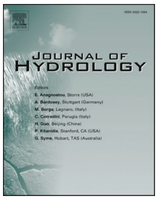
Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol



Exploring soil moisture dynamics and variability across scales and geological settings using gaussian mixture-long short-term memory networks

Balazs Bischof[a,*], Daniel Klotz[c], Hoshin V. Gupta[b], Erwin Zehe[a], Ralf Loritz[a]

[a] Institute of Water and Environment, Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany
[b] Department of Hydrology and Atmospheric Sciences, The University of Arizona, Tucson 85721, United States
[c] Interdisciplinary Transformation University Austria, Linz, Austria

ARTICLE INFO

This manuscript was handled by Huaming Guo, Editor-in-Chief, with the assistance of Heng Dai, Associate Editor

Keywords:
Soil moisture
Small-scale variability
Gaussian mixture models
Long-short term memory networks
Probabilistic deep learning
Variability quantification

ABSTRACT

Soil moisture is a key variable for a range of hydrological and ecological processes, yet capturing its small-scale variability and preferential flow phenomena remains challenging. Recent advancements in deep learning have demonstrated potential in predicting hydrological variables, but conventional data-driven models often struggle to represent small-scale variability effectively. In this study, we integrate Long-Short Term Memory (LSTMs) networks and Gaussian Mixture Models (GMMs) to simulate soil moisture dynamics while explicitly quantifying its associated variability. Unlike deterministic approaches, our probabilistic framework accounts for nonlinear relationships between inputs and outputs while modeling the inherent small-scale variability in soil moisture. We apply this methodology to a comprehensive in-situ soil moisture dataset from the Attert experimental basin, where the experimental design incorporates three replicated soil moisture profiles at each location and depth within a 5-meter radius. These replications are fundamental to our probabilistic framework: they provide direct, co-located observations of the natural spread in soil moisture under identical boundary conditions, allowing the model to learn the statistical structure of small-scale variability. This design enables disentangling sensor noise from genuine spatial heterogeneity and provides an empirical basis for training models that capture both temporal dynamics and local-scale variability.. Our results demonstrate that the proposed model reproduces soil moisture dynamics across multiple depths and scales, achieving an average Kling-Gupta Efficiency (KGE) of 0.52, Rank Correlation of 0.72, and Root Mean Squared Error of 0.036 m^3m^{-3}, while also capturing the key aspects of small-scale variability and sensor uncertainty. Furthermore, the modeled distributions offer new insights into the spatiotemporal structure of soil moisture and underscore the value of probabilistic modeling in hydrological approaches. By explicitly incorporating small-scale variability into the modeling process, our approach enhances both the interpretability and reliability of soil moisture predictions. While LSTMs effectively capture temporal dynamics, our findings underscore the necessity of incorporating variability quantification to improve model accuracy and generalization. This study highlights the potential of probabilistic deep learning frameworks in hydrological modeling and supports their broader application for improved soil moisture estimation and variability assessment.

1. Introduction

Soil moisture is a key state variable of wide-ranging importance. It influences groundwater recharge, infiltration, overland flow, and water supply to terrestrial biomass, thereby shaping global food security and agricultural sustainability (Zehe et al., 2010; Bronstert et al., 2012; Tietjen et al., 2009, 2010; Kaur et al., 2020). Furthermore, soil moisture regulates the partitioning of net radiation into sensible and latent heat, thereby controlling land–atmosphere interactions. It affects microbiological activities crucial for carbon cycling and contaminant transformation (Koehler et al., 2010; Köhne et al., 2006), and plays an important role in formation of flood runoff, erosion and the emergence of preferential flow and transport (Zehe et al., 2005; Grayson and Western, 1998). Although process-based hydrological and land-surface

[*] Corresponding author.
E-mail address: balazs.bischof@kit.edu (B. Bischof).

https://doi.org/10.1016/j.jhydrol.2025.134364
Received 17 April 2025; Received in revised form 21 August 2025; Accepted 30 September 2025
Available online 8 October 2025

models that rely on numerical solutions of the Darcy-Richards equation are commonly used to simulate soil moisture dynamics and their underlying controls, these models can struggle to adequately represent ubiquitous preferential flow phenomena (Beven and Germann, 1982).In the light of the recent success of Deep Learning (DL) in hydrology, we thus explore the potential of DL models as an alternative here.

In a broader context, the use of statistical learning methods to characterize soil moisture dynamics in space and time is of course not new. Several studies have relied on the use of geo-statistical methods to simulate soil moisture behavior (Bárdossy and Lehmann, 1998; Western et al., 2003; Brocca et al., 2010; Zehe et al., 2010; Mälicke et al., 2020; Herbst and Diekkruger, 2003). DL models, particularly Long Short-Term Memory (LSTM) networks, have recently emerged as one of the leading modeling approaches in hydrology due to their ability to capture complex temporal dependencies in hydrological data. LSTMs have been successfully applied to predict the dynamical evolution of hydrological variables such as streamflow (Kratzert et al., 2019), groundwater levels (Wunsch et al., 2021), and sap flow (Loritz et al., 2024) across different spatial and temporal scales. Other studies have explored the development of DL models, such as LSTMs and transformers, to predict soil moisture dynamics. Those studies have addressed various aspects of soil moisture modeling, including a general assessment of the predictive capabilities of different DL approaches (Wang et al., 2024), predictions of soil moisture dynamics at larger spatial scales (Liu et al., 2022), across different soil depths (Karthikeyan et al., 2021), and over extended time periods (Datta et al., 2023). However, all these studies have applied DL in a deterministic manner, predicting a single, fixed value rather than explicitly accounting for uncertainty and variability inherent in soil moisture measurements. This deterministic approach can be problematic, particularly when training models on in-situ soil moisture observations, as it fails to capture the multiscale variability and measurement uncertainties present in such data.

Predicting soil moisture dynamics in both space and time is challenging due to the inherent uncertainties in in-situ measurements and the multiscale spatial variability of soil moisture (Brocca et al., 2010; Robinson et al., 2008). To accurately capture these, it is essential to explicitly incorporate them into the modeling approaches. Factors such as sensor type, soil heterogeneity, land-management practices and installation practices can introduce variability and potential errors. While in-situ methods can provide relatively accurate point measurements, with a measurement accuracy of about 0.02 $\mathrm{m}^3\ \mathrm{m}^{-3}$ (Robinson et al., 2008; see also Jackisch et al., 2020), they sample only a small volume of soil, typically just a few cubic centimeters or decimeters. Due to this limited support volume and the small-scale variability of soil texture and related soil water retention properties, Time-Domain Reflectometry (TDR) measurements might exhibit enormous spreading at small scales. Measurements can exhibit offsets in their temporal mean of up to 0.2 $\mathrm{m}^3\ \mathrm{m}^{-3}$, even if they are located at separating distances of only a few meters (Zehe et al., 2010). Needless to mention that such small-scale differences are not explainable given the spatial resolution of data characterizing relevant physiographic differences (e.g. land use, soil type, etc.), even when working in comprehensively instrumented research catchments like the Attert experimental basin (Pfister et al., 2017), which is our study area. Despite this challenge, in-situ measurements are among the only sensors capable of measuring soil moisture at deeper depths (Demand et al., 2019) and essentially provide the uncertain ground-truth for training remote sensing products (Bronstert et al., 2012) and therefore remain indispensable. Soil moisture sensors in different depths, can furthermore provide important information about preferential flow phenomena. This manifests as a “non-sequential” event, where deeper sensors show a quicker response to rainfall than shallower ones, as shown by Demand et al. (2019) for the Attert experimental watershed.

Addressing the challenge of multiscale soil moisture variability requires suitable datasets in combination with probabilistic modeling approaches that can account for this form of variability, which is posing great challenges for model training and validation. Sensor readings often differ in ways that seem arbitrary and not fully explained by available data (e.g., precipitation, land use, elevation; Hosseini et al., 2015). As a result, a model might encounter identical input features but be confronted with largely different target values. These variabilities are not random; rather, they reflect the small-scale variability of e.g. soil texture, land use and micro-topography, which largely control soil water retention and thus soil moisture dynamics.

In the present study we analyze an in-situ soil moisture dataset from the Attert experimental basin. The experimental design, which clusters three soil moisture profiles, provides an ideal framework for examining the multiscale variability of soil moisture across three distinct geological and pedological settings. To effectively deal with the corresponding variability during model training, we propose using Gaussian Mixture Models combined with LSTMs (GMMs; Bishop, 1994; Klotz et al., 2022). This combination (hereafter referred to as the ‘model’) is capable of estimating probability distributions while accounting for autocorrelated, nonlinear interactions between inputs and outputs. By training on in-situ soil moisture observations, our model aims to generalize soil moisture dynamics across spatial and temporal scales, and depths. Unlike models that predict a single, deterministic value, the proposed representation generates weighted probabilistic distributions as outputs, which is a promising way to capture small-scale soil moisture variability at the plot scale. Our study aims to answer the following questions:

1. How effectively can GMMs predict in-situ soil moisture dynamics across different spatial and temporal scales and soil depths?
2. To what extent can these models capture and quantify measurement variabilities and uncertainties inherent in in-situ soil moisture observations?
3. Does the GMMs provide new insights about the nature of soil moisture variability and dynamics and, if so, how can this be explained?

2. Material and methods

2.1. *Attert catchment – a unique, natural laboratory*

The soil moisture data employed in this study is derived from hourly observations within the Attert experimental watershed, located in western Luxembourg and Belgium (Fig. 1a; Pfister et al., 2017). Geologically, the catchment comprises three primary formations: Devonian slates of the Ardennes massif in the northwest, Triassic sandy marls in the central region, and a small segment of Luxembourg sandstone along the southern border (Martínez-Carreras et al., 2012). The geological contrasts are closely reflected in differences in soil formation, topography, and land use patterns, which in turn shape distinct hydrological behaviors (Jackisch et al., 2017). In the schist-dominated uplands, steep forested slopes and shallow, stony Cambisols with limited storage capacity promote rapid runoff generation during rainfall events, while agricultural fields and pastures are confined to the flatter plateaus. The marl region, by contrast, is characterized by gentler slopes, silty to clay-rich soils with low permeability, and a predominance of pasture and crop cultivation, leading to slower infiltration rates and a greater propensity for surface runoff. The sandstone area in the south, though spatially limited, supports well-developed sandy Cambisols and extensive beech forests; its highly permeable substrate facilitates rapid vertical infiltration and deep percolation, sustaining baseflow during dry periods. These sharp contrasts in geology, soils, and land cover give rise to markedly different runoff generation mechanisms, storage capacities, and seasonal flow regimes across the basin. Coupled with its dense sensor network and high-resolution hydro-meteorological monitoring, the Attert watershed serves as a unique natural laboratory for investigating how lithology, land use, and climate interact to control soil moisture dynamics and streamflow responses at multiple spatial and temporal scales.

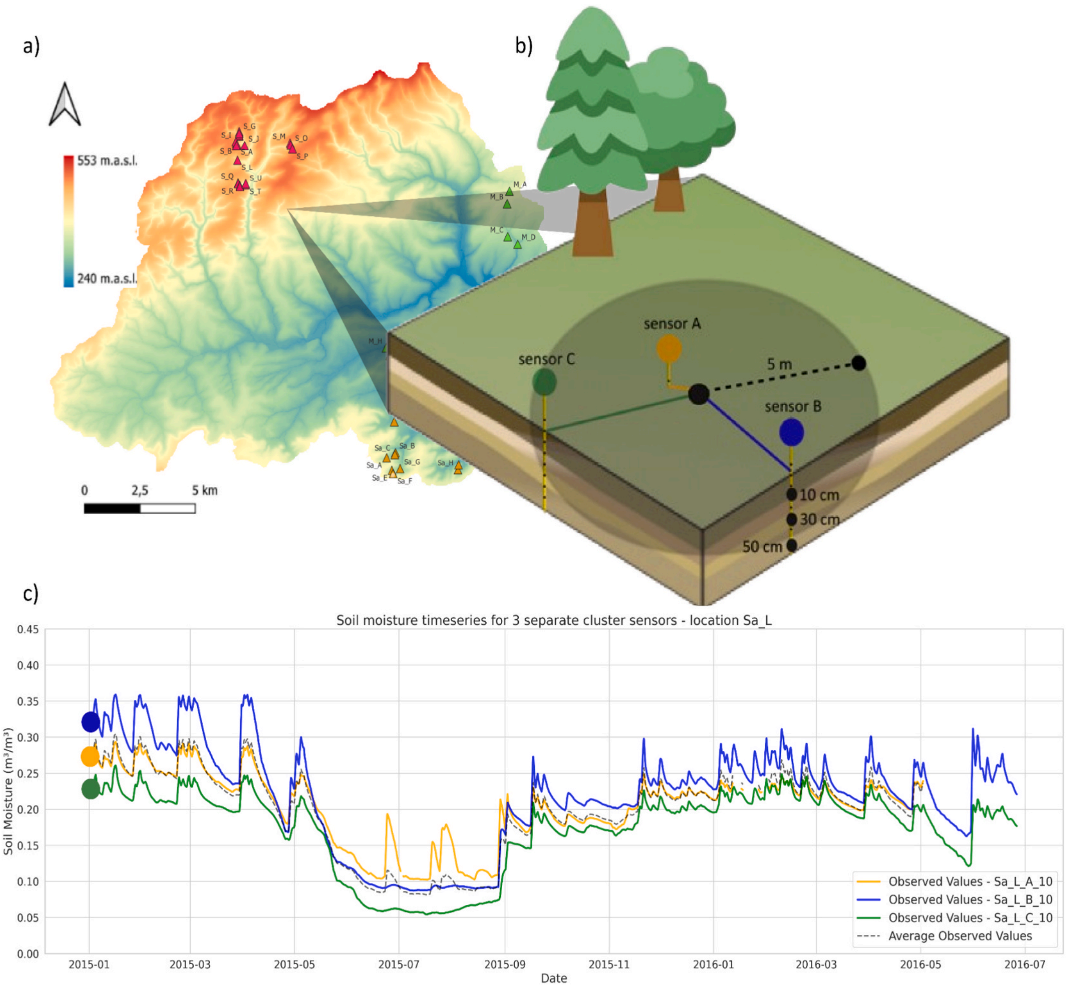


Fig. 1. The Attert experimental basin and the underlying variability of soil moisture measurements: a) the Attert experimental basin and measurement locations (S – schist, Sa – sandstone, and M – marl); b) representation of cluster measurement allocation (one measurement location consist of 3 randomly allocated sensors within a 5-meter radius circle; each measuring soil moisture at three different depths; c) underlying soil moisture variability of three sensor measurements belonging to location Sa_L at 10 cm.

2.1.1. *In-situ soil moisture measurements in the attert basin*

In-situ soil moisture measurements were taken from spatially dispersed point measurements arranged in “so-called” clusters (Fig. 1b; see also Zehe et al., 2014). These sensor clusters are strategically positioned across hillslopes, encompassing the upper, middle, and lower sectors along the expected flow trajectory in all three geologies of the Attert basin. At each of the 43 cluster locations, soil moisture dynamics are measured at three locations and at three depths of 10, 30, and 50 cm, totaling nine sensors per cluster. The sensor profiles are randomly allocated within a circle of approximately 5-meter radius (Fig. 1b). Assuming that the differences between measurements from sensors within the same cluster offers a reliable indication of existing small-scale variability at a specific location, this arrangement provided valuable insights into small-scale variability. In total, soil moisture time series from 43 distinct measurement locations at different depths were collected using a sum of 387 (43 locations – 3 depths – 3 sensors) soil moisture sensors between the years 2012 and 2017. All time series underwent initial quality control procedures prior to analysis, including the removal of physically implausible values, elimination of duplicated timesteps, and exclusion of data from malfunctioning sensors. These steps ensured that only physically consistent and reliable observations were retained. We only used data from places where all sensor readings (i.e., A, B, and C) were available for the same timestep in order to maintain consistency and facilitate fair model comparison. It is important to note that not all sensors provided data for the entire observation period. Some recorded data for only a few months up to the entire period due to maintenance requirements and occasional technical issues, resulting in varying data availability across the dataset (see section 2.3). In addition, the different geological formations were represented by varying amounts of available data, reflecting both the logistical challenges and the practical realities of long-term field monitoring. While this uneven coverage means that some regions contribute more observations than others, it also provides an opportunity to assess model performance under a broad range of data densities and environmental conditions, thereby offering a richer basis for evaluating the robustness and transferability of our approach.

2.1.2. *Characterizing small-scale variability in soil moisture*

The sensor measurements described in Section 2.1.1, generally exhibit strong alignment across most cluster locations, as indicated by high correlations (> 0.9) between the readings (Loritz et al. 2017). However, in certain instances, notable biases are observed (at some locations exceeding 20 % Vol; Fig. 1c). These differences stem partly from the limited accuracy of the sensors and partly from the pronounced multiscale variability of soil moisture (Mälicke et al., 2020). Such variability arises because time-domain reflectometry (TDR) probes measure over small support volumes, typically much smaller than the representative elementary volume of soil water storage in heterogeneous soils (Famiglietti et al., 2008). As a result, distributed TDR measurements can exhibit a spread of up to 0.23 m^3 m^{-3}, even with areas as small as 20 x 20 km^2 (Zehe et al., 2010), and variability at the 5–10 m scale can be as large as across an entire 20 km^2 catchment (Mälicke et al., 2020). This reflects both small-scale heterogeneity in soil texture, vegetation cover, and microtopography, as well as large-scale differences in soil and vegetation characteristics and meteorological forcing. Microtopographic features, such as slopes and depressions, affect local water redistribution and drainage; variations in soil texture influence water retention and permeability; and vegetation cover alters moisture dynamics via root water uptake and evapotranspiration. In addition to these spatial effects, soil moisture exhibits strong temporal variability driven by precipitation events, evapotranspiration, and longer-term climatic fluctuations. The magnitude and persistence of these changes can differ between locations due to differences in storage capacity or infiltration rates further increasing the challenge of capturing soil moisture dynamics and variability.

Importantly, while individual point measurements may not be representative of spatially averaged storage, distributed observations are often rank stable at both small and large scales, meaning they can capture temporal changes in relative soil moisture well despite large differences in absolute values (Mälicke et al., 2020). Since this variability results from a superposition of factors – many of which are unmeasured in our dataset – models that can learn and represent such variations and associated uncertainties are essential for accurate soil moisture predictions. In this study, we use the term **small-scale variability** to refer to spatial variations in soil moisture occurring over distances from a few meters up to tens of meters, typically within a single hillslope or field, and driven by fine-scale differences in soil, vegetation, and microtopography. Small-scale *spatial* variability arises from heterogeneities in soil texture, vegetation cover, and microtopography, which influence infiltration, retention, and redistribution. Small-scale temporal variability refers to short-term fluctuations (hours to days) at a given location, often linked to localized precipitation, evaporation pulses, or lateral water movement at the plot or hillslope scale. Both forms of variability overlay larger-scale spatial gradients (e.g., soil and vegetation types across catchments) and longer-term temporal changes (e.g., seasonal or multi-year trends).

2.1.3. *Input variables for the LSTM*

We considered four dynamic features, at an hourly resolution, comprising meteorological variables available in the database: air temperature (°C), precipitation (mm/h), and discharge (m^3/s). Air temperature was measured at site-level, while discharge measurements were used from locations in the proximity of soil moisture sensors (CAOS ('Catchments as Organized Systems') dataset, Neuper and Ehret, 2019; Kaplan et al., 2019; Luxembourg Institute of Science and Technology (LIST), 2023). Potential evapotranspiration (PET) data (mm/h), estimated using the Penman-Monteith equation (Pfister et al., 2017; Allen, Pereira, Raes and Smith, 1998), was obtained from multiple meteorological locations within the catchment. The PET data corresponding to the nearest meteorological station for each cluster site location was identified and utilized. Precipitation data is based on radar measurements across different location groups, encompassing a total of seven pixels in the catchment (Neuper and Ehret, 2019; Administration des services techniques de l'agriculture (ASTA), 2023).

For each soil moisture observation site, the distance to each measurement location was calculated, allowing the nearest station data to be selected and associated with that specific location. Furthermore, we included static attributes such as soil type, land use, and elevation data as input variables (European Environment Agency (EEA), 2020). To differentiate between data from different depths an additional input column was introduced: sensor depth (in cm), containing information about the actual depth of the sensor. Table A1, Appendix A contains all the relevant details about the implemented features, including information on the benchmark model (see section 2.4.2).

2.2. *DL model— Gaussian mixture long short-term memory*

In the following sections we introduce the LSTM and GMM architecture and explain how we combine these two approaches.

2.2.1. *Long short-term memory (LSTM)*

LSTMs are a type of recurrent neural networks (RNNs) specifically designed to address the vanishing and exploding gradient problems commonly encountered in traditional RNNs (Hochreiter, 1998; Hochreiter and Schmidhuber, 1997). This is achieved through the introduction of a cell state, which enables the network to capture and retain long-term dependencies, an essential feature when dealing with environmental and sequential data. The memory cell is regulated by "gates" that control the flow of information, allowing the network to evolve its memory and outputs over time while ensuring that errors can propagate consistently through the network, thus enhancing the learning process. LSTMs have demonstrated their effectiveness in hydrological modeling and are recognized as high-performing models for simulating various hydrological variables (e.g., Nearing et al. 2021).

2.2.2. *Gaussian mixture models (GMM)*

GMMs, introduced in the context of Mixture Density Networks (MDNs) by Bishop (1994), combine a neural network with a Gaussian mixture model to enable the generation of multiple outputs in response to one or more inputs. This approach allows for the inclusion of confidence intervals in regression tasks, which is crucial for accounting for uncertainties, an essential aspect of hydrological modeling (Liu et al., 2007), particularly for variables like soil moisture that exhibit significant multiscale variability. In MDNs, the mixture model is a probabilistic model constructed from a weighted sum of probability distributions. Mathematically, for each input x_t (representing each time step in our regression problem), the model predicts a weighted probability density function $P(y_t)$:

$$P(y_t|x_t) = \sum\nolimits_k^K \pi_k(x_t) N(y_t|\mu_k(x_t), \sigma_k(x_t))\pi_k(x_t) \tag{1}$$

In essence, we sum the total number K of mixture components (which are Gaussians in our case), using the weight π_k. The weights represent the relative importance or contribution of each Gaussian distribution in the mixture, and must sum to 1 (i.e., $\sum_k^K \pi_k(x_t) = 1$). The term $N(y_t, \mu_k(x_t), \sigma_k(x_t))$ is the Gaussian probability density function, where parameters μ_k and σ_k represent the mean and the variance of the individual distributions.

In other words, unlike the typical use of LSTMs in hydrology, where a single LSTM layer with a linear headlayer provides one deterministic time-evolving prediction, the Gaussian Mixture-LSTM (GM-LSTM) model outputs multiple parameters that define a time-evolving mixture of Gaussian probability distributions (Fig. 2). Specifically for this study, instead of producing a single point estimate of soil moisture, the model outputs several time-evolving means and standard deviations for a set of Gaussian distributions, along with their corresponding weighting factors. This approach allows the model to produce probabilistic outputs, representing not just a single outcome but a distribution of possible outcomes, a particularly valuable feature when dealing with heavily variable, uncertain, or multimodal data.

2.3. *Model hyperparameters and loss function*

The general model setup presents an LSTM combined with a GMM as a head-layer to model time-evolving probabilistic outputs in the context of sequential data (see section 2.2.2). We tested several different hyperparameter settings and found that the performance remained relatively stable across a broad range of values. The hyperparameter setting used for all model variants in this study are: number of hidden layers = 1; hidden layer neurons = 64; learning rate = 0.0005; dropout rate = 0.4; batch size = 128; sequence length = 128 hr; epochs = 10;

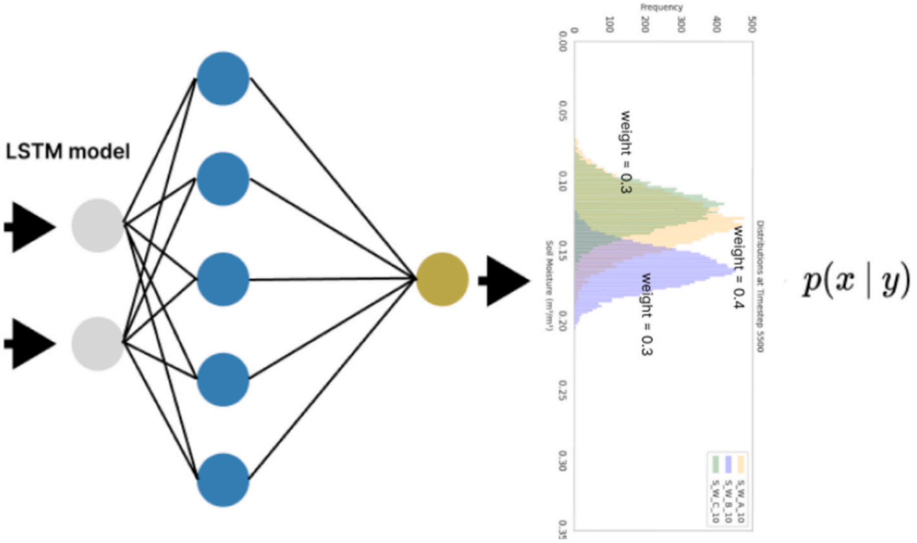


Fig. 2. Illustration of a GMM combined with an LSTM architecture. The LSTM model processes sequential input data, which are then used to predict the probability distribution $p(x|y)$ as a mixture of Gaussian components.

iterative optimization algorithm = ADAM. We selected Gaussian distributions in the network based on the assumption that the variability in our soil moisture data can be effectively represented by a Gaussian mixture. The learning rate was adapted after each epoch starting from a learning rate of 0.0005.

For loss function we employ the negative log-likelihood (NLL) which, as noted in Klotz et al. (2022), differs from the standard loss functions such as mean squared error (MSE) typically used in LSTM regression tasks. This loss function measures how well the predicted mixture of Gaussians aligns (in a probabilistic manner) with the target data. Accordingly, it is designed to maximize the likelihood of the target values y_t given a mixture of Gaussian distributions with mixing coefficients $\pi(x_t)$, mean $\mu(x_t)$, and standard deviations $\sigma(x_t)$. Mathematically (2), the negative log-likelihood loss can be written as:

$$NLL_{loss}(x_t, y_t) = -\frac{1}{M}\sum_{i=1}^{M} log\left(\sum_{k=1}^{K}\pi_k^{(i)}(x_t)\cdot N\left(y_t^{(i)}|\mu_k^{(i)}(x_t), \sigma_k^{(i)}(x_t)\right)\right), \qquad (2)$$

where M is the total number of sensor readings in the mini-batch (which equals the number of timesteps across the mini-batch multiplied by the number of sensors), K is the number of Gaussian components in the mixture, $\pi_j(x_t)$ is the mixing coefficient for the j-th Gaussian component, satisfying $\sum_{j=1}^{K}\pi_j(x_t) = 1$ (enforced by a soft-max function). The sum inside the logarithm represents the probability of observing the true value $y_t^{(i)}$ under the predicted mixture model.

This equation encapsulates the concept of a probabilistic mixture model: it expresses the likelihood of observing the target value $y_t^{(i)}$ as a weighted sum of Gaussian distributions, each associated with a weighting coefficient $\pi_j(x_t)$, summarizing to one, and characterized by its own mean $\mu_j(x_t)$ and standard deviation $\sigma_j(x_t)$. Taking the log-sum-exp (log-sum of exponentials) of the weighted Gaussian probability densities provides numerical stability and mitigates underflow issues, especially when dealing with very small values in the probability densities. In general, the loss is computed by calculating the log-likelihood for each sensor independently, then averaging them across all samples in the mini-batch. The negative sign is applied to convert the likelihood into a loss to be minimized (since optimization typically involves minimizing a cost function). Minimizing the loss maximizes the likelihood of the observed data, encouraging the model to assign higher probabilities to the correct target values.

In our case the models are designed to predict not just a single soil moisture value but a probability distribution, both capturing the expected value and the variability around it. To train such models effectively, the loss function must evaluate the entire predictive distribution and reward both the accurate central predictions and well-calibrated, sharp variability estimates. The NLL does exactly this: it measures how much probability the model assigns to the observed value, penalizing predictions that are either biased or miscalibrated. It is also a proper scoring rule, meaning that its expected value is minimized when the predicted distribution matches the true one. This property ensures that minimizing NLL leads the model to learn the correct distribution. By contrast, traditional losses such as mean squared error (MSE) only optimize for the mean and provide no way to learn input-dependent variances or multimodal distributions, both of which are crucial for our approach.

2.4. *Model experiments*

We use two model configurations to evaluate the GM-LSTM's performance in predicting hourly soil moisture dynamics and replicating observed variability at cluster sites. These setups are compared to a) a *statistical baseline model* and b) ERA5-Land (Muñoz-Sabater et al., 2021) to assess performance and variability estimates against measurements.

2.4.1. *Model configurations: Conditioned 1G-LSTM vs. Unconditioned 3G-LSTM*

We designed two LSTM model configurations (Table 1): one incorporating only a single Gaussian component (*conditioned 1G-LSTM*, hereafter referred to as *condLSTM*) and the other employing a three-Gaussian mixture (*unconditioned 3G-LSTM*, hereafter referred to as *3G-LSTM*). The first utilizes a single Gaussian distribution and, besides the previously outlined input variables (section 2.1.3), incorporates an additional feature representing the sensor ID (e.g., A, B, or C) at each location via one-hot encoding. Formally:

$$P(y_t|x_t) = N(y_t|\mu_k(x_t; i(ID)), \sigma_k(x_t; i(ID))) \qquad (3)$$

where $i(ID)$ returns an encoding of the sensor ID. The *condLSTM* threads each sensor as a separate prediction task. A given prediction is a construction by conditioning on the given task. Hence, the model is global in the sense that the same parameters are used for each task/sensor, which, in theory, allows for learning the overall representation of the soil moisture process.

The *3G-LSTM* does not encode sensor ID information. Instead, this model was designed to directly infer the relevant dynamics and

Table 1
Summary of the two model configurations.

Model	Sensor ID information	Input variables	Hyperparameter setup	Output
condLSTM	Used	Precipitation, evaporation, discharge, air temperature, DEM, soil type, land use	Same (except number of Gaussians = 1)	single probabilistic distribution for each sensor measurement
3G- LSTM	Not used	Precipitation, evaporation, discharge, air temperature, DEM, soil type and soil texture, land use	Same (except number of Gaussians = 3)	a mixture of three weighted probabilistic distributions for each sensor measurement

variability solely from the three soil moisture measurements at each cluster, without guidance from sensor-specific input. That is, the *3G-LSTM* learns a marginal distribution over all 3 sensors (while the model *condLSTM* learns a conditionalized distribution for each sensor).

In theory, both modelling approaches should be able to learn the soil moisture process with the same fidelity. In practice, it is however not clear which approach is able to extract the most information from the data. Comparing these two setups is useful for evaluating different modeling strategies: the first setup provides explicit control and interpretability by using information about sensor ID's, while the second one explores the model's ability to learn these differences directly from the data. This comparison helps identify the most effective approach for capturing site-level soil moisture dynamics and understanding the underlying variability.

2.4.2. Model evaluation and benchmarking

Due to discrepancies in the temporal lengths of available soil moisture time series, the conventional practice of partitioning data (i.e., chronological) into training, selection, and evaluation subsets poses difficulties. This creates a challenge where data from some measurement locations are confined to either the training, selection, or evaluation sets, but lack sufficient duration to span all subsets. Consequently, this deficiency manifests as incomplete data coverage within the evaluation subset, complicating the assessment of the developed model's performance regarding dynamics and variability as well. To circumvent this issue, we follow the approach of Loritz et al. (2024) and partition soil moisture time-series from all available locations into segments of uniform time duration of 2160 h (90 days). These segments are then randomly distributed across training and evaluation subsets (in a 60 %-40 % ratio) prior to each model initialization. To ensure uniform representation within each subset, and to assess the model's stability across different time segments during training, a unique random seed is used for each segmentation and changed before each model initialization. This strategy ensures that training consistently occurs with distinct, randomly chosen time segments.

To make our model results better comparable with the literature we evaluated the performance of the different model setups based on the Kling-Gupta Efficiency (KGE; Gupta et al., 2009), the Spearman's Rank Correlation, and the Mean Squared Error (MSE). Using these metrics for model evaluation allows us to assess both the accuracy of the model and its ability to preserve temporal dynamics and rank order, particularly when dealing with non-linear and variable soil moisture data. As recommended by Maier et al. (2023) during evaluation we bootstrap by repeatedly drawing random subsets (80 %) of the evaluation data with replacement and calculate the KGE and rank correlation on each subset for 50 iterations. This approach provides more reliable estimates of model accuracy and generalization, especially with sparse or heavily variable data like soil moisture. The models were evaluated by comparing the average soil moisture content at a cluster site per depth against the weighted average of the predicted distribution. For the *condLSTM* model it is straightforward to compare the model's predictions with actual sensor observations. However, this is not the case for the *3G-LSTM* model. Therefore, to ensure a fair model comparison and gain a general understanding of how well the model captures the dynamics, we assessed its performance by comparing the average of the three sensor measurements with the average of the model predictions across all three sensors (i.e., weighted means of simulated mixture distributions).

We benchmark our model against a statistical baseline model and against the ERA5-Land land surface model (Muñoz-Sabater et al., 2021). For the statistical baseline model, we computed the spatially averaged soil moisture for each timestep and location within the same geological formation group (e.g., schist, marl, and sandstone) using the data from the training subset. This provides the average annual cycles of soil moisture, which is the null model recommended by Schaefli and Gupta (2007) for data with strong seasonality. Furthermore, we compared our model with the volumetric water content estimated by the ERA5-Land model. The model provides volumetric water content estimations for four depth ranges: 0–7 cm, 7–28 cm, 28–100 cm, and 100–289 cm. For our analysis, we excluded the deepest layer and compared the first three layers with our three corresponding depths of 10, 30, and 50 cm. Despite its relatively coarse 9x9 km resolution (which limits the ability to reflect fine-scale variations in areas with heterogeneous soil types), it has been shown that ERA5-Land effectively captures the general trends and seasonal patterns in soil moisture (Lees et al., 2022; Manoj et al., 2024). This makes comparisons with the product useful as it provides a reliable reference for assessing our model's performance in representing the overall dynamics of soil moisture across different locations.

2.4.3. Capturing small scale variabilities of in-situ soil moisture measurements

To compare the observed variability with variability learned by the models we calculated the former among sensor readings at each timestep, depth, and cluster site, using data only when all three sensors were active. To account for measurement uncertainties, random noise with a standard deviation of 0.02 m^3/m^3 was added to each sensor measurement (Robinson et al., 2008; Jackisch et al., 2020). Variability at a cluster site was then quantified as the difference between the 80th and 20th percentiles of the combined sensor distribution. We selected the percentiles to focus on the central part of the data, excluding extreme values in the top and bottom percentages. To evaluate the variability replicated by the *condLSTM* model for each cluster site, we treated the predictions from the three sensors as distributions with equal weights. For the *3G-LSTM*, the components were aggregated according to their respective weights, resulting in a single, weighted Gaussian mixture distribution for each cluster location and depth. The resulting distributions for both cases were then aggregated into a single probability distribution representing the cluster site itself. Similar to the observations we calculated the predicted variability by subtracting the 20th percentile from the 80th percentile of these aggregated distributions.

To achieve a more comprehensive representation, we then grouped the measured and predicted cluster site variabilities according to their geological location and depths. Additional to the visual inspection in Fig. 4, to quantify how effectively the small-scale variability at an individual cluster location is captured by the two model setups we utilize the Wasserstein distance (Villani, 2003), the Continuous Rank Probability Score (CRPS) (Gneiting and Raftery, 2007) and the log-likelihood (Bishop, 2006). Also known as the Earth Mover's Distance, the Wasserstein-distance metric from optimal transport theory measures the distance between two probability distributions. It captures both the magnitude and structure of discrepancies between distributions, making it particularly useful for comparing modeled and observed variability in soil moisture. Specifically, it quantifies the minimum “cost or work” needed to transform one distribution into another by shifting probability mass. The CRPS is a strictly proper scoring rule that evaluates the accuracy of probabilistic predictions by comparing the cumulative distribution function (CDF) of the forecast to the observed value. CRPS generalizes the Mean Absolute Error (MAE) to probabilistic forecasts, offering a measure of sharpness and calibration. Lower CRPS values indicate better probabilistic performance. The log-likelihood quantifies how well a model represents observed data by evaluating the probability of measurements under its predicted variability. To ensure comparability across different depths and geologies of varying lengths, we use log-likelihood per timestep. This prevents biases due to differing time series lengths (and differing amount of locations per geology). The computed log-likelihood ratio directly compares the models, with a positive value indicating superior performance of the *condLSTM*, while a negative value suggests better performance of the *3G-LSTM*.

Additionally, for both model configurations, we constructed an aggregated probability density function (PDF) by combining the Gaussian components, each weighted by its corresponding mixture coefficient, effectively representing the model’s predicted distribution. To compare this predicted PDF with the empirical distribution of observed soil moisture, we used a kernel density estimate (KDE) to approximate the observed data’s histogram. We then plotted the aggregated PDF alongside the observed KDE, enabling a visual assessment of how well the model’s probabilistic predictions align with the actual soil moisture data distribution.

3. Results

3.1. Model performances

The performance metrics for the two DL-based models and the two baseline models are shown in Table 2. The results indicate that:

1) The *cond* and *3G-LSTMs* perform similarly, as measured by KGE and rank correlation, in predicting average soil moisture at the different cluster sites. Since both DL models are trained directly on local soil moisture readings they outperform or match the *ERA5-Land* and the *statistical baseline* models on average.
2) While the average performance of the *cond* and *3G-LSTMs* match closely, some differences become apparent when analyzing specific geologies and depths. For instance, the *condLSTM* outperforms the *3G-LSTM* in the schist geologies, while the *3G-LSTM* performs better in marl geologies.
3) Overall, KGE values tend to decline with increasing depth, particularly in sandstone and marl formations, while rank correlation values generally increase. This contrast could arise because, although the dynamics in deeper soil layers are easier to predict as they are much smoother and often represent long-term seasonal patterns (hence higher rank correlation), the absolute values may deviate more due to potentially reduced sensitivity to surface-level influences at some sites (Demand et al., 2019).

Both LSTM models consistently achieve higher KGE values across nearly all locations compared to *ERA5-Land*, highlighting a significant bias in the latter (also see Fig. C1, Appendix C). This bias is problematic, as plant water stress and soil moisture droughts depend on absolute soil water content rather than its statistical distribution. Soil water retention is primarily governed by capillary forces, which intensify as moisture levels decrease and as soil texture becomes finer. Consequently, plant roots must overcome increasing capillary suction to extract water, with uptake stopping entirely at the permanent wilting point.

While *ERA5-Land* exhibits slightly higher rank correlations in some locations, this metric alone is insufficient for assessing hydrological and agricultural relevance, as it does not capture magnitude-dependent processes. The elevated rank correlations can be attributed to measurement characteristics in regions dominated by sandstone or deeper marl formations, where soil moisture variability is naturally low, often appearing as near-linear time series. Additionally, *ERA5-Land*’s estimates are inherently smoother due to the aggregation and interpolation of meteorological and hydrological variables over broad spatial and temporal scales. This smoothness can lead to an artificial alignment with low-variability measurements, inflating rank correlation values. However, despite these slightly higher rank correlations, both LSTM models outperform *ERA5-Land* in KGE (Table 2) and RMSE (Table F1, Appendix F). This indicates that while *ERA5-Land* may align with low-variability measurement profiles in certain geologies, the LSTM models more effectively capture soil moisture dynamics and variability, making them better suited for hydrological and agricultural applications.

The model performance, as measured by the KGE, appears to be highest in schist, followed by sandstone and lowest in marls. This can be well explained by the distinct soil moisture dynamics associated with these geological formations. In clay-rich marls, low permeability and high water retention lead to slower infiltration, producing relatively stable soil moisture levels during winter but pronounced drying and autumn recharge. Additionally, TDR measurements in clay-rich soils like marls tend to be more error-prone than in other soils, which more likely introduces uncertainty in the observed data. In contrast, schist and

Table 2
Performance of the two DL models (*cond* and *3G-LSTMs*) and the two baseline models (*ERA5-Land* and *statistical baseline*) measured by the KGE and Rank Correlation against the average soil moisture at a given geology and depth.

Location	KGE(1 is good, 0 is poor)				Rank Correlation(1 is good, 0 is poor)			
	cond LSTM	*3G-LSTM*	*ERA5 −Land*	*Statistical baseline*	*cond LSTM*	*3G-LSTM*	*ERA5 −Land*	*Statistical baseline*
Total average	0.511	0.516	0.406	0.208	0.704	0.721	0.757	0.461
Marls − 10 cm	0.282	0.547	0.352	0.203	0.649	0.661	0.623	0.407
Marls − 30 cm	0.578	0.573	0.452	0.198	0.702	0.740	0.731	0.324
Marls − 50 cm	0.300	0.472	0.612	0.137	0.716	0.709	0.762	0.416
Schist − 10 cm	0.640	0.572	0.413	0.083	0.722	0.706	0.713	0.436
Schist − 30 cm	0.719	0.655	0.326	0.169	0.723	0.737	0.759	0.457
Schist − 50 cm	0.605	0.515	0.223	0.253	0.735	0.760	0.736	0.512
Sandstone − 10 cm	0.552	0.545	0.560	0.316	0.713	0.766	0.840	0.478
Sandstone − 30 cm	0.522	0.499	0.343	0.289	0.730	0.718	0.866	0.554
Sandstone − 50 cm	0.404	0.267	0.366	0.221	0.650	0.721	0.785	0.566

sandstone exhibit much more dynamic soil moisture responses at the event scale, with faster fluctuations in response to precipitation and evapotranspiration. Schist geologies have moderate permeability, allowing the models to capture both surface and subsurface signals, resulting in consistently good performance. Sandstone, with its high permeability and rapid drainage, shows strong responsiveness in shallow layers to precipitation, but deeper layers are harder to predict because of fast vertical fluxes. Furthermore, the overall lower KGEs in the deeper depths can be partly caused by the chosen input features, which primarily represent the meteorological boundary conditions at a cluster site. Deeper soil layers in these geologies are likely influenced by subsurface processes that are not fully accounted for by surface meteorological inputs.

3.2. *Characterizing soil moisture variability across depths and geologies*

To evaluate the representation of underlying soil moisture variability, we compared the aggregate distributions of mixtures (combining the predicted distributions) for the *cond* and the *3G-LSTMs* with the total data distribution of actual soil moisture observations. The results show that both models align with the actual spread and shape of the soil moisture observations (Fig. D1, Appendix D). However, while this general comparison confirms that both models capture the overall distribution of the data well, the differences between the two setups in this aggregated view are negligible.

To better understand this, we examined the heteroscedasticity (i.e., state dependent variance) in both the measurements and the model outputs. As described in section 2.4.3, we compared the differences between measurements taken from the same cluster site (i.e., measurements of different sensors) with the variability estimates produced by the two models. To enable a broader and more general comparison while still accounting for site-specific variations, we aggregated these distributions in Fig. 3. within each individual geological formation (e.g., schist) and depth (e.g., 10 cm) separately. This allowed us to examine potential performance and variability differences in relation to different geological contexts and soil moisture measurement depths.

Fig. 3 shows comparisons between the distributions of measurement differences and the outputs from both models. It also displays the distributions of actual soil moisture values, predicted values, and associated uncertainties. This visualization highlights how both models successfully capture variability, effectively representing the overall spread and variability distribution of the data.

1) The visual inspection of the *3G-LSTM* indicates a slightly better alignment with observed measurements in terms of distribution.
2) In contrast, the *condLSTM* produces wider variability estimates, potentially overestimating the variability in soil moisture measurements.

These are underpinned by the Wasserstein distances and CRPS scores between the distributions, with the *3G-LSTM* showing a small improvement (around 18 %) in accuracy considering the Wasserstein-distance values, and having a slightly better accuracy based on the CRPS scores (Table 3). The *condLSTM* leverages specific information

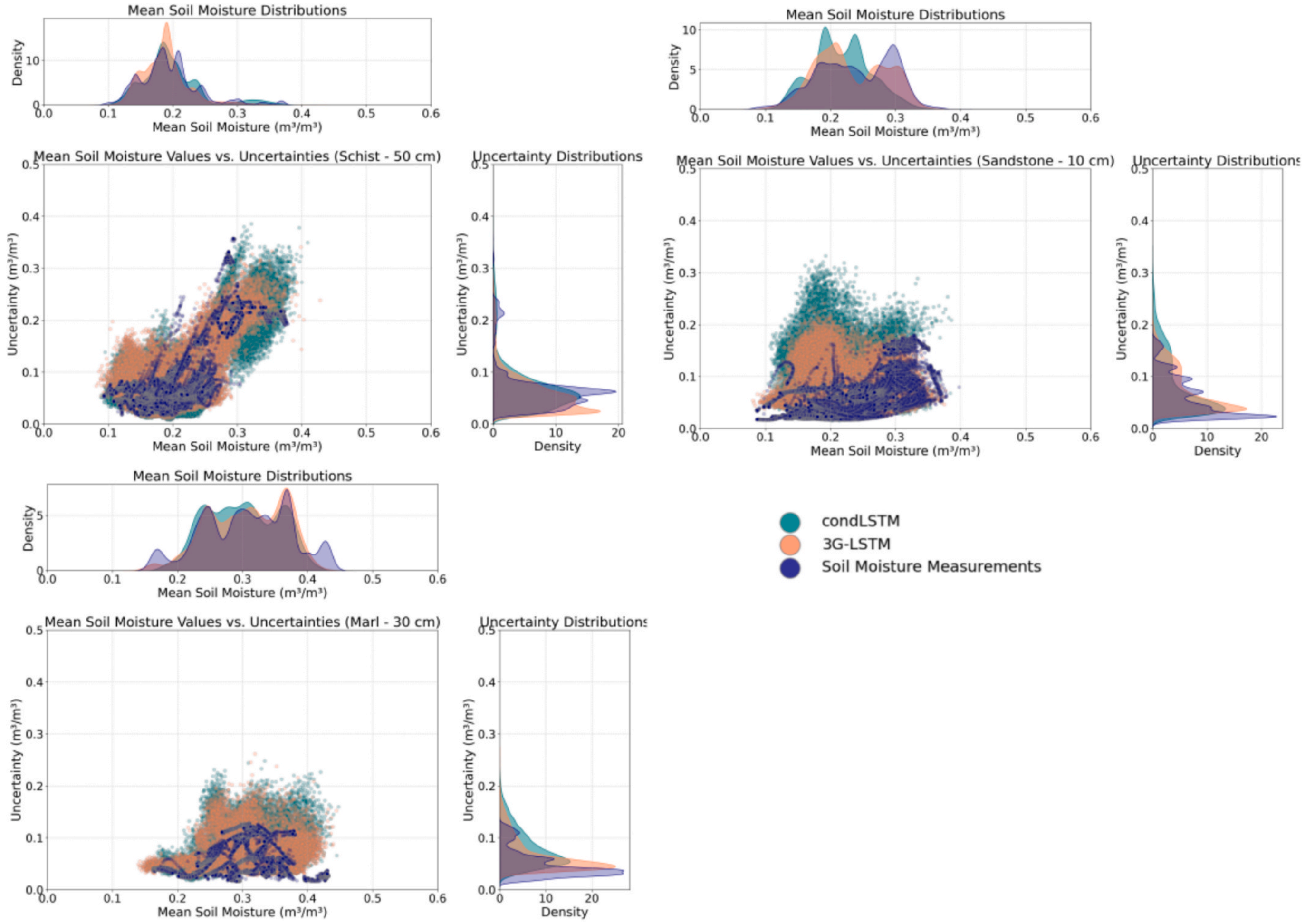


Fig. 3. Three comparisons showcasing the distributions of measurement differences alongside the outputs from both models: a) Schist formation – 50 cm depth; b) Sandstone formation – 10 cm depth; c) Marl formation – 30 cm depth. Figures of the remaining six geology-depth combinations can be found in Appendix G.

Table 3
Performance of the two DL models (cond and 3G-LSTMs) representing site-level soil moisture variability. The table shows the Wasserstein-distance (low values indicate good fit), the Continuous Rank Probability Score (CRPS, low values indicate good fit), and the average log-likelihood (high values indicate good fit) calculations for both models. Interpretation of 'good' and 'poor' values for the metrics (Wasserstein-distance and CRPS) can be found in Fig. B/1., Appendix B. The table showing all the values for different depths and geologies can be found in Table F2, Appendix F.

Location	CRPS		Wasserstein-distance		Log-likelihood		
	condLSTM	*3G-LSTM*	*condLSTM*	*3G-LSTM*	*cond LSTM*	*3G-LSTM*	*Ratio*
Total average	0.029	**0.028**	0.022	**0.018**	1.326	**1.371**	−0.042
Marlaverage	0.027	**0.025**	0.033	**0.025**	1.454	**1.555**	−0.091
Schistaverage	0.040	**0.039**	**0.016**	0.018	**0.778**	0.722	0.057
Sandstone average	0.021	**0.019**	0.017	**0.011**	1.746	**1.837**	−0.091

about sensor IDs, which one might expect to provide an advantage in representing the inherent variability of the data. The *3G-LSTM* model exhibits slightly higher log-likelihood values than the *condLSTM*, with differences mostly ranging from 0.04 to 0.09. These small differences suggest both models achieve similar accuracy in capturing site-level soil moisture variability. However, as demonstrated in the plots of Fig. 3, the *3G-LSTM* achieves a similar or even better representation without any explicit knowledge of sensor-specific information. This capability highlights the model's strength in incorporating variability through its design, which means learning three distinct inputs corresponding to the same output (i.e., three different sensor measurements belonging to the same cluster site). This is accomplished without relying on additional data about the differences in between the sensor measurements. Instead, the *3G-LSTM* implicitly identifies and integrates these differences through its training process, showing its robustness and adaptability in handling complex datasets with large variability, such as soil moisture.

Furthermore, the performance with respect to the variability varies when examining different geological formations and depths. Generally, the models perform better in deeper soil layers (Table F2, Appendix F), likely due to reduced noise and more stable soil moisture dynamics at greater depths, as opposed to the more variable and transient conditions near the surface. However, this trend is not uniform across all geologies.

1) For sandstone, variability remains relatively consistent across different depths. This can be explained by the fact soil properties do not change significantly with increasing depth in this geology.
2) In contrast, both schist and marls show a decrease in variability as depth increases.

These differences highlight the importance of considering geology-specific behavior when evaluating model performance across varying depths. Their performance in accurately reflecting site-specific changes

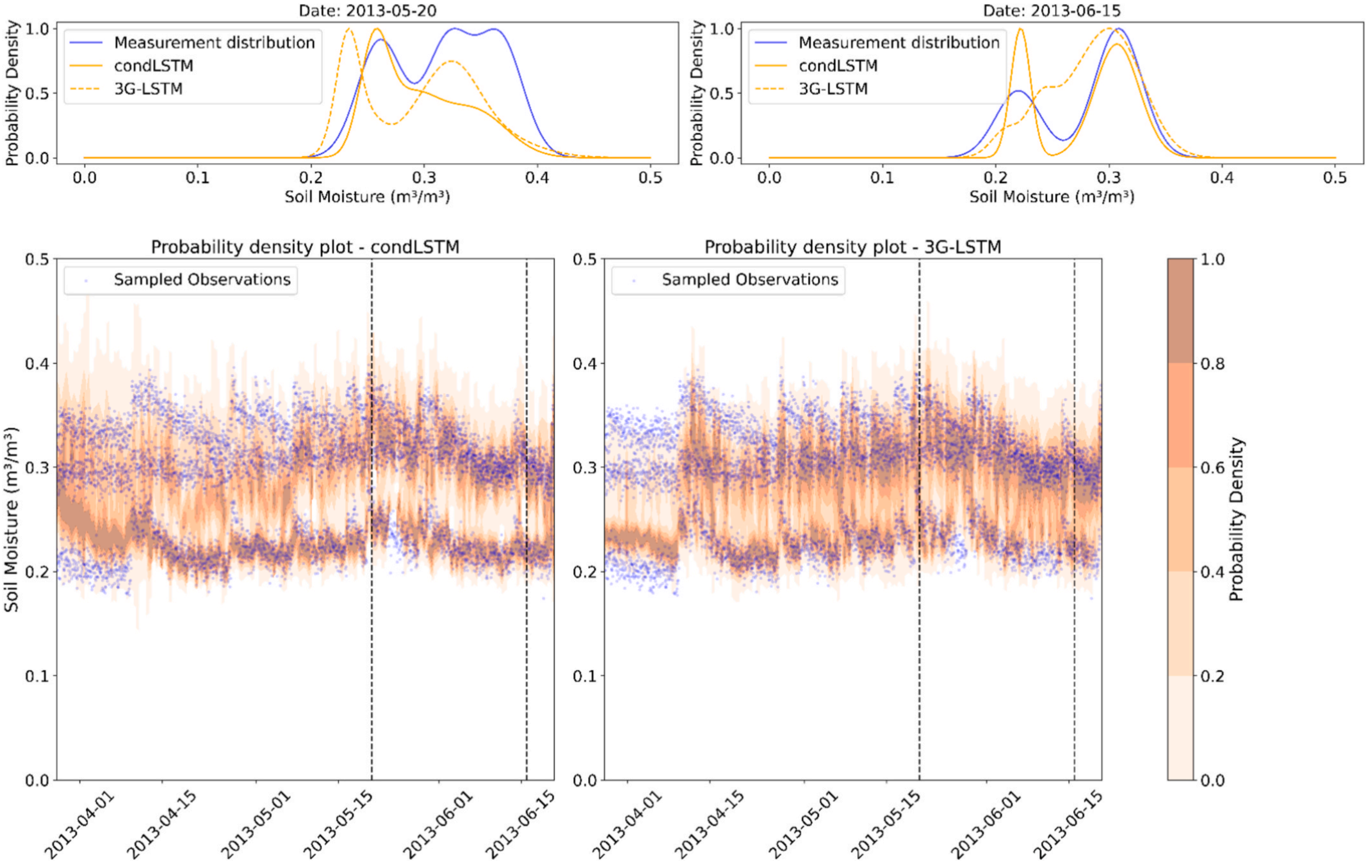


Fig. 4. Example time series plot of a sandstone location at 10 cm depth. The plot shows randomly drawn samples from the measurement distribution (combined distribution of the three sensor measurements from a single site – scattered blue points) and the weighted distribution of the *condLSTM* (left) and *3G-LSTM* (right). The upper two distribution plots show two randomly selected timesteps and the corresponding distribution of soil moisture measurements (blue line), *condLSTM* predictions (solid orange line), and *3G-LSTM* predictions (dotted orange line). Additional figures for different geologies and depths can be found in Appendix C.

is influenced by the unique characteristics of each geological formation. The ability of the *3G-LSTM* to capture variability at individual sites is demonstrated in Fig. E1, Appendix E, highlighting its effectiveness not only at a broader scale but also at the site level.

The previous visualizations and metrics provided a comprehensive evaluation of overall model performance and the extent to which both models capture variability at a broader and local scale. While the variability belts (as described in section 2.4.3.) effectively represent general performance, they do not account for potential differences in probability density distributions within the constructed belt. To address this limitation, we introduce a more granular assessment and visualization of the predicted variability. Specifically, we generated random samples from the combined distribution of the measurements, incorporating a standard deviation of 0.02 m^3 m^{-3} for each sensor individually (measurement uncertainty; Robinson et al., 2008; see also Jackisch et al., 2020), and compared these samples with the probability density estimates derived from the predicted mixture distributions for each timestep (Fig. 4). The figure demonstrates that, for both models, the estimated probability densities closely align with the distribution of the measurement data, indicating that the models effectively capture both the spread and associated probabilities. Furthermore, the upper two distribution plots reveal that the distributions at randomly selected timesteps exhibit a similar shape, suggesting that the models not only capture the overall variability across aggregated geologies (Fig. 3) and locations (Fig. E1, Appendix E) but also maintain consistency on the timestep level.

3.3. *Seasonal dynamics of multimodal soil moisture predictions*

We next explored the temporal evolution of the distributional characteristics of our models to assess whether changes in the statistical distributions of model outputs could be associated with varying soil moisture states and environmental conditions. We investigated whether the probability of multimodal soil moisture state occurrence exhibits seasonal trends.

As illustrated in Fig. 5, the modeled distributions – focusing on the outputs of the *3G-LSTM* due to its slightly higher accuracy in replicating observed soil moisture variability and to ensure clearer visualization – demonstrate seasonal dependency. Multimodality is more frequently observed during winter months compared to the late summer period (August-September). This seasonal pattern aligns with the increased discharge measurements and a higher frequency of rainy days during winter, which we hypothesize to be driven by enhanced variability in hydrological and meteorological forcings, such as snowmelt, precipitation events, and elevated groundwater levels. These factors contribute to a greater complexity in hydrological responses, leading to the emergence of multimodal distributions as the model captures multiple possible outcomes under similar conditions. Conversely, during late summer, the reduced frequency of rainy days and lower discharge rates indicate more stable hydrological conditions, which are reflected in a predominantly unimodal soil moisture pattern.

This relationship between multimodality, discharge dynamics, and precipitation frequency highlights the sensitivity of the models to environmental variability. Notably, multimodal predictions emerge as an indicator of periods with heightened variability and complexity in hydrological responses, reflecting the models' ability to represent alternative flow pathways under varying conditions. This interpretation aligns with the study on sequential and non-sequential soil moisture responses of Demand et al. (2019), where non-sequential responses – often indicative of preferential flow bypassing the upper soil layers – were observed more frequently under certain seasonal and moisture conditions. The presence of both sequential and non-sequential responses within a cluster site may provide a physical explanation for the observed multimodality, further highlighting the role of heterogeneous infiltration dynamics and preferential flow pathways in shaping soil moisture distributions. Importantly, Demand et al. (2019) observed a distinct seasonal pattern in the percentage of the total 135 TDR profiles, with peaks occurring in summer and early fall and the lowest values during the wet season. This pattern is essentially the inverse of the seasonal trend in multimodality depicted in Fig. 5.

4. Discussion

In the following we discuss our findings regarding the three research

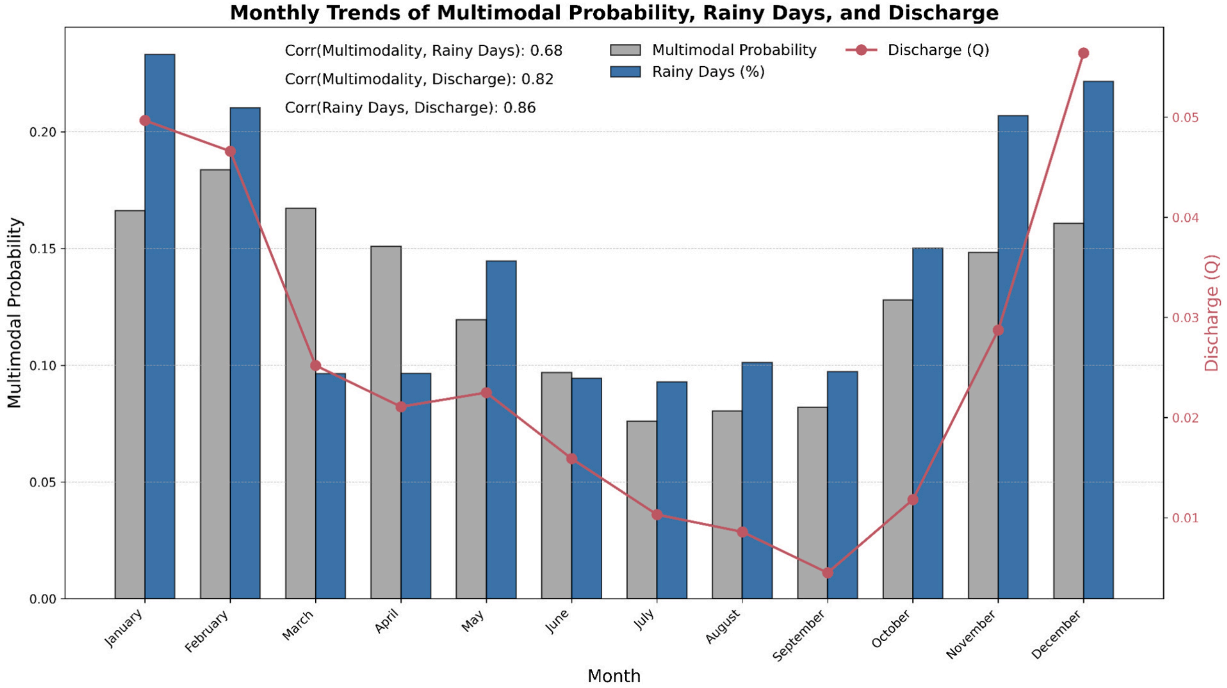


Fig. 5. Representation of the probability of multimodality occurrence and its correlation with discharge measurements and number of rainy days per month. Grey bar – probability of multimodality occurrence; blue bar – percentage of rainy days; red line – average monthly discharge measurements.

questions stated at the end of the introduction.

4.1. *Dynamical accuracy – how effectively can GMMs coupled with LSTMs predict in-situ soil moisture dynamics across different spatial and temporal scales and soil depths?*

This study finds that the two developed models are capable of matching or outperforming the selected benchmark models while also delivering accurately replicated estimates of variability. Analysis of the performance with respect to dynamics shows that the model is able to track the dynamical changes, while also accounting for the correlations and representing the absolute values of soil moisture conditions better than the benchmarks.

As mentioned in Sec. 1., several studies have investigated the application of DL for soil moisture modeling (Wang et al., 2024; Liu et al., 2022; Karthikeyan and Mishra, 2021; Datta and Faroughi, 2023). Specifically, Liu et al. (2022) reported an average root mean square error of 0.034 m^3 m^{-3} across all modeled sites, which aligns closely with our findings of an average error of 0.033 m^3 m^{-3} for the *condLSTM* and 0.036 m^3 m^{-3} for the *3G-LSTM* (calculated mean squared error (MSE) and root mean squared error (RMSE) values can be found in Table F1, Appendix F). Karthikeyan and Mishra (2021) achieved RMSE values in the range of approximately 0.04 m^3 m^{-3}, further supporting the consistency of these results across studies. Specifically, their models reported RMSE values of 0.034 m^3 m^{-3} at a depth of 10 cm, 0.030 m^3 m^{-3} at 20 cm, and 0.032 m^3 m^{-3} at 50 cm. Overall, these results both demonstrate a similar magnitude of error and reveal a consistent trend across varying soil depths, highlighting the robustness of the modeling approaches across different scenarios and conditions. Building on their work we integrate GMMs to ensure the model effectively captures soil moisture dynamics, allowing for a more accurate assessment of its ability to replicate fine-scale variations — the core focus of our research.

Our results show that our models outperform the *ERA5-Land* benchmark when assessed using KGE. However, when comparing rank correlation values, the results are more nuanced — in some instances the *ERA5-Land* achieves higher rank correlation scores. This suggests that while our models perform reasonably well in representing variability and matching key statistical properties of the data (as represented by the KGE), the ability to preserve the relative ordering of data points (as represented by rank correlation) is comparable between our models and *ERA5-Land*, with no dominance in either direction. Higher rank correlation values for *ERA5-Land* in certain sandstone and deeper marl areas might stem from the minimal variability in soil moisture measurements, which often appear as flatter time series. Its smoother estimates, due to broad-scale aggregation and interpolation, align better with these low-variability conditions, leading to higher rank correlation values. Another potential explanation for why *ERA5-Land*, despite being a coarse-resolution product, achieves comparable or occasionally superior performance lies in the differences of model structure and input data. First, *ERA5-Land* incorporates a significantly larger number of input variables (compared to our models), which allows it to leverage a broader range of physical and environmental factors, potentially enhancing its ability to capture certain dynamics. Second, the *ERA5-Land* model is constrained by physical boundaries, including porosity (upper bounds) and the permanent wilting point (lower bounds), which restricts the range of possible outcomes. However, Zehe et al. (2019) reported observed values that fall below the permanent wilting point, using the same dataset (i.e., data from the Attert experimental basin). Hence, it might in actuality be possible for our models to learn and preserve such information. Since our evaluation of rank correlation is based on site-level averages, the metric becomes sensitive to such outliers, potentially leading to a decrease in the performance of our models relative to *ERA5-Land*. However, while achieving reasonably strong rank correlation values is important, our model provides a significant advantage over *ERA5-Land:* it effectively learns the distribution and variability of the data. Unlike *ERA5-Land*, which focuses primarily on reproducing absolute values, our model emphasizes reproducing the actual spread of the data. This characteristic is critical for a wide range of applications, particularly those that depend on understanding variability and extremes rather than just average conditions.

On top of that, the performance of our models varies with different geological conditions and depths. The *condLSTM* performs well in schist geologies, while the *3G-LSTM* performs better in marl formations. Minimal differences are observed in sandstone, potentially due to its uniform soil moisture distribution, where soil moisture ranges are more evenly spread. KGE values tend to decline with depth, likely reflecting reduced sensitivity to surface meteorological inputs, while rank correlation tends to increase, suggesting that deeper trends are easier to predict despite deviations in absolute values. Interestingly, this depth-related drop in predictive performance (shown in Table 2) does not occur in the schist region. Previous work in the same experimental catchment (Demand et al., 2019; Jackisch et al., 2017) has demonstrated the significance of macropore flow and bypass mechanisms in schist geology, mechanisms that can preserve meteorological signals deeper in the soil profile. Thus, the upper boundary conditions may have a more pronounced and sustained impact on deeper soil moisture in schist, explaining why deeper layers there remain better correlated with surface forcing data. However, for the marl and sandstone regions, the decline in model performance at greater depths highlights the limitations of simplistic digital filtering techniques (Albergel et al., 2008) that attempt to infer deeper soil moisture conditions from surface-based measurements (e.g., satellite-derived moisture estimates). Such methods, which often assume a uniform vertical propagation of signals through the soil profile, may fail to capture the complex interplay of subsurface processes, geological heterogeneities, and preferential flow pathways, which lead to non-sequential responses. Consequently, such approaches may not reliably substitute more robust, process-based models or localized observations when estimating deeper soil moisture conditions, especially in areas characterized by complex subsurface processes, geological heterogeneities, and preferential flow pathways.

4.2. *Small-scale variability – how accurately can GMMS replicate and interpret inherent soil moisture variability?*

This study set out to push the boundaries of scale by focusing on replicating variability at the local level. Using a dataset with a unique measurement setup, we were able to introduce three time series for the same locations (within a 5-meter radius circle) and depths, allowing the model to train on these localized variations. By doing so, we addressed a key challenge in soil moisture modeling: representing small-scale variability next to dynamical outputs. This allowed the models to effectively capture variations across different geologies and depths (aggregated cluster sites) (Fig. 3), as well as at the site level (Fig. E1, Appendix E) and individual timesteps (Fig. 4).The models not only predict soil moisture dynamics but also provide insights into the potential range of changes within a 5-meter radius area, as well as for different geological subgroups. This dual capability – understanding both expected behavior and variability – is important for variability quantification and risk assessment. Our results reveal that soil moisture variability shows distinct behavior across different moisture states, suggesting that variability is not uniform but rather dependent on prevailing soil moisture conditions (Western et al., 1998). Under drier conditions, particularly when soil moisture is at or below 0.2 m^3 m^{-3}, spatial variability is generally low (Fig. 4). A likely explanation for this is that during summer, transpiration and root water uptake predominantly control soil water dynamics, and root water extraction is known to smooth out soil moisture spatially (Mälicke et al., 2020; Hildebrandt et al., 2016). As a result, drier soils tend to exhibit lower spatial variability.

In contrast, variability peaks in winter at around 0.3—–0.35 m^3 m^{-3}, where soil is close to the threshold at which macropore flow might emerge (Zehe et al., 2005). During these conditions, some sites might already be operating in preferential flow mode, while others remain dominated by matrix flow. This transition phase, where both flow mechanisms coexist, leads to maximum variability, aligning with previous findings of Zehe and Blöschl, (2004). However, similar peaks in variability are also evident in reanalysis products such as ERA5-Land (Li et al., 2020), which do not explicitly represent macropore processes. In-situ analyses of non-sequential sensor reactions (Graham and Lin, 2011) nevertheless confirm that preferential flow does occur in the field, with up to 25 % of the events at forested sites showing such signatures (Demand et al., 2019). The coexistence of sequential and non-sequential responses, and their seasonal modulation, thus provides a plausible explanation for the observed peaks in variability. As soil moisture approaches saturation, variability tends to decline, either because most sites operate under similar high-conductivity conditions allowing rapid percolation to depth, or because additional inputs cause only small relative changes in storage. Ryu and Famiglietti (2005) examined satellite-derived probability density functions of surface soil moisture at the footprint scale, and found similar soil moisture behavior – variability generally peaked in the midrange of mean soil moisture content and decreased toward the wetter and drier ends. They suggested that higher variability in the midrange is attributed to the multimodality of the soil moisture PDFs, which supposedly results from fractional precipitation within the footprint scale-fields.

We implemented two distinct modeling approaches: the *condLSTM*, which utilizes a single Gaussian distribution conditioned on the sensor ID information, and the *3G-LSTM*, which applies a mixture of three Gaussians without any additional sensor identifiers. In terms of overall predictive performance, both models show similar skill, with no statistically significant difference in their dynamical accuracy, training time, model complexity, or computational requirements. When analyzing performance across different depths and geologies, minor differences emerge: The *condLSTM* demonstrates superior performance in schist geologies, whereas the *3G-LSTM* performs slightly better in marl formations. Considering the models' ability to replicate the variability of soil moisture, the *3G-LSTM* provides slightly improved accuracy, suggesting that incorporating multiple Gaussian components enhances the representation of local-scale heterogeneity. To improve prediction accuracy across diverse geological settings, several strategies might be pursued: (1) employing an ensemble of the two models to combine their respective strengths, (2) fine-tuning model parameters using local training data to better capture site-specific interactions, and (3) incorporating geology-sensitive static features (e.g., permeability or porosity) into the input space to help the models adapt to different formations. Despite these variations, the observed differences remain rather small, indicating that either model is suitable for most applications.

A practical distinction lies in the type of information required: the *condLSTM* explicitly leverages site-specific identifiers, whereas the *3G-LSTM* achieves comparable results without them, instead learning spatial and temporal patterns directly from the data. Our primary aim in comparing these models was not to determine which performs better, but to assess whether such explicit information is necessary. The results indicate that the model can successfully infer this information from the data without manual input.

4.3. Multimodality and seasonal dependency of soil moisture variability

We found that our data-driven approach is able to uncover structures and relationships within the system, providing deeper insights into its dynamics than would be possible by training a deterministic model. We investigated the evolution of the distributional characteristics of our models over time in order to assess whether variations in the output distributions could be associated with changes in soil moisture states and environmental conditions. Fig. 5 illustrates the monthly variations in the likelihood of multimodality occurrence throughout the year, along with the similarity of these occurrences to discharge patterns and number of rainy days. During winter months the effect of precipitation, snowmelt and temperature changes can result in complicated soil moisture patterns. Apparently, this complexity cannot be properly described with unimodal outputs, causing the model to represent the predictions as a mixture of modes that reflect different plausible states of the soil moisture system. A possible reason for this is the dynamic nature of environmental conditions. Factors such as precipitation, evaporation, plant water uptake, and seasonal changes can influence soil moisture levels, making them more variable at certain times, which increases variability. The investigation by Albano et al. (2024) of how seasonality influences the performance of microwave satellite soil moisture products found that, under wetter conditions, soil moisture distributions exhibit more variability due to the increased water presence across all investigated eco-regions. Western et al. (2003) demonstrated through geo-statistical analysis that the spatial structure of soil moisture changes significantly between dry and wet conditions. Specifically, the nugget-to-sill ratios and the variogram ranges, which describe the degree of spatial variability and the scale of spatial correlation, show distinct differences when transitioning from dry to wet states. The analysis of temporal persistence of soil moisture patterns by Mälicke et al. (2020), revealed that distinct patterns persist in the wet and dry seasons, while soil water contents in the latter are spatially much more uniform compared to the wet season. All of those results are in line with our findings that months with increased water input result in highly variable soil moisture, with an enlarged probability of multimodality occurrence. However, during months like August and September, precipitation is lower, discharge decreases, and soil moisture dynamics are more stabilized so that the system tends to be more predictable, allowing our model to produce outputs with a smaller spread that reflect a single, dominant state. While the model effectively captures multimodal distributions, which may be physically explained, its relatively high capacity – 64 LSTM cell states – raises questions about which input factors contribute to the observed multimodality. A more interpretable framework would involve complexity control mechanisms to constrain the network's degrees of freedom. Potential approaches include regularizing the LSTM cell-state activations to encourage sparsity, progressively reducing hidden layer dimension until performance degrades, or introducing a bottleneck layer to project the hidden states onto a lower-dimensional space, facilitating insight into the drivers of multimodality (De la Fuente et al., 2023). While these strategies could improve interpretability by limiting model complexity and refining the explanation of multimodal behavior, they fall beyond the scope of this study and remain promising directions of future research.

4.4. Model limitations and future outlook

Accurate modeling of in-situ soil moisture dynamics remains complex due to several factors. The spatial variability of soil moisture can make it difficult for models to generalize across different scales and capture localized processes that influence moisture retention and movement. Additionally, the performance of soil moisture models depends on the quality and reliability of observational data. Issues such as calibration errors with soil moisture sensors, sensor-to-sensor variability, and sensitivity to soil properties can lead to measurement errors and biases, potentially degrading model performance (Jackisch et al., 2020). The availability of high-quality, high-resolution soil moisture data is crucial for training and evaluating DL models, but such data are often limited, particularly in remote or under-monitored regions. Data gaps, inconsistencies, and errors can complicate the training process, making it challenging for the model to generalize across various environments and conditions. In our case the aforementioned issues occurred in multiple ways. Many time series belonging to different site locations had contained gaps of various sizes, caused by sensor maintenance or changes. To tackle this issue, we applied temporal segmentation of the data (to avoid having data from a specific location only in training or evaluation subset), but due to the frequency of these gaps, this approach did not provide us a perfect solution, resulting in a few locations where either the evaluation or the training data is partly or completely absent.

In addition, as mentioned in Section 2.1, and due to the previously described issue, geological formations with differing characteristics often had uneven data availability, with some locations providing more reliable data than others. This disparity makes it difficult to conduct a fair and consistent performance comparison across all locations. The imbalance in data availability can potentially introduce bias into the model evaluation, as areas with more abundant or higher-quality data may yield better predictions, while regions with sparse or inconsistent data might appear to underperform. Nevertheless, we assume that the model setup and structure presented here should be applicable to regions and catchments with similar hydrological, geological and meteorological conditions. Accordingly, these models may not perform as effectively if these conditions differ. To overcome this limitation, a critical step towards developing a more generalized model will be to ensure it can provide accurate predictions for ‘unseen’ locations. Kratzert et al. (2019) showed that LSTMs trained on large, diverse datasets from different regions (in their study, used to predict runoff in river basins) can perform well even at ungauged locations. Further, their models outperformed traditional benchmark models that were calibrated locally. We therefore anticipate that training the models on an expanded dataset, such as the International Soil Moisture Network (ISMN), would enhance its generalizability and robustness across diverse conditions. Incorporating extensive soil moisture observations would enable the model to better capture a broader range of spatial and temporal variability, improving its applicability to larger-scale drought monitoring and forecasting. This approach aligns with efforts like the UFZ Drought Monitor (near real-time, high-resolution drought assessments for Germany) (Zink et al., 2016), but goes further by integrating local-scale soil moisture variability in a more nuanced manner. By bridging larger-scale drought assessment with local-scale variability, this extension should advance the accuracy and utility of such monitoring and early-warning systems.

Future work could include comparisons with the results of process-based hydrological models to further contextualize the performance of the proposed approach. Our current use of ERA5-Land as a benchmark already partially addresses this, as the products embed physical relationships through the 1D Darcy-Richards equation. Nonetheless, direct comparison with physically based models that are locally trained and set up, or integration within hybrid frameworks such as physics-informed neural networks (PINNs), could provide deeper insights and strengthen the link between data-driven predictions and hydrological process understanding.

5. Summary and conclusions

In this study, we developed and evaluated two different GMM setups to capture the dynamics and variability of soil moisture in the Attert experimental basin. Leveraging a unique dataset of nine sensor measurements belonging to the same site – three sensors at three depths within the same cluster – we modeled variability on different scales on 43 locations with 387 sensors. The use of GMMs allowed us to represent soil moisture as a distribution rather than a single estimate, making it a feasible approach that aligns with the variability inherent in soil moisture data. We demonstrated that our models effectively capture both the temporal dynamics and spatial variability of soil moisture across different cluster site locations, geologies, and depths. Our evaluation of *ERA5-Land* reanalysis data showed that while *ERA5-Land* performs remarkably well in capturing rank and non-exceedance probabilities, it exhibits a strong bias – a particularly critical issue in the context of increasing global drying trends. The found multimodality in predictions reflects the models’ sensitivity to variability, serving as an indicator of heightened hydrological complexity and alternative flow pathways. This aligns with the findings of Demand et al. (2019), where non-sequential soil moisture responses – linked to preferential flow bypassing upper layers – were more frequent under specific seasonal and moisture conditions. The coexistence of sequential and non-sequential responses within a cluster site suggests that heterogeneous infiltration and preferential flow play a key role in shaping soil moisture distributions. Moving forward, the development of soil moisture reanalysis data using deep learning and the ISMN holds potential to improve the accuracy and representation of soil moisture dynamics and variability. By integrating such models with large-scale, high-resolution observational datasets, future research can increase soil moisture simulation accuracy, better quantify variability, and refine our understanding of the underlying hydrological processes.

CRediT authorship contribution statement

Balazs Bischof: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Daniel Klotz:** Writing – review & editing. **Hoshin V. Gupta:** Writing – review & editing. **Erwin Zehe:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Ralf Loritz:** Writing – review & editing, Validation, Supervision, Software, Project administration, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge Theresa Blume (GFZ German Research Center for Geosciences) for collecting and providing the in-situ soil moisture data used for this study.

Appendix A. . Input data

.

Table A1
Summary table of the used input and target variables including general information, source and resolution.

Name	Source	Resolution	Supplementary information	Units
Elevation	Luxembourg Institute of Science and Technology	5x5 meter	DEM of the Attert basin. LIDAR scan, provided in EPSG:2169 with cm precision.	m
Soil type	Origine Service de pedologie (Administration des services techniques	100 x 100 m	Pedological GIS data. The map is a spatial combination of the best available data merged from the sources. Here 6 classes were	Class

(continued on next page)

Table A1 (*continued*)

Name	Source	Resolution	Supplementary information	Units
	de l'agriculture); Service Public de Wallonie		employed, namely: silty soils, gravelly soils, sandy-loamy soils, loamy-sandy soils, clayes soils, and peat soils.	
Land use	European Environmental Agency (EEA)	500 x 500 m	CORINE Land Cover dataset, including satellite observations from 2011 to 2012. Here 6 classes were employed. namely complex cultivation patterns, pastures, broad-leaved forests, mixed forests, agricultural land, and coniferous forests.	Class
Volumetric water content	ERA5-Land	0.1° x 0.1°; hourly	Volume of water in four different soil layers (0–7 cm, 7–28 cm, 28–100 cm, and 100–289 cm) of the ECMWF Integrated Forecasting System.	m^3m^{-3}
Precipitation	CAOS project dataset; ASTA (Administration des services techniques de l'agriculture, Luxembourg)	Pixel; nearest pixel to site location; hourly	Rainfall radar data for the Attert-catchment; nearest pixel to site location	mm/h
Potential evapotranspiration	CAOS project dataset	11 stations; nearest station to site location; hourly	Evaporation data estimated with the Penman-Monteith equation across 11 meteorological measurement stations	mm/h
Air temperature	CAOS project dataset	Site-level measurements; hourly	Site-level hourly air temperature measurements (43 cluster-site locations – Attert-catchment)	K
Discharge	Luxembourg Institute of Science and Technology	Closest location to cluster site; hourly	Hourly discharge measurements within 4 catchments	m^3s^{-1}
Soil moisture	CAOS project dataset	43 cluster. site locations; hourly	Soil observations from cluster stations in the Attert basin, 43 locations – each location consist of three sensor measurements and three depths	m^3m^{-3}

Appendix B. . Variability metrics: Wasserstein-distance and Continuous rank probability Score

Wasserstein-distance:

The Wasserstein-distance quantifies the cost of transforming one probability distribution into another. For two probability distributions P and Q, the first order Wasserstein-distance is defined as:

$$W_1(P,Q) = inf_{\gamma\in\Gamma(P,Q)}\int\left|x-y\right|d\gamma(x,y)$$

where $\Gamma(P,Q)$ represents the set of all joint distributions with marginals P and Q.

Continuous Rank Probability Score (CRPS).

The CRPS assesses the accuracy of the probabilistic forecast by comparing the cumulative distribution function (CDF) of a forecasted distribution $F(x)$ to the observed value x_0. It is defined as:

$$CRPS(F,x_0) = \int_{-\infty}^{\infty}(F(x) - 1(x \geq xo))^2 dx$$

where $1(\cdot)$ is the Heaviside step function. The CRPS generalizes the Mean Absolute Error (MAE) to probabilistic forecasts, producing sharp and well-calibrated predictions.

Log-likelihood and log-likelihood ratio

The log-likelihood at each timestep t for a model with predicted variability σ_t and observed variability x_t is:

$$logL_t = -\frac{1}{2}\left[log(2\pi\sigma_t^2) + \frac{(x_t-\mu)^2}{\sigma_t^2}\right]$$

where μ is the expected mean of the distribution. The average log-likelihood per timesteps over N timesteps is:

$$\underline{l}ogL_t = \frac{1}{N}\sum_{t=1}^{N}logL_t$$

For two models (model 1 – *condLSTM*; model 2 – *3G-LSTM*), the log likelihood ratio per timestep is:

$$\Delta\underline{l}ogL = \underline{l}ogL_1 - \underline{l}ogL_2$$

where $\Delta\underline{l}ogL > 0$ favors the *condLSTM*, and $\Delta\underline{l}ogL < 0$ favors the *3G-LSTM*.

What are the metrics telling us: example for ‘good’ and ‘poor’ values.

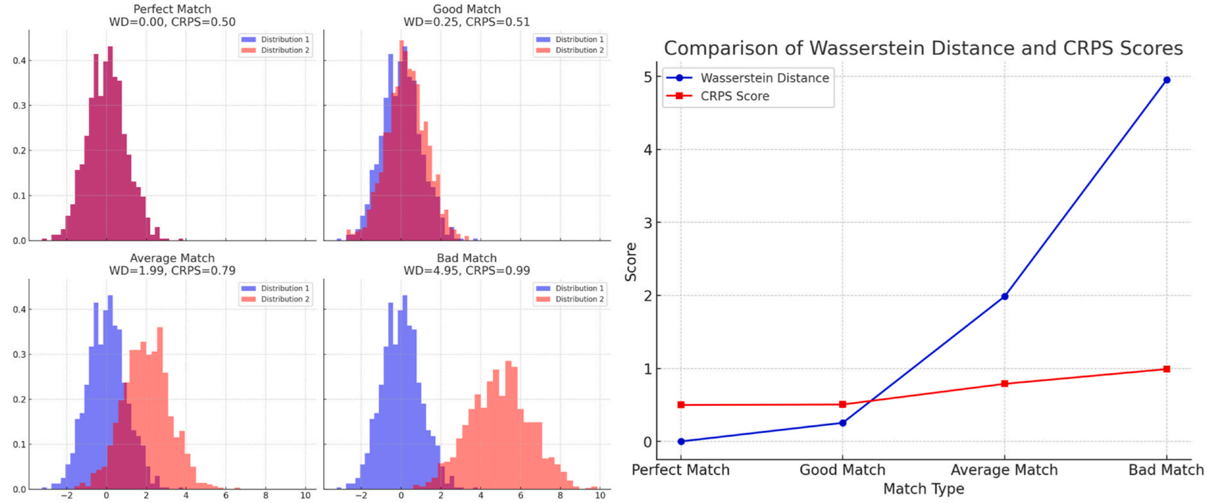


Fig. B1. Randomly generated distributions to showcase a perfect, a good, an average, and a bad match. The calculated values of the corresponding Wasserstein-distance and CRPS scores provide a guideline to assess performance accuracy and quality.

Appendix C. . Example time series comparison

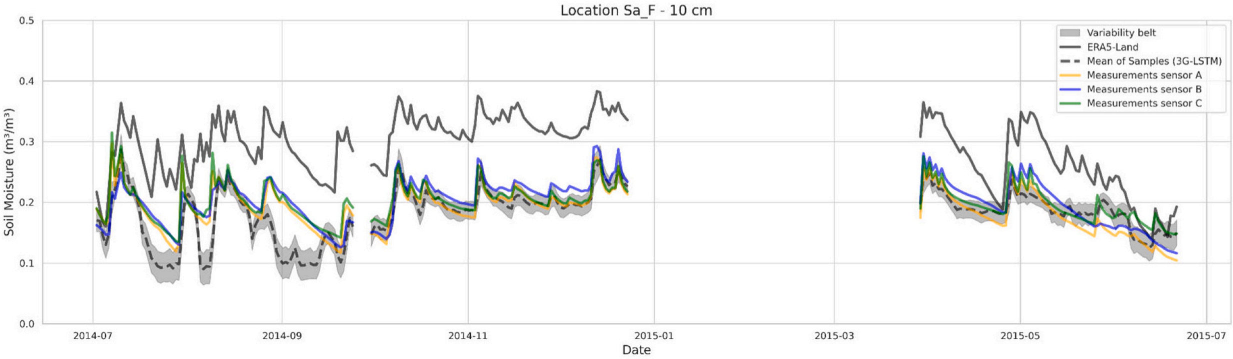


Fig. C1. Example time series plot of the calculated 80-20th percentile variability belt of the *3G-LSTM* (grey shaded area), the mean of predictions (dashed black line), the three sensor observations belonging to the same cluster site (yellow, green, and blue lines), and the corresponding *ERA5-Land* volumetric water content simulation. The plot shows an example of a sandstone location at 10 cm, aiming to visualize the existing bias in between *ERA5-Land* and observations.

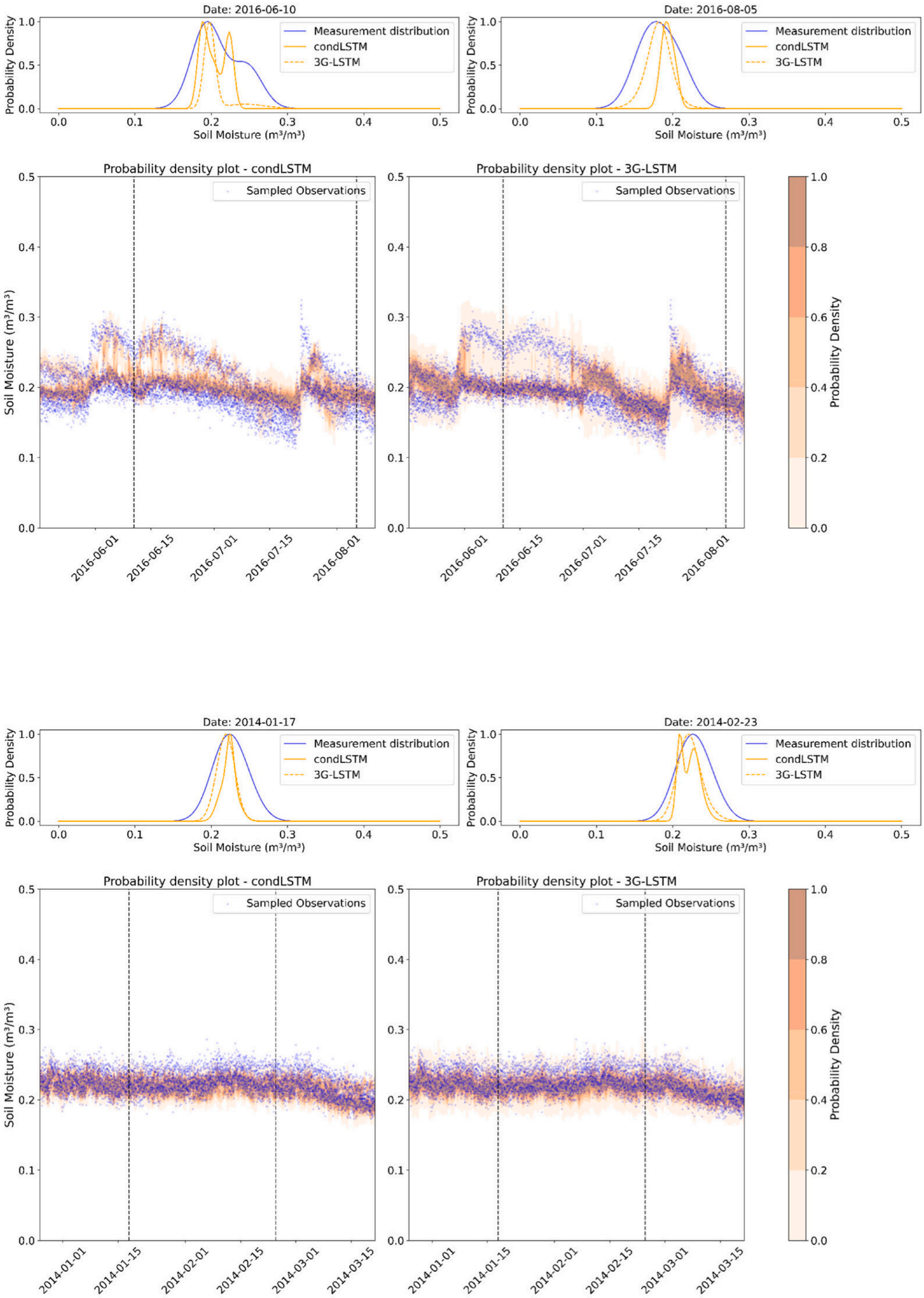


Fig. C2. Two additional example time series plots of schist location at 30 cm depth (upper plot), and sandstone location 50 cm depth (bottom plot). The plot shows randomly drawn samples from the measurement distribution (combined distribution of the three sensor measurements from a single site – scattered blue points) and the weighted distribution of the *condLSTM* (left) and *3G-LSTM* (right). The upper two distribution plots show two randomly selected timesteps and the corresponding distribution of soil moisture measurements (blue line), *condLSTM* predictions (solid orange line), and *3G-LSTM* predictions (dotted orange line).

Appendix D. . Aggregated distribution of model mixture outputs and data measurements

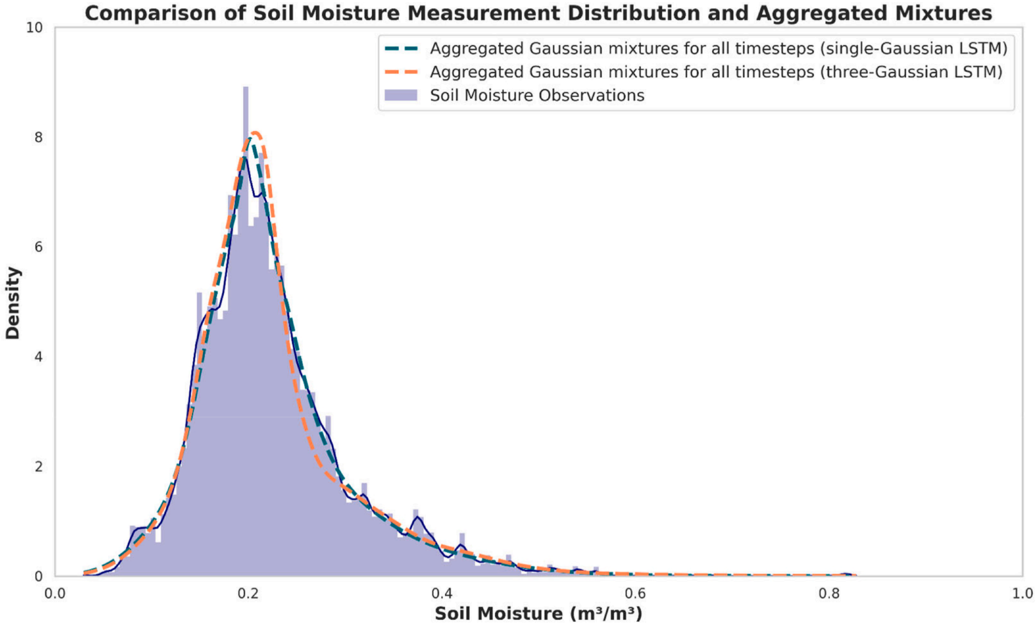


Fig. D1. Comparison of the aggregated mixture distributions of the developed models with the actual soil moisture data distribution.

Appendix E. . Cluster-site level variability of soil moisture

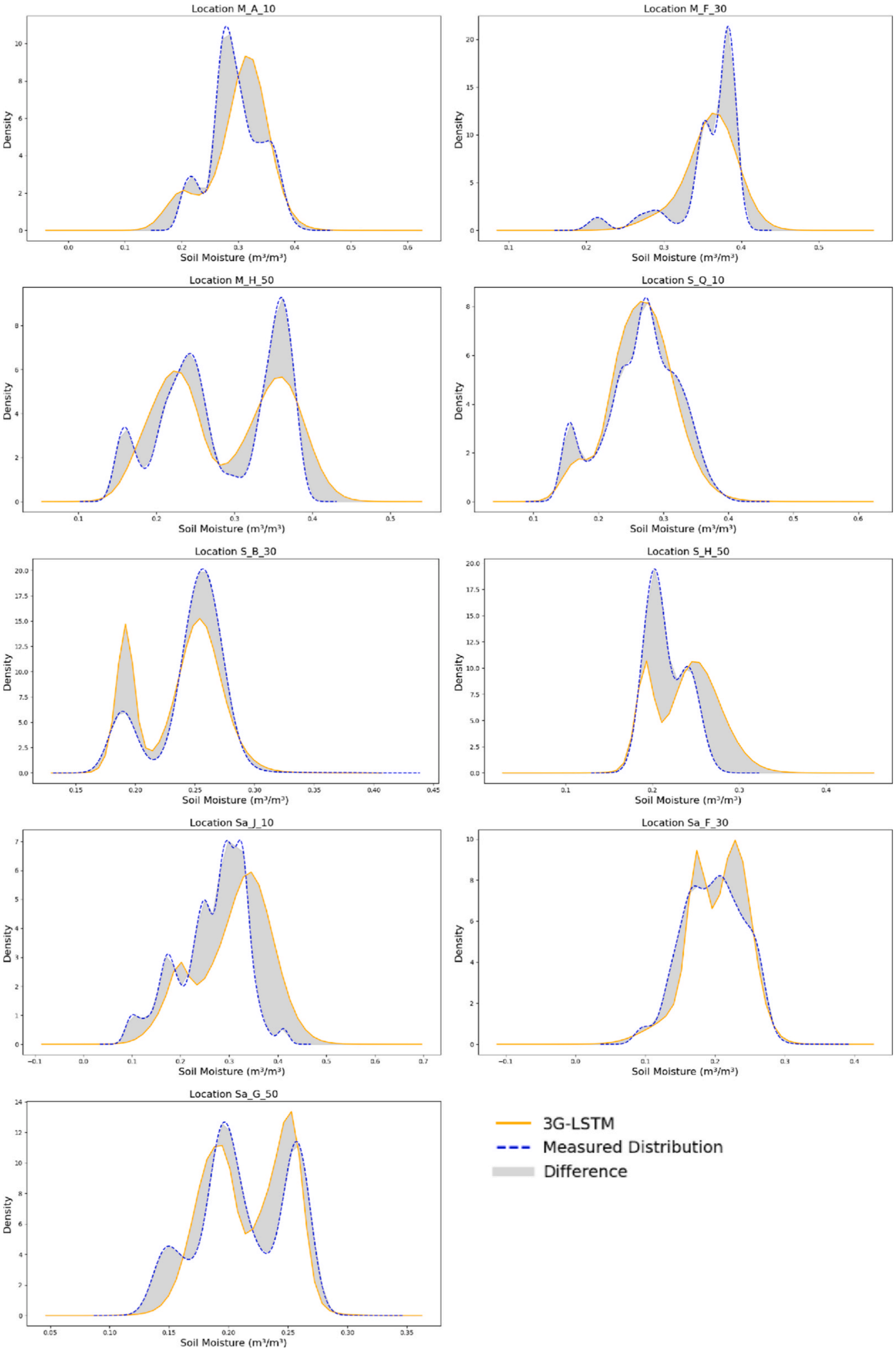


Fig. E1. Site-level variability comparison for example sites from all geologies and depths between the aggregated mixtures of the *3G-LSTM* and the distribution of observational data sensor measurements.

Appendix F. . Complete tables for the used evaluation metrics

.

Table F1
MSE and RMSE values of the developed GM-LSTM models and *ERA5-Land* for each geology and depth.

Location	MSE			RMSE		
	cond LSTM	*3G-LSTM*	*ERA5 −Land*	*cond LSTM*	*3G-LSTM*	*ERA5 −Land*
Total average	0.0011	0.0013	0.0120	0.0334	0.0364	0.1097
Total average 10 cm	0.0017	0.0017	0.0103	0.0410	0.0414	0.0958
Total average 30 cm	0.0008	0.0011	0.0125	0.0278	0.0332	0.1058
Total average 50 cm	0.0009	0.0011	0.0134	0.0282	0.0332	0.1067
Marls − 10 cm	0.0020	0.0020	0.0051	0.0438	0.0441	0.0704
Marls − 30 cm	0.0011	0.0015	0.0033	0.0330	0.0378	0.0577
Marls − 50 cm	0.0017	0.0019	0.0033	0.0412	0.0435	0.0574
Schist − 10 cm	0.0015	0.0016	0.0204	0.0387	0.0396	0.1429
Schist − 30 cm	0.0006	0.0012	0.0201	0.0239	0.0340	0.1416
Schist − 50 cm	0.0003	0.0007	0.0272	0.0181	0.0265	0.1650
Sandstone − 10 cm	0.0017	0.0017	0.0055	0.0405	0.0407	0.0742
Sandstone − 30 cm	0.0007	0.0008	0.0140	0.0263	0.0278	0.1181
Sandstone − 50 cm	0.0007	0.0008	0.0095	0.0254	0.0364	0.0977

.

Table F2
CRPS and Wasserstein-values for the developed models including all depths and geologies.

Location	CRPS		Wasserstein-distance		Log-likelihood		
	condLSTM	*3G-LSTM*	*condLSTM*	*3G-LSTM*	*condLSTM*	*3G-LSTM*	*Ratio*
Total average	0.029	0.028	0.022	0.018	1.326	1.371	−0.042
Marls − 10 cm	0.028	0.026	0.046	0.040	1.403	1.478	−0.075
Marls − 30 cm	0.022	0.020	0.027	0.017	1.641	1.788	−0.148
Marls − 50 cm	0.031	0.029	0.024	0.017	1.318	1.399	−0.051
Schist − 10 cm	0.052	0.052	0.029	0.025	0.181	0.262	−0.081
Schist − 30 cm	0.043	0.041	0.012	0.014	0.479	0.255	0.224
Schist − 50 cm	0.024	0.023	0.006	0.015	1.675	1.648	0.026
Sandstone − 10 cm	0.026	0.025	0.024	0.018	1.516	1.603	−0.088
Sandstone − 30 cm	0.019	0.018	0.014	0.008	1.786	1.899	−0.113
Sandstone − 50 cm	0.016	0.015	0.012	0.007	1.937	2.009	−0.072

Appendix G. . Figures of sensor measurement difference distributions and those of both model outputs

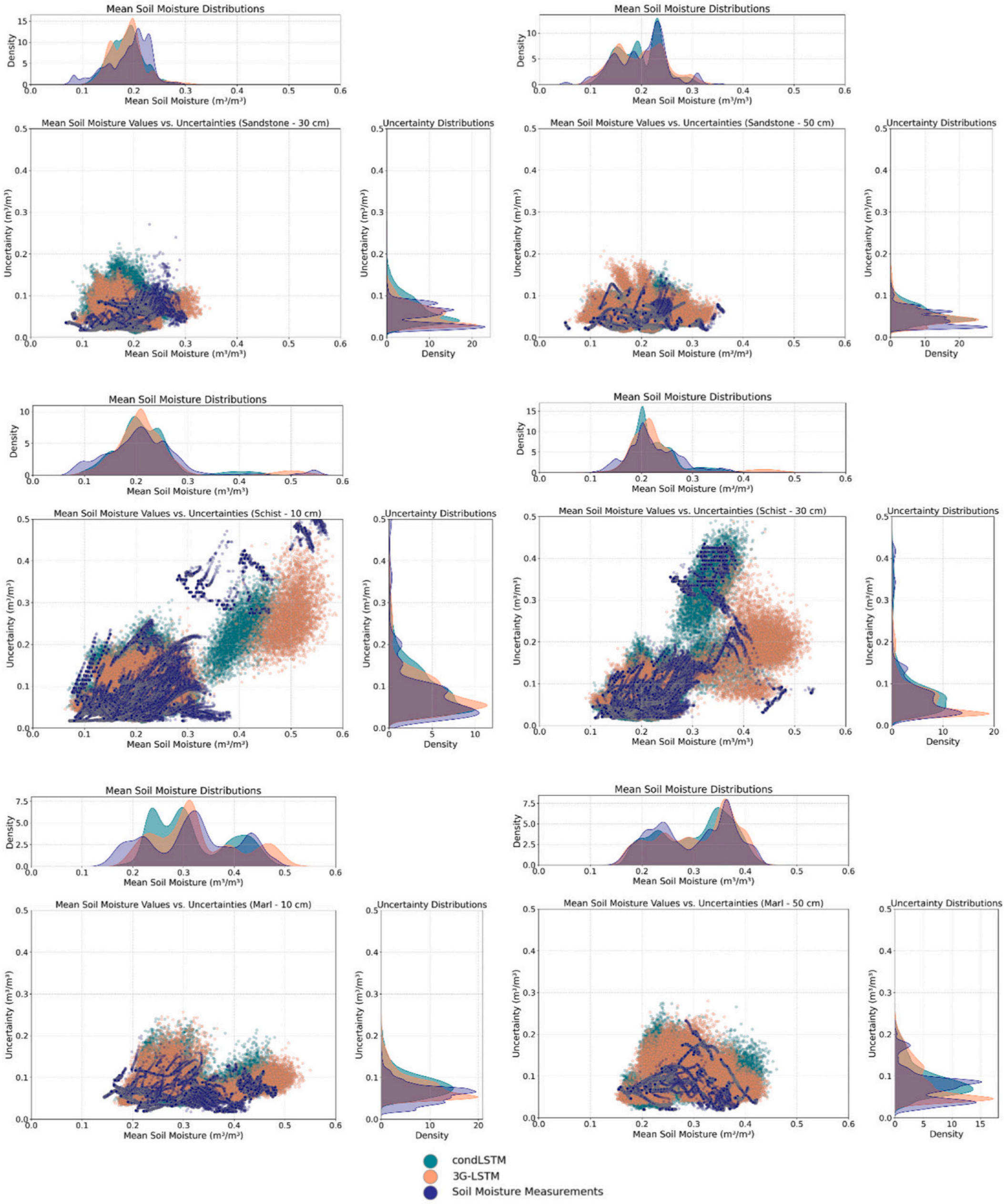


Fig. G1. Geology level variability comparison for the remaining geology-depth combinations between the mixtures of both models and the distribution of observational data sensor measurements.

During the preparation of this work the author(s) used Chat GPT in order to improve language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Data availability

The authors do not have permission to share data.

References

Albano, R., Lacava, T., Mazzariello, A., Manfreda, S., Adamowski, J., Sole, A., 2024. How can seasonality influence the performance of recent microwave satellite soil moisture products? Remote Sens. 16, 3044. https://doi.org/10.3390/rs16163044.

Albergel, C., Rüdiger, C., Pellarin, T., Calvet, J.-C., Fritz, N., Froissard, F., Suquia, D., Petitpa, A., Piguet, B., Martin, E., 2008. From near-surface to root-zone soil moisture using an exponential filter: an assessment of the method based on in-situ observations and model simulations. Hydrol. Earth Syst. Sci. 12, 1323–1337. https://doi.org/10.5194/hess-12-1323-2008.

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M.: Crop evapotranspiration—Guidelines for computing crop water requirements, FAO Irrigation and Drainage Paper 56, Food and Agriculture Organization of the United Nations, Rome, Italy, 1998.

Administration des services techniques de l'agriculture (ASTA): AgriMeteo Luxembourg, available at: http://www.agrimeteo.lu/, last access: 2023.

Bárdossy, A., Lehmann, W., 1998. Spatial distribution of soil moisture in a small catchment. Part 1: geostatistical analysis. J. Hydrol. 206, 1–15.

Beven, K., Germann, P., 1982. Macropores and water flow in soils. Water Resour. Res. 18 (5), 1311–1325. https://doi.org/10.1029/WR018i005p01311.

Bishop, C. M.: Mixture density networks, Technical Report, Aston University, Birmingham, UK, 1994.

Bishop, C. M.: Pattern Recognition and Machine Learning, Springer, 2006.

Brocca, L., Melone, F., Moramarco, T., Morbidelli, R., 2010. Spatial-temporal variability of soil moisture and its estimation across scales. Water Resour. Res. 46, W02410. https://doi.org/10.1029/2009WR008016.

Bronstert, A., Creutzfeldt, B., Graeff, T., et al., 2012. Potentials and constraints of different types of soil moisture observations for flood simulations in headwater catchments. Nat. Hazards 60, 879–894. https://doi.org/10.1007/s11069-011-9874-9.

Datta, P., Faroughi, S.A., 2023. A multihead LSTM technique for prognostic prediction of soil moisture. Geoderma 433, 116452. https://doi.org/10.1016/j.geoderma.2023.116452.

De la Fuente, L.A., Ehsani, M.R., Gupta, H.V., Condon, L.E., 2023. Towards interpretable LSTM-based modelling of hydrological systems. Egusphere [preprint]. https://doi.org/10.5194/egusphere-2023-666.

Demand, D., Blume, T., Weiler, M., 2019. Spatio-temporal relevance and controls of preferential flow at the landscape scale. Hydrol. Earth Syst. Sci. 23, 4869–4889. https://doi.org/10.5194/hess-23-4869-2019.

European Environment Agency (EEA): CORINE Land Cover 2018 (vector), Europe, 6-yearly - version 2020_20u1, available at: , last access: May 2020.

Famiglietti, J.S., Ryu, D., Berg, A.A., Rodell, M., Jackson, T.J., 2008. Field observations of soil moisture variability across scales. Water Resour. Res. 44, W01423. https://doi.org/10.1029/2006WR005804.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102, 359–378. https://doi.org/10.1198/016214506000001437.

Graham, C.B., Lin, H.S., 2011. Controls and frequency of preferential flow occurrence: a 175-event analysis. Vadose Zone J. 10, 816–831. https://doi.org/10.2136/vzj2010.0119.

Grayson, R.B., Western, A.W., 1998. Towards areal estimation of soil water content from point measurements: time and space stability of mean response. J. Hydrol. 207, 68–82. https://doi.org/10.1016/S0022-1694(98)00096-1.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J. Hydrol. 377, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.

Herbst, M., Diekkruger, B., 2003. Modeling the spatial variability of soil moisture in a micro-scale catchment and comparison with field data using geostatistics. Phys. Chem. Earth, Parts a/b/c 28, 239–245.

Hildebrandt, A., Kleidon, A., Bechmann, M., 2016. A thermodynamic formulation of root water uptake. Hydrol. Earth Syst. Sci. 20, 3441–3454. https://doi.org/10.5194/hess-20-3441-2016.

Hochreiter, S., 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int. J. Uncertainty Fuzziness Knowledge-Based Syst. 6, 107–116. https://doi.org/10.1142/S0218488598000094.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1785. https://doi.org/10.1162/neco.1997.9.8.1735.

Hosseini, R., Newlands, N.K., Dean, C.B., Takemura, A., 2015. Statistical modeling of soil moisture, integrating satellite remote-sensing (SAR) and ground-based data. Remote Sens. 7, 2752–2780. https://doi.org/10.3390/rs70302752.

Jackisch, C., Angermann, L., Allroggen, N., Sprenger, M., Blume, T., Tronicke, J., Zehe, E., 2017. Form and function in hillslope hydrology: in situ imaging and characterization of flow-relevant structures. Hydrol. Earth Syst. Sci. 21, 3749–3775. https://doi.org/10.5194/hess-21-3749-2017.

Jackisch, C., Germer, K., Graeff, T., et al., 2020. Soil moisture and matric potential – an open field comparison of sensor systems. Earth Syst. Sci. Data 12, 683–697. https://doi.org/10.5194/essd-12-683-2020.

Kaplan, N.H., Sohrt, E., Blume, T., Weiler, M., 2019. Monitoring ephemeral, intermittent, and perennial streamflow: a dataset from 182 sites in the Attert Catchment, Luxembourg. Earth Syst. Sci. Data 11, 1363–1374. https://doi.org/10.5194/essd-11-1363-2019.

Karthikeyan, L., Mishra, A.K., 2021. Multi-layer high-resolution soil moisture estimation using machine learning over the United States. Remote Sens. Environ. 266, 112706. https://doi.org/10.1016/j.rse.2021.112706.

Kaur, G., Singh, G., Motavalli, P.P., Nelson, K.A., Orlowski, J.M., Golden, B.R., 2020. Impacts and management strategies for crop production in waterlogged/flooded soils: a review. Agron. J. 112, 1475–1501. https://doi.org/10.1002/agj2.20093.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., Nearing, G., 2022. Uncertainty estimation with deep learning for rainfall–runoff modeling. Hydrol. Earth Syst. Sci. 26, 1673–1693. https://doi.org/10.5194/hess-26-1673-2022.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., Nearing, G., 2019. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrol. Earth Syst. Sci. 23, 5089–5110. https://doi.org/10.5194/hess-23-5089-2019.

Koehler, B., Zehe, E., Corre, M.D., Veldkamp, E., 2010. An inverse analysis reveals limitations of the soil-CO_2 profile method to calculate CO_2 production and efflux for well-structured soils. Biogeosciences 7, 2311–2325. https://doi.org/10.5194/bg-7-2311-2010.

Köhne, S., Lennartz, B., Köhne, J.M., Simunek, J., 2006. Bromide transport at a tile-drained field site: experiment, and one- and two-dimensional equilibrium and nonequilibrium numerical modeling. J. Hydrol. 321, 390–408. https://doi.org/10.1016/j.jhydrol.2005.08.010.

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., Dadson, S.J., 2022. Hydrological concept formation inside long short-term memory (LSTM) networks. Hydrol. Earth Syst. Sci. 26, 3079–3101. https://doi.org/10.5194/hess-26-3079-2022.

Li, M.X., Wu, P.L., Ma, Z.G., Lv, M.X., Yang, Q., 2020. Changes in soil moisture persistence in China over the past 40 years under a warming climate. J. Clim. 33, 9531–9550. https://doi.org/10.1175/JCLI-D-19-0900.1.

Liu, J., Rahmani, F., Lawson, K., Shen, C., 2022. A multiscale deep learning model for soil moisture integrating satellite and in-situ data. Geophys. Res. Lett. 49 (Issue 7). https://doi.org/10.1029/2021GL096847.

Liu, Y., Gupta, H.V., 2007. Uncertainty in hydrological modeling: toward an integrated data assimilation framework. Water Resour. Res. 43 (Issue 7). https://doi.org/10.1029/2006WR005756.

Loritz, R., Hassler, S.K., Jackisch, C., Allroggen, N., van Schaik, L., Wienhöfer, J., Zehe, E., 2017. Picturing and modeling catchments by representative hillslopes. Hydrol. Earth Syst. Sci. 21, 1225–1249. https://doi.org/10.5194/hess-21-1225-2017.

Loritz, R., Wu, C.H., Klotz, D., Gauch, M., Kratzert, F., Bassiouni, M., 2024. Generalizing tree–level sap flow across the European continent. Geophys. Res. Lett. 51, e2023GL107350. https://doi.org/10.1029/2023GL107350.

Luxembourg Institute of Science and Technology (LIST): available at: https://www.list.lu/, last access: 2023.

Maier, H.R., Zheng, F., Gupta, H., Chen, J., Mai, J., Savic, D., Loritz, R., Wu, W., Guo, D., Bennett, A., Jakeman, A., Razavi, S., Zhao, J., 2023. On how data are partitioned in model development and evaluation: confronting the elephant in the room to enhance model generalization. Environ Model Softw. 167, 105779. https://doi.org/10.1016/j.envsoft.2023.105779.

Mälicke, M., Hassler, S.K., Blume, T., Weiler, M., Zehe, E., 2020. Soil moisture: variable in space but redundant in time. Hydrol. Earth Syst. Sci. 24, 2633–2653. https://doi.org/10.5194/hess-24-2633-2020.

Manoj, J.A., Loritz, R., Villinger, F., Mälicke, M., Koopaeidar, M., Göppert, H., Zehe, E., 2024. Toward flash flood modeling using gradient resolving representative hillslopes. Water Resour. Res. 60. https://doi.org/10.1029/2023wr036420.

Martínez-Carreras, N., Krein, A., Gallart, F., et al., 2012. The influence of sediment sources and hydrologic events on the nutrient and metal content of fine-grained sediments (Attert River Basin, Luxembourg). Water Air Soil Pollut. 223, 5685–5705. https://doi.org/10.1007/s11270-012-1307-1.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D.G., Piles, M., Rodríguez-Fernández, N.J., Zsoter, E., Buontempo, C., Thépaut, J.-N., 2021. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. Earth Syst. Sci. Data 13, 4349–4383. https://doi.org/10.5194/essd-13-4349-2021.

Nearing, G.S., Kratzert, F., Sampson, A.K., Pelissier, C.S., Klotz, D., Frame, J.M., et al., 2021. What role does hydrological science play in the age of machine learning? Water Resour. Res. 57, e2020WR028091. https://doi.org/10.1029/2020WR028091.

Neuper, M., Ehret, U., 2019. Quantitative precipitation estimation with weather radar using a data- and information-based approach. Hydrol. Earth Syst. Sci. 23, 3711–3733. https://doi.org/10.5194/hess-23-3711-2019.

Pfister, L., Martínez-Carreras, N., Hissler, C., Klaus, J., Carrer, G.E., Stewart, M.K., McDonnell, J.J., 2017. Bedrock geology controls on catchment storage, mixing, and release: a comparative analysis of 16 nested catchments. Hydrol. Process. 31, 1828–1845. https://doi.org/10.1002/hyp.11134.

Robinson, D.A., Campbell, C.S., Hopmans, J.W., Hornbuckle, B.K., Jones, S.B., Knight, R., Ogden, F., Selker, J., Wendroth, O., 2008. Soil moisture measurement for ecological and hydrological watershed-scale observatories: a review. Vadose Zone J. 7, 358–389. https://doi.org/10.2136/vzj2007.0143.

Ryu, D., Famiglietti, J.S., 2005. Characterization of footprint-scale surface soil moisture variability using Gaussian and beta distribution functions during the Southern Great Plains 1997 (SGP97) hydrology experiment. Water Resour. Res. 41, W12402. https://doi.org/10.1029/2004WR003835.

Schaefli, B., Gupta, H.V., 2007. Do nash values have value? Hydrol. Process. 21, 2075–2080. https://doi.org/10.1002/hyp.6825.

Tietjen, B., Zehe, E., Jeltsch, F., 2009. Simulating plant water availability in drylands under climate change: a generic model of two soil layers. Water Resour. Res. 45, W01418. https://doi.org/10.1029/2007WR006589.

Tietjen, B., Jeltsch, F., Zehe, E., Classen, N., Groengroeft, A., Schiffers, K., Oldeland, J., 2010. Effects of climate change on the coupled dynamics of water and vegetation in drylands. Ecohydrology 3, 226–237. https://doi.org/10.1002/eco.70.

Villani, C., 2003. Topics in optimal transportation, graduate studies in mathematics. Bull. Amer. Math. Soc. (N.S.) Vol. 58.

Wang, Y., Shi, L., Hu, Y., Hu, X., Song, W., Wang, L., 2024. A comprehensive study of deep learning for soil moisture prediction. Hydrol. Earth Syst. Sci. 28, 917–943. https://doi.org/10.5194/hess-28-917-2024.

Western, A.W., Blöschl, G., Grayson, R.B., 1998. Geostatistical characterisation of soil moisture patterns in the Tarrawarra catchment. J. Hydrol. 205, 20–37. https://doi.org/10.1016/S0022-1694(97)00142-X.

Western, A. W., Grayson, R. B., Blöschl, G., and Wilson, D. J.: Spatial variability of soil moisture and its implications for scaling, in: Scaling Methods in Soil Physics, edited by: Pachepsky, Y., Radcliffe, D. E., and Selim, H. M., 119–142, CRC Press, Boca Raton, FL, 2003.

Wunsch, A., Liesch, T., Broda, S., 2021. Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). Hydrol. Earth Syst. Sci. 25, 1671–1687. https://doi.org/10.5194/hess-25-1671-2021.

Zehe, E., Blöschl, G., 2004. Predictability of hydrologic response at the plot and catchment scales: role of initial conditions. Water Resour. Res. 40, W10202. https://doi.org/10.1029/2003WR002869.

Zehe, E., Becker, R., Bárdossy, A., Plate, E., 2005. Uncertainty of simulated catchment runoff response in the presence of threshold processes: role of initial soil moisture and precipitation. J. Hydrol. 315, 183–202. https://doi.org/10.1016/j.jhydrol.2005.03.038.

Zehe, E., Loritz, R., Jackisch, C., Westhoff, M., Kleidon, A., Blume, T., Hassler, S.K., Savenije, H.H., 2019. Energy states of soil water – a thermodynamic perspective on soil water dynamics and storage-controlled streamflow generation in different landscapes. Hydrol. Earth Syst. Sci. 23, 971–987. https://doi.org/10.5194/hess-23-971-2019.

Zehe, E., Ehret, U., Pfister, L., Blume, T., Schröder, B., Westhoff, M., Jackisch, C., Schymanski, S.J., Weiler, M., Schulz, K., Allroggen, N., Tronicke, J., van Schaik, L., Dietrich, P., Scherer, U., Eccard, J., Wulfmeyer, V., Kleidon, A., 2014. HESS opinions: from response units to functional units: a thermodynamic reinterpretation of the HRU concept to link spatial organization and functioning of intermediate scale catchments. Hydrol. Earth Syst. Sci. 18, 4635–4655. https://doi.org/10.5194/hess-18-4635-2014.

Zehe, E., Graeff, T., Morgner, M., Bauer, A., Bronstert, A., 2010. Plot and field scale soil moisture dynamics and subsurface wetness control on runoff generation in a headwater in the Ore Mountains. Hydrol. Earth Syst. Sci. 14, 873–889. https://doi.org/10.5194/hess-14-873-2010.

Zink, M., Samaniego, L., Kumar, R., Thober, S., Mai, J., Schäfer, D., Marx, A., 2016. The German drought monitor. Environ. Res. Lett. 11, 074002. https://doi.org/10.1088/1748-9326/11/7/074002.