though still relatively small. Most sites exhibit a change of only ~0.1%–0.3% in predicted moisture at 5% noise. For example, under a ±5% structured drift, the median change in 10 cm moisture prediction across sites is about +0.19%, and under ±5% white noise it's about +0.13%. Indicating that a ±5% input uncertainty does not drastically alter model outputs. The effects become more pronounced in the ±10% noise columns. A ±10% error in surface moisture is quite significant, and accordingly a few sites show more noticeable changes in their deeper moisture predictions. Still, the magnitude of change is modest in percentage terms, for most sites, the average prediction shift is on the order of a few tenths of a percent. In the 10 cm depth panel, many sites have changes around 0.2%–0.8% under ±10% noise. The median impact at 10 cm across all sites is about +0.22% for AR(1) noise and +0.20% for white noise.

Although typical changes are small, the heatmap does highlight a few outlier cases at ±10% noise, where certain sites experienced a bigger shift. In the AR(1) ±10% column, for instance, one site's 10 cm moisture increased by about 3.4% on average, the largest seen, and similarly one site under a ±10% structured drift saw about a −3.3% change at 40 cm (indicating a slight underprediction) on average. These are exceptions, no site shows changes beyond this few-percent range. Importantly, even these "worst-case" sensitivity values are still quite low relative to the ±10% input error; no catastrophic divergences occur. There is no single site that consistently reacts dramatically more than others across all noise scenarios, the sites with the largest change under one type of noise are different from those under another, and overall all sites stay within a comparable range of sensitivity. This suggests no particular site's model is disproportionately fragile; differences in soil or climate at various sites do not produce fundamentally different vulnerability to input errors.



Site	Depth	WN1%	AN1%	SN1%	WN5%	AN5%	SN5%	WN10%	AN10%	SN10%
Site 1	10cm	0.1	0.1	0.1	0.1	0.1	0.0	0.1	0.3	0.4
	20cm	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	-0.1
	40cm	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1
Site 2	10cm	-0.0	-0.0	0.0	-0.0	-0.0	-0.1	-0.0	0.1	0.0
	20cm	-0.0	0.0	0.0	-0.0	-0.1	-0.0	-0.0	0.1	0.0
	40cm	-0.0	0.0	0.0	-0.1	-0.1	-0.0	-0.1	0.0	-0.1
Site 3	10cm	0.2	0.2	0.2	0.3	0.2	0.7	0.4	0.0	-0.5
	20cm	0.0	0.1	0.1	0.1	0.0	0.3	0.2	0.1	-0.1
	40cm	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0
Site 4	10cm	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.4	0.5
	20cm	0.1	0.1	0.1	0.2	0.2	0.2	0.4	0.6	0.2
	40cm	0.2	0.1	0.1	0.6	0.6	0.5	1.2	1.5	0.4
Site 5	10cm	0.2	0.1	0.1	0.3	0.1	0.1	0.6	0.7	0.1
	20cm	0.1	0.1	0.0	0.1	0.1	0.3	0.1	0.1	-0.4
	40cm	-0.0	0.1	-0.0	-0.1	0.2	0.4	-0.4	-0.4	-1.1
Site 6	10cm	-0.0	-0.0	0.1	-0.0	-0.1	0.5	-0.2	-0.4	-0.6
	20cm	-0.0	-0.0	0.1	0.0	-0.0	0.5	-0.0	-0.2	-0.6
	40cm	0.0	0.0	0.1	0.1	0.1	0.5	0.1	-0.2	-0.4
Site 7	10cm	0.2	0.2	0.0	0.3	0.9	-0.1	0.1	0.2	1.1
	20cm	0.2	0.2	0.0	0.3	1.1	-0.1	0.0	0.1	1.3
	40cm	0.3	0.2	0.1	0.4	1.4	0.1	-0.0	0.0	1.2
Site 8	10cm	0.3	0.4	-0.0	1.4	1.6	-0.3	2.4	3.4	1.2
	20cm	0.3	0.3	-0.1	1.0	1.2	-0.5	1.9	2.6	1.1
	40cm	0.3	0.2	-0.0	0.4	0.6	-0.6	1.1	1.3	0.8
Site 9	10cm	-0.1	-0.0	0.2	-0.4	-0.2	0.5	-0.4	-0.1	-0.9
	20cm	-0.1	-0.0	0.1	-0.3	-0.1	0.3	-0.4	-0.1	-0.5
	40cm	-0.0	-0.0	0.0	-0.3	-0.2	0.0	-0.8	-0.5	-0.2
Site 10	10cm	0.0	0.0	-0.0	0.6	0.6	0.1	0.8	0.8	-0.2
	20cm	0.0	-0.0	-0.0	0.9	1.0	1.1	1.7	1.7	0.9
	40cm	-0.0	-0.0	-0.0	1.2	1.3	2.4	2.6	2.4	2.2
Site 11	10cm	0.2	0.1	0.2	0.0	0.3	0.3	0.3	0.0	-0.3
	20cm	0.1	0.0	0.0	-0.2	0.4	0.2	0.2	-0.0	-0.2
	40cm	0.0	-0.0	0.0	-0.3	0.3	0.4	0.1	-0.1	-0.7
Site 12	10cm	0.0	0.0	0.1	-0.4	-0.5	0.9	-0.8	-0.5	-1.4
	20cm	-0.1	-0.0	0.0	-0.3	-0.3	0.5	-0.6	-0.4	-1.0
	40cm	-0.1	-0.0	0.0	-0.3	-0.2	0.4	-0.5	-0.4	-0.8
Site 13	10cm	0.1	0.1	-0.0	-0.2	-0.3	0.4	-0.7	-0.4	-0.4
	20cm	0.1	0.1	-0.0	-0.2	-0.2	0.4	-0.6	-0.4	-0.4
	40cm	0.1	0.1	-0.0	-0.1	-0.1	0.4	-0.1	-0.0	-0.2
Site 14	10cm	0.1	0.0	-0.1	0.1	0.2	0.2	0.2	0.2	1.0
	20cm	0.0	0.1	-0.1	0.1	0.2	0.2	0.3	0.3	0.8
	40cm	-0.0	0.0	0.0	0.1	0.1	-0.2	0.2	-0.0	0.5
Site 15	10cm	0.0	0.0	0.1	0.1	-0.0	0.3	-0.0	0.3	-0.0
	20cm	-0.1	0.1	0.0	0.1	-0.1	-0.3	-0.3	0.1	0.4
	40cm	-0.0	0.0	-0.0	-0.0	-0.1	0.1	-0.3	-0.1	-0.1
Site 16	10cm	0.1	0.0	0.1	0.2	0.4	-0.3	0.6	0.9	-0.2
	20cm	-0.1	-0.0	-0.1	-0.3	-0.2	-0.4	-0.4	-0.3	-1.0
	40cm	-0.3	0.0	-0.4	-1.0	-1.3	-1.7	-1.4	-1.7	-3.3
Site 17	10cm	0.2	0.1	0.1	0.6	0.5	-0.0	1.4	1.1	0.1
	20cm	-0.0	0.0	-0.0	0.2	0.1	0.2	0.7	0.7	0.4
	40cm	-0.0	0.1	-0.0	0.5	0.2	0.2	1.2	1.3	0.4
Site 18	10cm	0.1	0.0	0.0	-0.0	-0.2	-0.1	-0.2	-0.1	-0.5
	20cm	0.2	0.2	-0.0	0.4	0.0	-0.8	0.2	-0.1	-2.5
	40cm	-0.0	-0.0	-0.0	0.1	-0.1	-0.2	-0.2	-0.3	-1.0
Site 19	10cm	-0.0	0.2	0.0	0.5	0.3	0.6	0.6	1.3	0.1
	20cm	0.0	0.4	0.2	0.6	0.2	0.8	0.4	1.2	-0.5
	40cm	0.1	0.4	0.1	0.7	0.7	0.6	1.2	1.7	-0.4
Site 20	10cm	1.3	1.3	1.0	1.9	1.8	0.8	1.8	2.2	-0.7
	20cm	0.6	0.7	0.5	0.9	0.9	0.6	0.7	1.3	-0.1
	40cm	-0.0	0.1	0.1	-0.1	-0.1	0.5	-0.2	0.5	0.7

Figure 6. Heatmap of uncertainty propagation across sites. Cells show the mean percentage change in model output (relative to the noise-free baseline) under additive input perturbations of ±1%, ±5%, ±10% for three stochastic structures: white (WN), AR(1) autocorrelated (AN), and structured (SN). Columns represents noise type, rows denote target depth, and each panel corresponds to a site. The shared top colorbar indicates the global range across all panels.

Across noise types, the model exhibits distinct, interpretable responses. White noise is zero-mean and temporally independent, so its effects manifest as small, symmetric jitters that largely cancel over the evaluation window, yielding near-neutral averages. AR(1) perturbations introduce temporal memory; a sustained positive (or negative) offset tends to propagate into a consistent directional bias in the predictions, producing a clearer shift rather than balanced fluctuations. Structured (sinusoidal) drift alternates between positive and negative phases, so its influence is mixed across time and sites: it increases variability but, when averaged, shows little systematic bias because opposing phases counterbalance. Across the depth, deeper soil moisture predictions are less affected by noise in the surface layer, aligning with physical expectations that deeper soil states are insulated from short-term surface fluctuations.

3.6. Performance at High Temporal Frequency Sites

To further evaluate the robustness of the proposed algorithm, we tested the model performance using high-frequency (10-min interval) soil moisture observations from three sites in Zambia (Bbondo, Kapululira, and Margaret). This test extends earlier evaluations (which used hourly data from the ISMN network) to a finer temporal resolution and a different geographic region. The results indicate an excellent agreement between the model predictions and the in situ measurements (Figure 7). All three Zambian sites exhibit very high coefficients

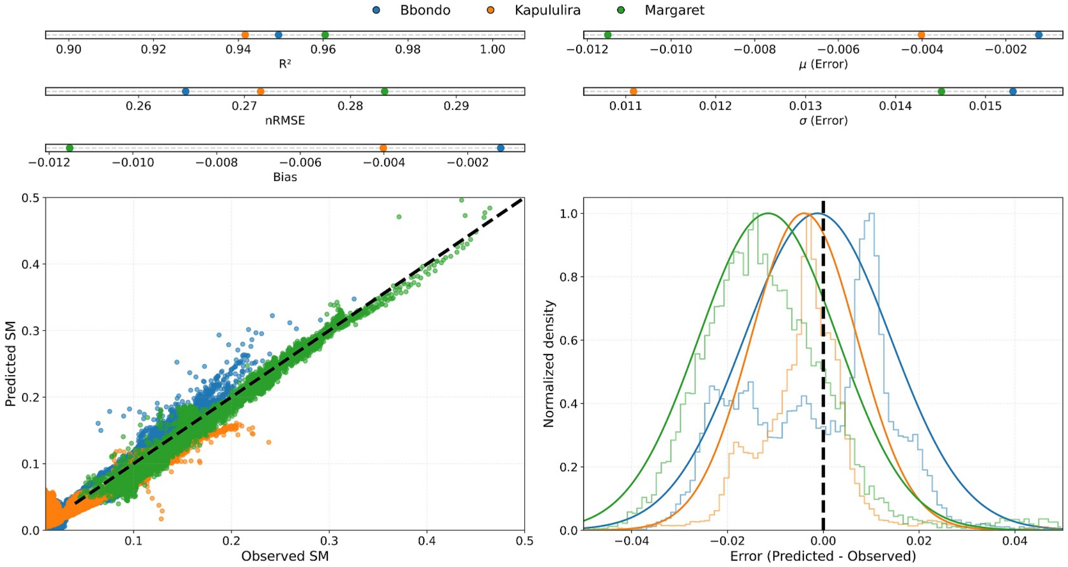


Figure 7. Observed versus predicted soil moisture at 10 cm for 10-min resolution across Bbondo, Kapululira, and Margaret. Left plot: Scatter of ensemble-mean predictions ($N = 100$) against observations with a 1:1 reference; inset panels summarize site-wise performance metrics. Right plot: Normalized error densities with Gaussian fits; inset panels report μ and σ. A zero-error reference and shared axes facilitate cross-site comparison.

of determination ($R^2 \approx 0.94$–0.96). Such R^2 values reflect a strong correlation between simulated and observed values and are slightly higher than the typical performance seen across a global set of 10 cm soil moisture sites (median $R^2 \sim 0.90$). This consistency in high R^2 at all sites demonstrates that the model captures the soil moisture dynamics in Zambia extremely well, even at a 10-min timestep. In addition to the strong correlation, the error metrics are low, highlighting the model's accuracy. The nRMSE ranges from approximately 0.26 to 0.30 across the three stations which are notably lower than the median nRMSE (~0.38) observed at 10 cm in a broad data set of global sites, highlighting the superior performance of our model in the Zambian context. The bias (mean error) is essentially zero for all sites which means the model does not consistently overestimate or underestimate soil moisture at these locations. Among the three stations, Margaret showed a slightly higher nRMSE (0.299) and a small negative bias (−0.013) compared to Bbondo (nRMSE ~0.264, bias ~−0.001) and Kapululira (nRMSE ~0.272, bias ~−0.004). However, these differences are marginal. Overall, all three sites maintain high accuracy, with the error statistics confirming that the model's predicted soil moisture closely matches the measured data in both magnitude and temporal patterns.

Operationally, these results imply that at a 10-min timestep the model achieves $R^2 \approx 0.94$–0.96, nRMSE ≈ 0.26–0.30, and essentially zero bias across all three sites, indicating reliable high-frequency tracking of soil moisture fluctuations. Given that the model requires only surface moisture as input and has low computational latency, it is feasible for near–real-time deployment in monitoring or early-warning systems where rapid updates are essential.

4. Conclusion

This study introduced a weakly physics-constrained, multi-agent diffusion model for subsurface soil moisture estimation by leveraging surface soil moisture information. It is designed to balance the flexibility of data-driven approaches with the physical consistency of constraints. By incorporating soft guidance terms that encourage smooth flux transitions and feasible soil moisture ranges, the model avoids dependence on site-specific hydraulic parameters while retaining hydro-physical consistency. The multi-agent structure, with specialized agents for dry, intermediate, and wet conditions, further enhances robustness in heterogeneous environments.

Evaluation across a globally distributed data set of 20 ISMN stations (hourly resolution) and three high-frequency sites in Zambia (10-min resolution) demonstrates the scalability and generalizability of the framework. At 10 cm depth, the model exhibits consistently high accuracy (median $R^2 \approx 0.91$, nRMSE ≈ 0.37), with marginal bias and stable performance across climates and soil types. Performance declines at greater depths (20 and 40 cm), reflecting the inherently buffered and lagged nature of deeper soil moisture dynamics, but predictive performance remains significant across most sites. High-frequency testing in Zambia confirmed the model's ability to capture soil moisture variability at fine temporal scales, with $R^2 > 0.94$ and low nRMSE.

Stochastic robustness experiments with 30 random seeds showed that site-level results are generally stable, though a few sites with limited data or extreme soil conditions exhibited larger variability. Input perturbation tests confirmed that the model is resilient to realistic levels of sensor noise, with only minor propagation of uncertainty to subsurface predictions. Together, these findings highlight that weak physics constraints combined with diffusion-based stochastic modeling offer a scalable, accurate, and physically consistent approach to subsurface soil moisture estimation.

A key limitation of the present study is that the model is evaluated only for temporal prediction at sites included in the training set. While the multi-agent design and weak physics guidance enhance robustness across depths, the framework has not yet been tested for its ability to generalize spatially to entirely unseen locations. The 20 sites used here, although spanning contrasting climates and soil conditions, may be insufficient to fully characterize spatial variability, and thus do not guarantee out-of-sample spatial transferability. Future work will examine cross-site and cross-region generalization to assess the model's applicability for true spatial prediction scenarios. Despite these limitations, the framework represents a promising step toward transferable machine learning tools for hydrology, helping bridge the gap between purely empirical models and physics-based formulations.

Conflict of Interest

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

Data Availability Statement

In situ data can be downloaded from Zenodo [A. Singh, Singh, & Kumar (2025b)]. The code developed for this study can be downloaded from Zenodo [A. Singh, Singh, & Kumar (2025a)].

Acknowledgments
We gratefully acknowledge the Geohydrology Observatory Facility at IISERB Bhopal, where the experiments were conducted, which played a key role in conceptualizing this study. A.S. thanks the Zambia Meteorological Department (ZMD) for providing the gauge station data used in this study, accessed through the University of Leeds. K.G. would like to acknowledge an institutional grant received from IISERB.

References

Agyeman, B. T., Orouskhani, E., Naouri, M., Appels, W. M., Wolleben, M., Liu, J., & Shah, S. L. (2025). Sensitivity-informed parameter selection for improved soil moisture estimation from remote sensing data. *Control Engineering Practice*, *165*, 106593. https://doi.org/10.1016/j.conengprac.2025.106593

Ahmad, J. A., Forman, B. A., & Kumar, S. V. (2022). Soil moisture estimation in south Asia via assimilation of SMAP retrievals. *Hydrology and Earth System Sciences*, *26*(8), 2221–2243. https://doi.org/10.5194/hess-26-2221-2022

Amirkhani, A., & Barshooi, A. H. (2022). Consensus in multi-agent systems: A review. *Artificial Intelligence Review*, *55*(5), 3897–3935. https://doi.org/10.1007/s10462-021-10097-x

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

Chavoshi, A., Dashtian, H., Bakhshian, S., Young, M. H., & Niyogi, D. (2024). Pinn-sm: A physics-informed neural networks model for vadose zone soil moisture profile prediction. *Authorea Preprints*. https://doi.org/10.22541/essoar.173437284.48336455/v1

Dente, L., Su, Z., & Wen, J. (2012). Validation of SMOS soil moisture products over the Maqu and Twente regions. *Sensors*, *12*(8), 9965–9986. https://doi.org/10.3390/s120809965

Dinesh, D., Kumar, S., & Saran, S. (2024). Machine learning modelling for soil moisture retrieval from simulated NASA-ISRO SAR (NISAR) l-band data. *Remote Sensing*, *16*(18), 3539. https://doi.org/10.3390/rs16183539

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 Discussions*, *25*(11), 5749–5804. https://doi.org/10.5194/hess-25-5749-2021

Dorigo, W., 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

Draper, C., & Reichle, R. (2015). The impact of near-surface soil moisture assimilation at subseasonal, seasonal, and inter-annual timescales. *Hydrology and Earth System Sciences*, *19*(12), 4831–4844. https://doi.org/10.5194/hess-19-4831-2015

Gruber, A., Dorigo, W. A., Zwieback, S., Xaver, A., & Wagner, W. (2013). Characterizing coarse-scale representativeness of in situ soil moisture measurements from the international soil moisture network. *Vadose Zone Journal*, *12*(2), vzj2012-0170. https://doi.org/10.2136/vzj2012.0170

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840–6851.

Jing, Y., Li, Y., Li, X., Lin, L., She, X., Jiang, M., & Shen, H. (2024). An integrated learning framework for seamless high-resolution soil moisture estimation. *IEEE Transactions on Geoscience and Remote Sensing*, *62*(4210917), 1–17. https://doi.org/10.1109/tgrs.2024.3461717

Lawston, P. M., Santanello, J. A., Jr., & Kumar, S. V. (2017). Irrigation signals detected from SMAP soil moisture retrievals. *Geophysical Research Letters*, *44*(23), 11860–11867. https://doi.org/10.1002/2017gl075733

Lebel, T., Cappelaere, B., Galle, S., Hanan, N., Kergoat, L., Levis, S., et al. (2009). Amma-catch studies in the Sahelian region of west-Africa: An overview. *Journal of Hydrology*, *375*(1–2), 3–13. https://doi.org/10.1016/j.jhydrol.2009.03.020

Li, W., Duveiller, G., Wieneke, S., Forkel, M., Gentine, P., Reichstein, M., et al. (2024). Regulation of the global carbon and water cycles through vegetation structural and physiological dynamics. *Environmental Research Letters*, *19*(7), 073008. https://doi.org/10.1088/1748-9326/ad5858

Liu, R., Lu, L., Ge, Y., Dong, L., & Zhou, J. (2024). An approach for multi-depth soil moisture prediction in alfalfa based on a dual-branch combined model. *Computers and Electronics in Agriculture*, *222*, 109038. https://doi.org/10.1016/j.compag.2024.109038

Malik, I., & Mishra, V. (2025). On the utility of soil moisture for monitoring and prediction of compound hot and dry extremes in India. *Water Resources Research*, *61*(10), e2025WR040174. https://doi.org/10.1029/2025wr040174

Nguyen, H. H., Kim, H., & Choi, M. (2017). Evaluation of the soil water content using cosmic-ray neutron probe in a heterogeneous monsoon climate-dominated region. *Advances in Water Resources*, *108*, 125–138. https://doi.org/10.1016/j.advwatres.2017.07.020

Rahmati, M., Amelung, W., Brogi, C., Dari, J., Flammini, A., Bogena, H., et al. (2024). Soil moisture memory: State-of-the-art and the way forward. *Reviews of Geophysics*, *62*(2), e2023RG000828. https://doi.org/10.1029/2023rg000828

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Singh, A., & Gaurav, K. (2024). Piml-sm: Physics-informed machine learning to estimate surface soil moisture from multi-sensor satellite images by leveraging swarm intelligence. *IEEE Transactions on Geoscience and Remote Sensing*, *62*(4416913), 1–13. https://doi.org/10.1109/tgrs.2024.3502618

Singh, A., Niranjannaik, M., & Gaurav, K. (2025). Overcoming data scarcity: A transfer learning framework with fine-tuned neural networks and multi-sensor satellite image fusion for soil moisture estimation. *Engineering Applications of Artificial Intelligence*, *159*, 111636. https://doi.org/10.1016/j.engappai.2025.111636

Singh, A., Singh, V., & Gaurav, K. (2025a). Leveraging neural operator and sliding window technique for enhanced subsurface soil moisture imputation under diverse precipitation scenarios. *Journal of Geophysical Research: Machine Learning and Computation*, *2*(3), e2025JH000730. https://doi.org/10.1029/2025jh000730

Singh, A., Singh, V., & Gaurav, K. (2025b). Physics-aware probabilistic modeling of subsurface soil moisture using diffusion processes across different climate settings. *Geophysical Research Letters*, *52*(20), e2025GL118607. https://doi.org/10.1029/2025gl118607

Singh, A., Singh, V., & Kumar, G. (2025a). Physics-informed multi-agent diffusion model for subsurface soil moisture estimation across contrasting climate and soil conditions [code]. *Zenodo*. https://doi.org/10.5281/zenodo.17012578

Singh, A., Singh, V., & Kumar, G. (2025b). Physics-informed multi-agent diffusion model for subsurface soil moisture estimation across contrasting climate and soil conditions [Datasets]. *Zenodo*. https://doi.org/10.5281/zenodo.17012472

Singh, V., Verma, D., Singh, A., & Gaurav, K. (2025). Multi-depth soil moisture dynamics to rainfall events: An automated machine learning approach. *Hydrology and Earth System Sciences Discussions*, *2025*, 1–40. https://doi.org/10.5194/egusphere-2025-961,2025

Su, Z., Wen, J., Dente, L., Van Der Velde, R., Wang, L., Ma, Y., et al. (2011). The Tibetan Plateau observatory of plateau scale soil moisture and soil temperature (Tibet-obs) for quantifying uncertainties in coarse resolution satellite and model products. *Hydrology and Earth System Sciences*, *15*(7), 2303–2316. https://doi.org/10.5194/hess-15-2303-2011

Tian, J., Han, Z., Bogena, H. R., Huisman, J. A., Montzka, C., Zhang, B., & He, C. (2019). Estimation of subsurface soil moisture from surface soil moisture in cold mountainous areas. *Hydrology and Earth System Sciences Discussions*, *2019*, 1–32.

Tripathy, K. P., Mishra, A. K., Kumar, S. V., & Miralles, D. G. (2025). Lagged soil moisture controls on the persistence of drought and heatwaves in the United States. *Geophysical Research Letters*, *52*(22), e2025GL115811. https://doi.org/10.1029/2025gl115811

van der Velde, R., Benninga, H.-J. F., Retsios, B., Vermunt, P. C., & Salama, M. S. (2023). Twelve years of profile soil moisture and temperature measurements in Twente, the Netherlands. *Earth System Science Data*, *15*(4), 1889–1910. https://doi.org/10.5194/essd-15-1889-2023

Verma, S., & Nema, M. K. (2022). Development of an empirical model for sub-surface soil moisture estimation and variability assessment in a lesser Himalayan Watershed. *Modeling Earth Systems and Environment*, *8*(3), 3487–3505. https://doi.org/10.1007/s40808-021-01316-z

Wu, L. (2003). Soil water diffusion. *Encyclopedia of Water Science*, 865–867.

Yang, K., Qin, J., Zhao, L., Chen, Y., Tang, W., Han, M., et al. (2013). A multiscale soil moisture and freeze–thaw monitoring network on the third pole. *Bulletin of the American Meteorological Society*, *94*(12), 1907–1916. https://doi.org/10.1175/bams-d-12-00203.1

Zhang, M., Ge, Y., & Wang, J. (2024). Integrating infiltration processes in hybrid downscaling methods to estimate sub-surface soil moisture. *Ecological Informatics*, *84*, 102875. https://doi.org/10.1016/j.ecoinf.2024.102875