ADVANCING EARTH AND SPACE SCIENCES

Geophysical Research Letters

RESEARCH LETTER
10.1029/2026GL124661

How Important Are the Critical Points in Selecting the Optimal Samples for Accurate Estimation of Subsurface Soil Moisture?

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

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

Key Points:

- Physics-aware sampling based neural operator to derive the minimal training subset that achieves reliable predictive accuracy
- Uncertainty- and distribution-based methods perform best at monsoon influenced subarctic stations and cold, arid sites, respectively
- Sampling decisions exhibit strong seasonal and diurnal controls

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, V., Singh, A., & Gaurav, K. (2026). How important are the critical points in selecting the optimal samples for accurate estimation of subsurface soil moisture? *Geophysical Research Letters*, *53*, e2026GL124661. https://doi.org/10.1029/2026GL124661

Received 28 MAY 2026
Accepted 29 JUN 2026

Author Contributions:

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

Abstract We propose a novel physics-aware sampling framework to extract the most informative training instances (critical points) from the conventional 70% training data set using six diversified sampling strategies. These critical points are used to train a spectral Fourier Neural Operator (FNO) augmented with a lagged term to predict subsurface soil moisture (20 and 40 cm) from near-surface observations (5 cm). This approach is evaluated at eight stations of the International Soil Moisture Network spanning two climatic regimes (Dwc: monsoon-influenced subarctic; and BWk: arid and cold desert) with hourly soil moisture measurements available at these depths. We demonstrate that judiciously selected training subsets achieve predictive accuracy comparable or even superior to the full-training-data baseline model. Uncertainty-based sampling is optimal for Dwc stations (with 10% training data). Distribution-based selection excels at BWk sites incorporating training proportions between 10% and 62% with minimal nRMSE. Temporal analysis of the selection of critical points reveals significant climatic controls.

Plain Language Summary Soil moisture plays an important role in soil and crop health monitoring, water resource management, and environmental and weather applications. However, measuring moisture in deeper layers is expensive and often limited by poor data availability. We designed a novel method that carefully selects only the most important instances (critical points) from soil moisture records using six different selection methods. We then used these data points to develop a model to estimate deeper soil moisture from surface measurements. Tested at eight sites with different climates, we found that critical samples help the model achieve accuracy comparable to that of the traditional approach while using significantly less data points. This means that we can effectively monitor the conditions of the underground soil with fewer measurements, saving time and resources. This intelligent data selection makes soil monitoring more practical and affordable, especially in regions where collecting long-term data is difficult, making it useful for both scientists and farmers.

1. Introduction

Subsurface soil moisture is a key component of terrestrial water, energy, and carbon cycles, strongly regulating land-atmosphere interactions (Liao et al., 2024; Wu et al., 2024; A. Singh et al., 2023). Recent machine learning approaches have improved subsurface soil moisture estimation (Chavoshi et al., 2025; A. Singh, Singh, & Gaurav, 2025b; A. Singh et al., 2026), overcoming several limitations of empirical, semi-empirical, and physically based models (Draper & Reichle, 2015; Verma & Nema, 2022; Zhang et al., 2024). For instance, Chavoshi et al. (2025) developed a physics-informed neural network (PINN) constrained by Richards' equation using time, depth, precipitation, vegetation indices, air temperature, wind speed, and cumulative precipitation as inputs. However, this approach requires site-specific hydraulic parameters that are difficult to obtain and generalize. To address this limitation, A. Singh, Singh, & Gaurav (2025b) proposed a physics-aware diffusion framework with weak physical constraints, removing the need for explicit soil hydraulic parameters and relying mainly on remotely available surface soil moisture observations (A. Singh et al., 2026; A. Singh & Gaurav, 2024; Xu et al., 2022; Ahmad et al., 2022; Li et al., 2021).

Despite these advances, learning-based models remain highly dependent on the quantity and representativeness of training data (Hu et al., 2025; A. Singh et al., 2025). Most studies use large training fractions, typically ≥60% of the available time series, to capture seasonal variability and event-scale dynamics (V. Singh, Verma, et al., 2025;

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

Houben et al., 2025; Wang et al., 2023). This requirement is challenging in practice because long-term, high-resolution subsurface observations are often limited by installation cost, sensor degradation, calibration drift, and data gaps, especially in remote or harsh environments. This raises an important question: can robust predictive models be developed using substantially reduced but strategically selected training samples?

To address this gap, we propose a domain-informed, physics-aware sampling framework that replaces the conventional fixed training fraction, for example, 70%, with a minimal yet dynamically informative subset of critical points. We used six different sampling strategies to extract important temporal patterns from near-surface soil moisture observations. These selected samples are then used with a spectral Fourier Neural Operator (FNO) to predict subsurface soil moisture at multiple depths. Here, "active learning" is formulated mathematically rather than as a conventional model-in-the-loop process. Training points are ranked using information-theoretic and statistical measures, including variance, entropy, spectral energy, distributional coverage, and predictive uncertainty, to identify samples with high informational density. These criteria are then linked to soil moisture processes such as wetting events, regime transitions, frequency attenuation with depth, and uncertainty under heterogeneous hydrological states. Thus, the framework combines mathematical methods with physics-aware sampling.

The framework is evaluated across eight International Soil Moisture Network (ISMN) stations representing contrasting hydroclimatic regimes, using hourly in situ observations. This study addresses the following research questions:

- Can a strategically selected subset of critical points achieve predictive skill comparable to models trained using the conventional 70% data fraction?
- How do hydroclimatic regimes influence the effectiveness of different sampling strategies, and do certain strategies perform better under specific climatic conditions?
- How do dominant monsoonal and diurnal processes control sampling relevance and model performance?

2. In Situ Observations

We used in situ soil moisture observations acquired from eight long-term monitoring stations (Figure S1 in the Supporting Information S1) within the ISMN (Dorigo et al., 2011, 2021; Gruber et al., 2013) distributed across three major networks in China (Dente et al., 2012; Su et al., 2011). The stations include MAQU_NST-24, NAQU_NQ01, NAQU_NQMS, NAQU_NQNorth, NGARI_SQ05, NGARI_SQ08, NGARI_SQ09, and NGARI_SQ14 (listed alphabetically). The naming convention consists of the network name followed by the station identifier (e.g., MAQU_NST-24, where "MAQU" is the network, and "NST-24" is the station ID). Based on the Köppen Geiger climate classification, MAQU and NAQU stations are located in regions with a cold, dry winter and cool summer climate (Dwc), and the NGARI stations are characterized by an arid, cold-desert environment (BWk) (Beck et al., 2023). The duration of monsoonal influence extends from April-October at MAQU_NST-24, May-October across NAQU stations, and June-September across NGARI sites. Soil moisture records are analyzed at standardized depths of 5, 20, and 40 cm. The selected depths offer a representative vertical sampling column to capture near-surface and subsurface soil moisture dynamics. The selected sites provide continuous hourly measurements of soil moisture data spanning a minimum of 3 years which prolongs up to a decade at certain stations (Figure S1a in Supporting Information S1). MAQU_NST-24 employs an ECH2O EC-TM sensor, while the remaining stations use 5TM sensors, all based on Frequency Domain Reflectometry (FDR). Collectively, these stations encompass a broad scope of hydroclimatic regimes, land cover types, and soil textural compositions (Figures S1b and S1c in Supporting Information S1). This diversity provides an effective testbed for evaluating model robustness and transferability across heterogeneous environmental settings.

3. Model Development

We develop a physics-aware sampling strategy to identify critical points (or samples) from time-series data. The critical points are instantaneous data instances representing distinct soil moisture regimes at 5 cm depth which can best estimate the soil moisture distribution at deeper soil depths (20 and 40 cm). These samples are then used to train a predictive model, aiming to achieve comparable or better accuracy than the conventional approach of using a large contiguous training set (e.g., the first 70% of the data). Our methodology consists of two components: (a) selective sampling of informative time points (critical points) based on multiple criteria that capture the underlying physics and dynamics of the system, and (b) training a spectral FNO model with a lagged input term on those selected critical points.

Consider a univariate time series $x(t)$ representing a measured physical variable (e.g., soil moisture at the surface layer) and a target time series $y(t)$ representing a related variable (e.g., soil moisture at a deeper layer) that we aim to estimate. We have observations $\{x_i, y_i\}_{i=1}^{N}$ over time (with i indexing discrete time steps). A conventional training approach might use the first 70% of the sequence ($i = 1$ to $i = N_{\text{train}}$ with $N_{\text{train}} \approx 0.7N$) to train a model and reserve the remainder for testing. In our approach, we do not use all N_{train} points for training. Instead, we identify a subset of "critical" time indices $S \subset \{1, \ldots, N_{\text{train}}\}$ that are particularly informative about the system's behavior. We employ several sampling strategies to score or rank each candidate time point in the training period by how "informative" or "critical" it is. Each strategy is motivated by physical insights into the system's dynamics. These include variance-, entropy-, Fourier-, distribution-, uncertainty-based, and combined sampling methods as described below;

- *Variance-based selection*: High variance in the input signal indicates rapid changes or transient events, such as rainfall-induced increases in soil moisture or rapid dry-downs. To capture these dynamics, we compute the local variance of $x(t)$ over a sliding window of length N_v, for example, 96 time steps. For each time index i, the window is defined as $W_i = \{i - N_v + 1, \ldots, i\}$, and the variance is computed using Equation 1:

$$\sigma^2(i) = \frac{1}{N_v} \sum_{j=i-N_v+1}^{i} (x_j - \bar{x}_i)^2, \tag{1}$$

where

$$\bar{x}_i = \frac{1}{N_v} \sum_{j=i-N_v+1}^{i} x_j \tag{2}$$

is the local window mean. Larger $\sigma^2(i)$ values indicate stronger fluctuations around the mean. Points are therefore ranked in descending order of $\sigma^2(i)$, so that periods with the highest variability are selected first. Physically, these points often correspond to important hydrological events, including wetting pulses and rapid drying phases. Selecting them ensures that the model is exposed to sharp changes and extremes that are critical for learning soil moisture response dynamics.

- *Entropy-based selection*: Irregular or unpredictable periods in natural systems can be identified using signal entropy. We compute Shannon entropy as defined in Equation 3 over a moving window of length N_e, for example, 96 points, ending at time index i. The values $\{x_j : j \in W_i\}$ are binned into B intervals, and the corresponding probabilities $p_{i,b}$ are estimated.

$$H(i) = -\sum_{b=1}^{B} p_{i,b} \ln p_{i,b} \tag{3}$$

Here, $H(i)$ measures the uncertainty or spread of recent values in the time series. Higher entropy indicates that the window contains more diverse, widely distributed, or multimodal values rather than a single dominant state. Points are ranked by $H(i)$, and high-entropy points are prioritized because they represent complex system behavior. Physically, such windows may correspond to transition periods, where soil moisture shifts between wet and dry states due to intermittent inputs. Including these points exposes the model to rapid regime changes and improves its ability to learn complex dynamics.

- *Spectral (Fourier) energy-based selection*: Soil moisture dynamics exhibit characteristic temporal frequency components, including diurnal and seasonal fluctuations, along with higher-frequency responses to transient rainfall events. To ensure that the model captures these dynamics, we select points associated with high spectral energy in $x(t)$. A short-time Fourier transform is applied over sliding windows. For a window of length N_f ending at time index i, let $\hat{X}_i(\omega_k)$ denote the Fourier coefficient at frequency ω_k. The spectral energy is defined according to Equation 4:

$$E(i) = \sum_{k=0}^{N_f - 1} \left| \hat{X}_i(\omega_k) \right|^2 \tag{4}$$

In practice, this can be computed efficiently using the real Fast Fourier Transform. Time indices are then ranked by $E(i)$, with higher values indicating windows containing strong oscillatory or transient behavior. Physically, large spectral energy may represent periodic soil moisture variability, such as diurnal evaporation cycles, or abrupt wetting events that generate broadband frequency content. Including these points helps the model learn the system's frequency response, including the attenuation of high-frequency surface fluctuations with depth.

- *Distribution-based (stratified) selection*: To ensure that the model observes the full range of input conditions, including physically important extremes, we use stratified sampling over the distribution of $x(t)$. The range of x is divided into D bins of equal width, and each training index i is assigned to the bin containing x_i. Let $B_1, B_2, \ldots, B_D$ denote the corresponding index sets, sorted by time or value. An ordered sequence is then constructed in a round-robin manner by repeatedly selecting one index from each non-empty bin until all bins are exhausted.

This produces an ordering that approximately balances coverage across the value distribution, ensuring that dry, wet, and intermediate soil moisture states are all represented. This prevents the model from being biased toward moderate conditions and improves its ability to learn from rare but important extremes, where prediction errors are often largest.

- *Uncertainty-based selection*: We use model uncertainty to identify critical points, assuming that highly uncertain predictions correspond to complex, rare, or poorly represented regions of the input space. A preliminary Random Forest (RF) regressor is trained on a small stratified subset of the training data, for example, 10% of the points stratified by input value, to predict y from x and, where applicable, recent input history. The RF consists of M decision trees $\{f_m\}_{m=1}^{M}$. For each training point i, we estimate ensemble disagreement using Equation 5:

$$\text{Uncert}(i) = \sqrt{\frac{1}{M} \sum_{m=1}^{M} f_m(x_i)^2 - \left(\frac{1}{M} \sum_{m=1}^{M} f_m(x_i) \right)^2} \tag{5}$$

Here, $\text{Uncert}(i)$ represents the standard deviation of predictions across the RF trees. Points are ranked in descending order of uncertainty, and the most uncertain points are selected. Physically, high uncertainty can indicate unusual or nonlinear conditions, such as a sharp wetting front or an atypical soil moisture profile across depths, where different trees produce inconsistent predictions. Including such points provides an active-learning-like mechanism that guides the final model toward regions where prediction errors are likely to be largest.

- *Combined critical sampling*: Each sampling strategy described above captures a different aspect of informativeness, including variability, entropy, frequency content, state coverage, value extremes, and model uncertainty. Since these criteria partly overlap but also identify unique points, we combine them using a voting scheme to obtain a robust, physics-aware training set. For each method m (variance, entropy, Fourier, distribution, and uncertainty), the top ϕ fraction of ranked points is selected as its critical candidate set C_m, for example, the top $\phi = 50\%$ points. The combined critical set is then defined as the points selected by at least V methods (Equation 6):

$$C_{\text{combined}} = \left\{ i : \sum_{\text{method } m} \mathbf{1}\{i \in C_m\} \geq V \right\}, \tag{6}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. In our implementation, we used $V = 4$ out of total methods. Thus, C_{combined} contains time indices consistently flagged as important across multiple criteria.

These consensus points form the "head" of the final ordering. The remaining points are appended as the "tail," sorted by an aggregate score such as the average rank across methods. This produces a complete ordered list of training points, with the most universally critical samples placed first. Physically, this strategy prioritizes points

10.1029/2026GL124661

that are important from several perspectives. For example, a sharp rain-induced spike in $x(t)$ may exhibit high variance, entropy, spectral energy, and model uncertainty, causing it to receive multiple votes and enter C_{combined}. Therefore, the combined strategy captures extreme events, regime shifts, and complex dynamics within a single sampling framework.

To model the relationship between the input time series $x(t)$ and target series $y(t)$, we use FNO, a deep learning architecture designed to learn mappings between function spaces and solution operators of partial differential equations using spectral representations (Azizzadenesheli et al., 2024; Cao et al., 2024; Wen et al., 2022). This choice is motivated by the physical nature of soil moisture propagation, where surface conditions are mapped to subsurface responses through an underlying operator (A. Singh, Singh, & Gaurav, 2025a; Wang et al., 2024). The spectral formulation of FNO is well suited for capturing multi-scale temporal patterns and frequency-dependent behavior common in physical systems (Azizzadenesheli et al., 2024; Liu et al., 2023; Wen et al., 2022).

At each time step t, the model uses a window containing the L most recent normalized input values together with a one-step lag. The window is transformed into the frequency domain, where the first K low-frequency modes are linearly transformed with complex parameters to capture temporal scales and suppress high-frequency noise. Reverted to the time domain, this signal is combined with linear current/lagged inputs and a bias to represent multi-scale dynamics, short-term dependencies and transfer characteristics. Model parameters are trained using MSE loss optimized with Adam (Kingma & Ba, 2014), followed by evaluating on a 30% test set using normalized RMSE, Pearson correlation coefficient, and bias.

The FNO is trained only on selected samples $\{x_i, y_i : i \in S\}$ obtained from each sampling strategy and compared with a baseline FNO trained on the full training data set (N_{train}). We further conduct an incremental analysis by increasing the selected training subset by 1% steps from an initial 10% data until baseline performance is matched or exceeded or until the complete baseline training set is used (70%). This quantifies how effectively each strategy reduces data requirements while maintaining predictive accuracy. Details of the FNO architecture are provided in Section S1 of Supporting Information S1.

4. Results and Discussion

4.1. Model Performance

We evaluated key performance metrics, including the correlation coefficient (R), normalized root-mean-square error (nRMSE), and bias, for each model scenario combining the sampling strategy with the FNO across eight ISMN stations at two soil depths (20 and 40 cm), using surface soil moisture as reference (Table S2 in Supporting Information S1). We further explain the model performance based on the test nRMSE values at both depths (Figure 1). For each approach, the nRMSE is computed using the incremental analysis. To maintain clarity and consistency, the performance at each ISMN site and corresponding subsurface depths is denoted using only the station ID followed by soil depth in parentheses. For instance, the site MAQU_NST-24 is represented as NST-24 (20) and NST-24(40) for depths of 20 and 40 cm, respectively.

At station NST-24(20), the uncertainty-based strategy most efficiently surpasses the baseline nRMSE threshold of 0.628 (black dashed line) which requires only 10% of the training data to attain an nRMSE of 0.326. The variance-, entropy-, and distribution-based methods demonstrate comparable performance utilizing 18%, 22%, and 61% of the baseline subset, showing nRMSE values of 0.625, 0.622, and 0.627, respectively. In contrast, the Fourier-based and combined methods applied the full training data set to meet the baseline performance. At NST-24(40), the variance-, entropy-, combined, and distribution-based strategies exhibit substantial improvement, achieving baseline accuracy with 10%, 10%, 28%, and 54% of the training data, respectively.

Across the remaining stations, sampling performance varies with site characteristics and depth, but several consistent patterns emerge. At NQ01, all methods utilize the complete baseline training subset, although uncertainty-based sampling shows improved efficiency at deeper depth. This method frequently performs best, notably at NQMS and NQNorth stations, where it consistently achieves baseline performance with minimal training data across both depths. In contrast, distribution-based sampling is most effective at stations SQ05, SQ09, and SQ14, where it consistently outperforms other methods. At SQ08, multiple strategies perform well, with uncertainty- and Fourier-based strategies achieving robust results at shallow and deeper layers, respectively. At this site, combined sampling shows improved efficiency at greater depths, whereas uncertainty and distribution-

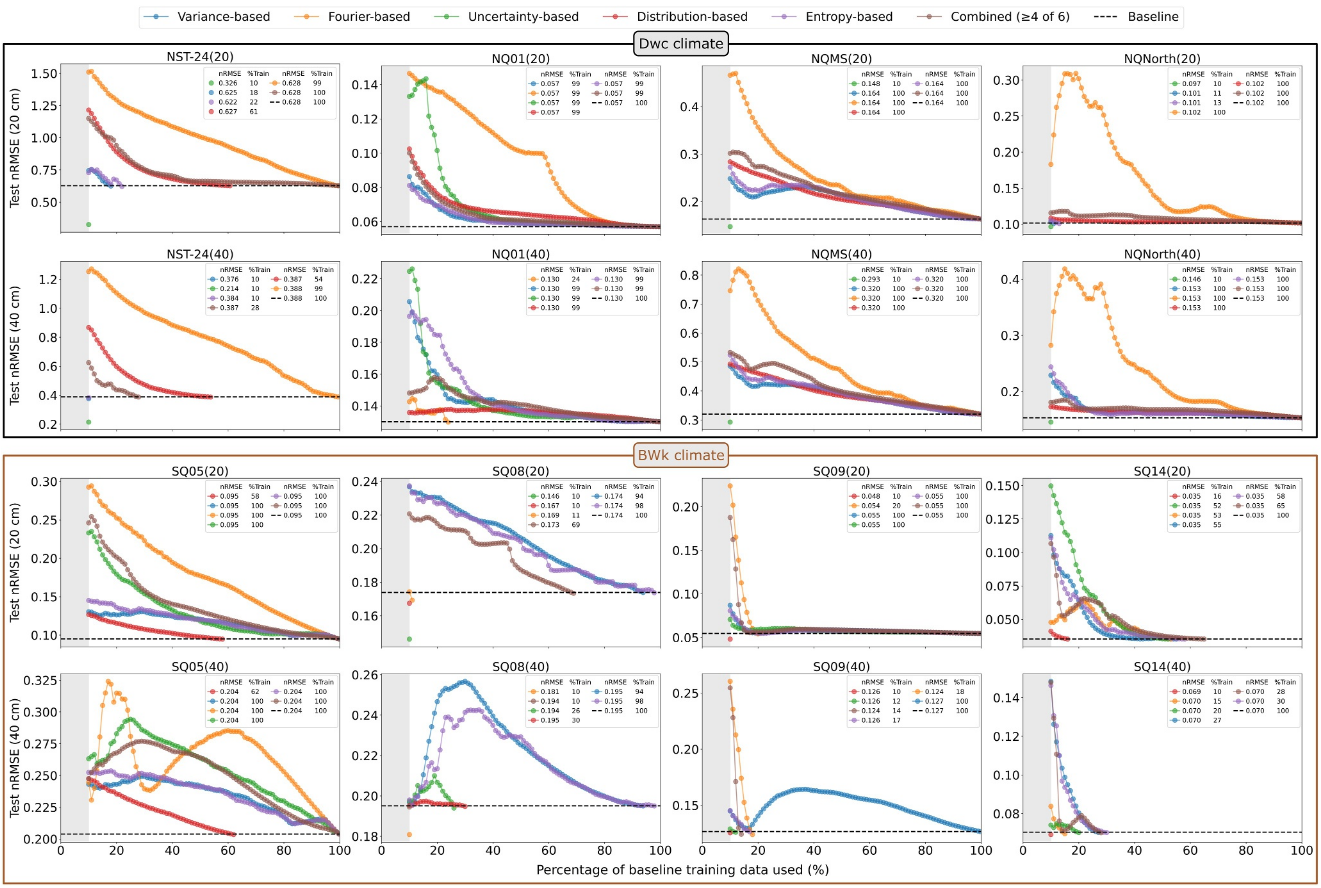

Figure 1. Performance comparison of sampling strategies across eight stations for subsurface soil moisture prediction at 20 and 40 cm soil depths. The plots present the variation in test normalized root-mean-square error (nRMSE) as a function of the proportion of baseline training data (% train) employed by each selection method. The baseline (dashed black line) denotes the reference model performance derived from the complete chronological training data set (70%). The gray shaded region marks the initial phase wherein a minimum of 10% of the baseline training data is utilized by each method. Beyond this threshold, the trajectories illustrate efficiency of each method in attaining or surpassing the baseline nRMSE with reduced training data fractions. Inset panels within each subplot provide station-specific summaries of model performance across depths and sampling approaches.

based methods exhibit increments up to 20%, indicating depth-sensitive variability in information gain. The corresponding selection of critical training points illustrating how each sampling strategy behaves under the varying moisture regime characteristic of this site is detailed in Section S2 in Supporting Information S1, accompanied by a comprehensive visual example (Figure S2 in Supporting Information S1). Other strategies like Fourier-based, followed by combined, uncertainty-, entropy-, and variance-based also show robust performance, with a general reduction in data requirements at deeper layer. This suggests improved representativeness and reduced redundancy at greater depths.

Our findings emphasize that the sampling efficiency is strongly modulated by climatic forcing. Uncertainty-based strategy proves most effective at the MAQU and NAQU stations, attaining benchmark accuracy with 10% of the baseline data. This indicates that the model benefits most from training points representing rapid transitions and high predictive ambiguity. Such conditions are common in snow-affected climates with pronounced seasonal cycles (Dwc). Conversely, at the NGARI stations (BWk), distribution-based method performs best, requiring only 10%–62% of the baseline training subset. This suggests that dry regions experiencing infrequent precipitation where temporal variability is slightly muted between monsoonal months with short-term wetting peaks, it becomes crucial to ensure uniform coverage of the full value range than targeting uncertainty peaks. Combined sampling further exemplifies depth-stratified efficiency across most sites, demonstrating maximal training fraction reductions. This highlights the need for climate specific optimization of sampling design to ensure robust and data efficient soil moisture prediction.

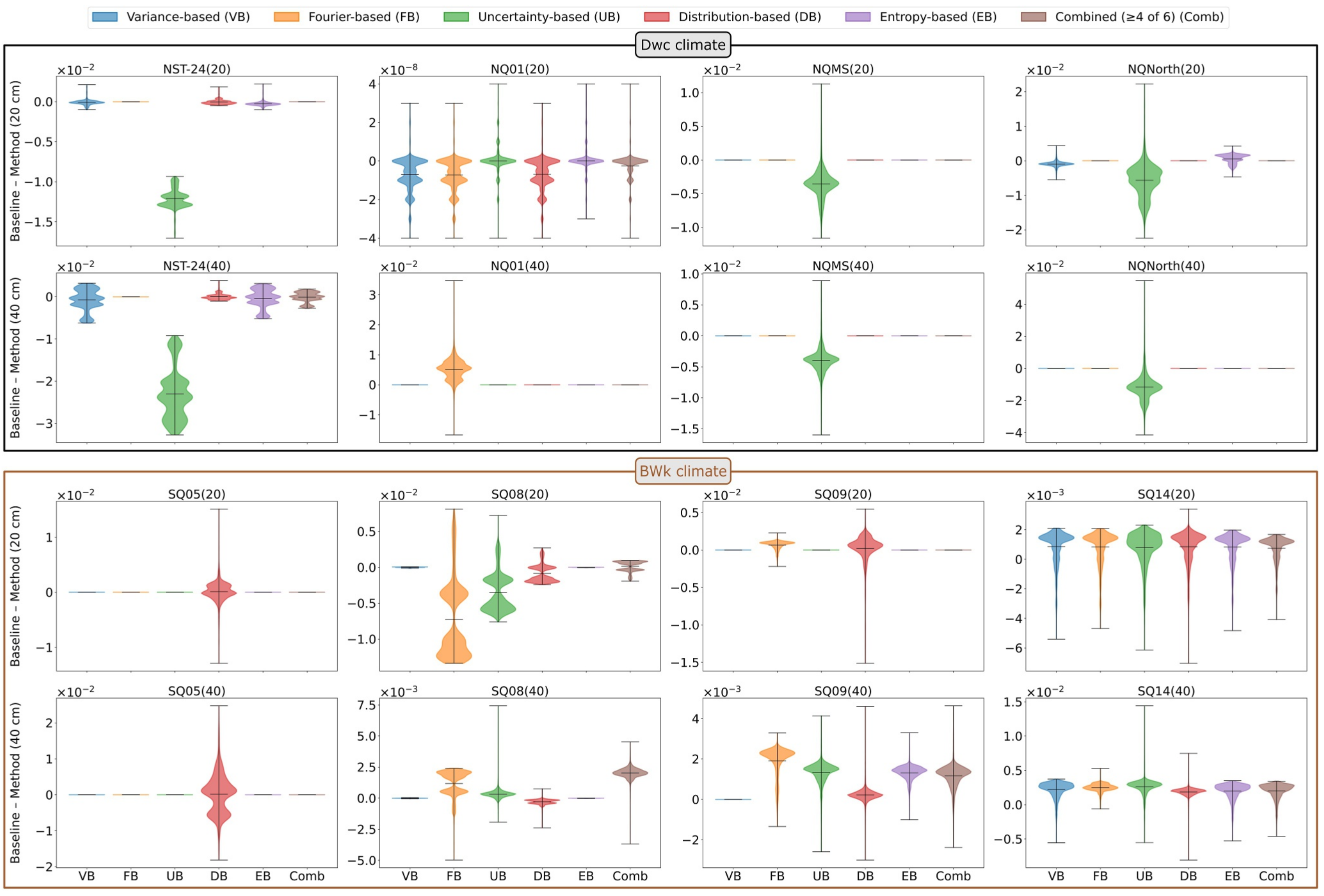

Figure 2. Distribution of deviations in predicted soil moisture obtained from different sampling strategies relative to the baseline method across eight stations and two subsurface depths (20 and 40 cm). Each violin depicts the spread and density of prediction differences, where the width corresponds to the frequency of occurrence around specific deviation magnitudes. Positive and negative values represent under- and overestimation compared to the baseline, respectively. The plots highlight inter-site and depth-wise variations, reflecting the consistency and bias tendencies of different selection methodologies.

4.2. Prediction Deviation Assessment Across Strategies

We analyzed the difference between the baseline-predicted soil moisture and that derived from six other sampling strategies for each station at near-surface (20 cm) and deeper layers (40 cm). To comprehensively illustrate this relationship, we used violin plots where the width of each violin represents the spread of deviation, while the median and standard deviation values quantify the central tendency and variability of the difference (Figure 2).

At the MAQU and NAQU stations, the deviation between baseline and alternative sampling methods remains minimal, indicating strong consistency in soil moisture prediction. Uncertainty-based sampling exhibits slightly larger variability, with standard deviation (σ) up to 0.005 at 20 cm and 0.008 at 40 cm depth. Other sampling methods yield narrower distributions with average σ of 1.5×10^{-4} at near-surface and 3.3×10^{-4} at deeper layers. The median of differences under the uncertainty-based method average -0.005 and -0.01 at 20 and 40 cm, respectively, while other sampling methods exhibit near-zero median values. This signify that it is effective in capturing dynamic soil moisture variability. However, it may slightly overestimate predictions relative to the baseline, specifically in the deeper layer where subsurface heterogeneity increases.

The NGARI stations show broader spreads, especially for distribution-based sampling. It records an average σ of 0.001 at 20 cm and 0.002 at 40 cm. Other sampling methods show comparatively lower variability (average σ of 4×10^{-4} at 20 cm and 7×10^{-4} at 40 cm). The median deviations for this strategy average 1.25×10^{-4} at shallower depth which subsequently increase to 6.5×10^{-4} at greater depth, reflecting a consistent positive bias (underestimation) relative to the baseline.

Overall, these results highlight a systematic inverse relationship between training data fraction and prediction dispersion relative to baseline performance. Stations located in Dwc climate exhibit remarkably stable predictions across all sampling methods. This stability arises from regular seasonal cycles and distributed precipitation, which provide abundant informative variability that supports efficient model training with minimal data. In contrast, the NGARI stations with BWk climate show a slightly greater dispersion and bias, a pattern attributable to the episodic nature of soil moisture dynamics in dry regions. Here, variability concentrates in rare precipitation events separated by extended dry periods. It creates a challenging sampling scenario where reduced training subsets risk underrepresenting critical wetting episodes. Moreover, the reduced variability range in arid systems limits the information content per observation. Consequently, larger samples are required to capture the full moisture spectrum to adequately represent event-driven dynamics and maintain predictive fidelity.

4.3. Seasonal Dependence on Sampling Selection

We evaluated the efficacy of multiple sampling strategies, in conjunction with the baseline approach, for identifying monthly selection of critical training points. To visualize this temporal distribution of selected samples, we used stem plots, wherein the baseline selections leverages the full training data set and inherently captures all data points each month (Figure 3a).

At station NST-24(20), both Fourier-based and combined methods exhibit comparable count of training points as elected by the baseline. The remaining methods demonstrate varying degrees of data reduction. A distinctive seasonality in sampling density emerges across all strategies, with substantially fewer training samples selected during the monsoon regime (April-October) than in non-monsoon months. Specifically, uncertainty-based sampling incorporates fewest number of samples, about 117–370 (12%–25%) in the dry period and only 4–35 (0.5%–5%) samples during the monsoon followed by variance-, entropy- and distribution-based strategies. A consistent pattern persists at NST-24(40), where the reduction is even more pronounced. This suggests a diminished need for the training subset size in the deeper layer, likely due to weaker hydrological variability. At NQ01, most methods use the full training data set at both depths, with only minor reductions observed (about 20%) for Fourier-based method at 40 cm during the monsoon months. At NQMS, uncertainty-based sampling consistently selects the smallest subset, especially during the monsoon. Similarly, at NQNorth(20), multiple methods (variance-, entropy- and uncertainty-based) reduce data usage, particularly during the monsoon regime. At NQNorth(40), only variance-based strategy exhibits this behavior, highlighting depth-dependent stability in sampling behavior.

The NGARI stations exhibit a contrasting monsoonal response. At SQ05 (both depths) distribution-based method distinctly surpasses other methods utilizing fewer training samples while maintaining predictive robustness, with an increase in selected samples during the monsoon to capture enhanced soil moisture variability. Other stations (SQ08, SQ09, and SQ14) show similar seasonal modulation, with higher sampling density during the wet season compared to the dry period across both depths.

Overall, a distinct climatic control governs the selection behavior of critical training points across the study sites. The MAQU and NAQU stations reveal a prominent reduction in sample utilization during the monsoon where uncertainty-based sampling yields superior performance. However, the NGARI stations present an alternate pattern, with a higher points selection during the same period. For these sites, distribution-based method proves most effective, emphasizing the climatic dependence of optimal data selection frameworks. This divergence persists consistently across both surface and subsurface layers. The reduced sampling demand at the first four stations reflects the higher temporal persistence of soil moisture under sustained precipitation and limited variability during the wet season. Conversely, the NGARI sites experience greater short-term fluctuations in soil moisture due to sporadic convective rainfall and rapid evaporation, necessitating a denser sampling representation.

4.4. Diurnal Controls on Critical Sampling Efficiency

We employed polar plots to exemplify the diurnal distribution of critical training points identified by each sampling strategy (Figure 3b). At station NST-24(20), uncertainty-based strategy selects the fewest samples (37–81) while the Fourier-based, combined, and baseline methods integrate a substantially higher number of

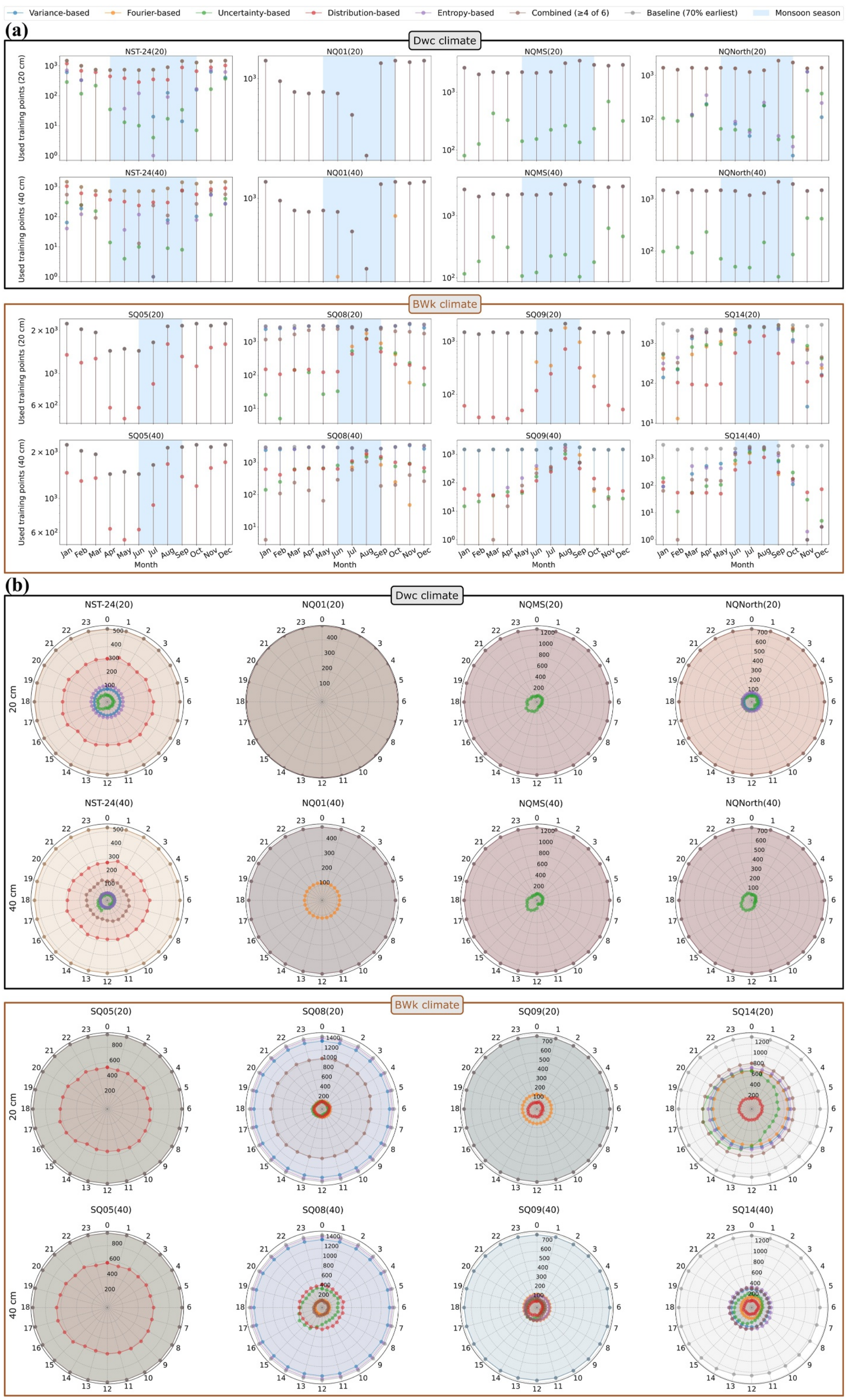

Figure 3. (a) Monthly distribution of selected training samples across eight stations at 20 and 40 cm soil depths. Each stem and marker denote the total number of critical points identified in a given month by different sampling strategies. The plots demonstrate the varying emphasis of different selection methods on distinct seasonal periods, with the blue-shaded region highlighting the monsoon season for each station. (b) Polar plots showing the diurnal distribution of critical training samples across eight stations at 20 and 40 cm soil depths. The angular axis represents the 24-hr cycle (at one-hour intervals) and the radial axis indicates the number of samples selected during each hour by different sampling methods. The plots reveal the capacity of each strategy to capture diurnal soil moisture variability and its temporal prioritization across depths and hydro-meteorological settings.

10.1029/2026GL124661

samples (529–531) across the day. A distinct pattern is identified under uncertainty-based method, wherein the 14:00–23:00 interval exhibits an increase in selected points (50–81) compared to earlier hours (37–50), reflecting enhanced data prioritization during the post-noon to evening period. At NST-24(40), most methods (combined, entropy-, variance-, distribution-based) show a marked reduction in selected samples relative to the shallower depth, while uncertainty-based sampling continues to emphasize greater selection after 14:00 hr. This indicates depth specific variability in sampling dynamics. At the NAQU stations, most sampling strategies produce analogous performance with a few notable deviations. At NQ01(20), Fourier-based method performs best whereas uncertainty-based strategy consistently selects fewer samples at NQMS and NQNorth stations, with noticeable increases during midday to evening hours.

At the NGARI stations, stronger spatial and depth-dependent variations are observed. Distribution-based method is most efficient at SQ05 and SQ09, consistently selecting the fewest critical points, with increased sampling between noon to evening. At SQ08, multiple methods perform well, with combined and Fourier-based strategies showing substantial reductions in sample usage (up to 88%) at 40 cm, while uncertainty-based and combined methods exhibit pronounced diurnal sample selection. At SQ14, all strategies use fewer samples than the baseline, particularly at greater depth, with predominant sample selection exhibited by uncertainty-, distribution-based and combined methods during 11:00–20:00 hr.

In summary, uncertainty-based strategy is the best performer at the MAQU and NAQU stations, integrating a smaller fraction of samples during the morning hours that progressively increases after noon. At the NGARI stations, distribution-based method uses the fewest critical training points, while Fourier- and uncertainty-based strategies also demonstrate strong performance. Uncertainty-, distribution-based, and in some cases, combined sampling exhibit a coherent diurnal modulation in sample selection. Among these, uncertainty-based method presents the most significant response across both depths. This pattern likely reflects soil-atmosphere coupling processes. During the afternoon, intensified evapotranspiration and radiative heating enhance soil moisture variability and surface energy exchange. This increases the informational value of samples selected during these hours. Alternatively, the nocturnal stabilization of soil moisture near midnight reduce variability, leading to a lower frequency of training point selection.

5. Conclusion

The proposed approach demonstrates that optimized critical subsets can replicate, and in some cases surpass, the predictive accuracy of a conventional model trained with 70% of the data. This indicates that reliable performance does not necessarily require extensive training fractions. Evaluation across eight ISMN stations spanning two contrasting hydroclimatic regimes reveals a strong climate-dependent control on sampling efficiency. Uncertainty-based sampling performs best under continental cold monsoonal conditions, where pronounced seasonal and diurnal variability creates informative transition states. In contrast, distribution-based and combined strategies are more effective in high-altitude desert environments, where soil moisture variability is limited and full-range state representation becomes essential. Although the 70% baseline is used here to benchmark sampling efficiency, it is not a prerequisite for future application. In practice, the framework is most suitable for data-sparse regions where some long-term reference stations exist within comparable hydroclimatic regimes, enabling transfer of the most effective strategy to nearby or similar sites. For regions without such reference data, the optimal strategy cannot be known with certainty a priori; however, physically interpretable approaches such as distribution-, variance-, entropy-, or combined sampling provide a practical starting point. Overall, the framework reduces dependence on extensive training data while retaining the value of representative observations for regime-specific calibration.

Conflict of Interest

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

Availability Statement

In situ data can be downloaded from Zenodo (V. Singh et al., 2025b). The code developed for this study can be downloaded from Zenodo (V. Singh et al., 2025a).

Acknowledgments

KG would like to acknowledge IISER Bhopal for institutional support. VS's PhD is supported through institutional fellowship from IISERB. The initial experiments that shaped the concept of this study were conducted at the Geohydrology Observatory Facility at IISER Bhopal, whose support we gratefully acknowledge.

References

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

Azizzadenesheli, K., Kovachki, N., Li, Z., Liu-Schiaffini, M., Kossaifi, J., & Anandkumar, A. (2024). Neural operators for accelerating scientific simulations and design. *Nature Reviews Physics*, *6*(5), 320–328. https://doi.org/10.1038/s42254-024-00712-5

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

Cao, Q., Goswami, S., & Karniadakis, G. E. (2024). Laplace neural operator for solving differential equations. *Nature Machine Intelligence*, *6*(6), 631–640. https://doi.org/10.1038/s42256-024-00844-4

Chavoshi, A., Dashtian, H., Bakhshian, S., Young, M. H., & Niyogi, D. (2025). Pinn-sm: A physics-informed neural networks model for vadose zone soil moisture profile prediction. *Journal of Geophysical Research: Machine Learning and Computation*, *2*(4), e2024JH000547. https://doi.org/10.1029/2024jh000547

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

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., 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–vzj2170. https://doi.org/10.2136/vzj2012.0170

Houben, T., Ebeling, P., Khurana, S., Schmid, J. S., & Boog, J. (2025). Machine-learning based spatiotemporal prediction of soil moisture in a grassland hillslope. *Vadose Zone Journal*, *24*(2), e70011. https://doi.org/10.1002/vzj2.70011

Hu, J., Deng, C., Zhang, Q., & Pang, A. (2025). Physics-informed neural networks enhanced by data augmentation: A novel framework for robust soil moisture estimation using multi-source data fusion. *Journal of Hydrology*, *663*, 134320. https://doi.org/10.1016/j.jhydrol.2025.134320

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Liao, D., Niu, J., Du, T., & Kang, S. (2024). Improving subsurface soil moisture estimation using a 2-dimensional data assimilation framework incorporated with a dual state-parameter scheme. *Water Resources Research*, *60*(8), e2023WR035771. https://doi.org/10.1029/2023wr035771

Liu, N., Jafarzadeh, S., & Yu, Y. (2023). Domain agnostic Fourier neural operators. *Advances in Neural Information Processing Systems*, *36*, 47438–47450.

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*. https://doi.org/10.1109/TGRS.2024.3502618

Singh, A., Gaurav, K., Sonkar, G. K., & Lee, C.-C. (2023). Strategies to measure soil moisture using traditional methods, automated sensors, remote sensing, and machine learning techniques: Review, bibliometric analysis, applications, research findings, and future directions. *IEEE Access*, *11*, 13605–13635. https://doi.org/10.1109/access.2023.3243635

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., & 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*(1), e2025JH001039. https://doi.org/10.1029/2025jh001039

Singh, V., Singh, A., & Gaurav, K. (2025a). Physics-aware sampling critical points [code]. *Zenodo*. https://doi.org/10.5281/zenodo.18103122

Singh, V., Singh, A., & Gaurav, K. (2025b). Physics-aware sampling critical points datasets [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.18103707

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

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. *Earth Systems and Environment*, *8*(3), 3487–3505. https://doi.org/10.1007/s40808-021-01316-z

Wang, Y., Shi, L., Hu, Y., Hu, X., Song, W., & Wang, L. (2023). A comprehensive study of deep learning for soil moisture prediction. *Hydrology and Earth System Sciences Discussions*, *2023*, 1–38.

Wang, Y., Zhang, H., Lai, C., & Hu, X. (2024). Transfer learning Fourier neural operator for solving parametric frequency-domain wave equations. *IEEE Transactions on Geoscience and Remote Sensing*.

Wen, G., Li, Z., Azizzadenesheli, K., Anandkumar, A., & Benson, S. M. (2022). U-FNO—An enhanced Fourier neural operator-based deep-learning model for multiphase flow. *Advances in Water Resources*, *163*, 104180. https://doi.org/10.1016/j.advwatres.2022.104180

10.1029/2026GL124661

Wu, Z., Cui, N., Zhang, W., Liu, C., Jin, X., Gong, D., et al. (2024). Estimating soil moisture content in citrus orchards using multi-temporal sentinel-1A data-based LSTM and PSO-LSTM models. *Journal of Hydrology*, *637*, 131336. https://doi.org/10.1016/j.jhydrol.2024.131336
Xu, Z., Man, X., Duan, L., & Cai, T. (2022). Improved subsurface soil moisture prediction from surface soil moisture through the integration of the (de) coupling effect. *Journal of Hydrology*, *608*, 127634. https://doi.org/10.1016/j.jhydrol.2022.127634
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