

Article

Quantifying Field Soil Moisture, Temperature, and Heat Flux Using an Informer–LSTM Deep Learning Model

Na Li ^{1,2} , Xiaoxiao Sun ^{1,2}, Peng Wang ^{1,2}, Wenke Wang ^{3,4,*} and Zhitong Ma ⁵

¹ School of Water Resources & Environment, China University of Geosciences, Beijing 100083, China; lina@cugb.edu.cn

² Key Laboratory of Groundwater Conservation of MWR, China University of Geosciences, Beijing 100083, China

³ Key Laboratory of Subsurface Hydrology and Ecological Effects in Arid Region, Ministry of Education, Chang'an University, Xi'an 710054, China

⁴ School of Environmental Science and Engineering, Chang'an University, Xi'an 710054, China

⁵ College of Geology and Environment, Xi'an University of Science and Technology, Xi'an 710054, China

* Correspondence: wenkew@chd.edu.cn

Abstract

Understanding water and heat transport through soils is vital for managing soil and groundwater resources, agricultural irrigation, and ecosystem protection. This paper aims to explore the potential application of deep learning methods in simulating water and heat transport processes within soils. It also examines the interactions between soil hydrological processes and environmental factors, including meteorological conditions and groundwater levels. To achieve these, we develop a hybrid model Informer–LSTM by combining two powerful architectures: Informer, a Transformer-based model essentially designed for long-sequence time-series forecasting, and Long Short-Term Memory (LSTM), a neural network that is great at learning short-term patterns in sequential data. The model is applied to field measurements from Henan Township in Ordos, Inner Mongolia, China, for training and testing, to simulate three key variables: soil water content, temperature, and heat flux at different depths in two soil columns with different groundwater levels. Our results confirm that Informer–LSTM is highly effective at simulating the soil water and heat transport. Simultaneously, we evaluate its performance by incorporating various combinations of input data including meteorological data, soil hydrothermal dynamics, and groundwater level. This reveals the relationship between soil hydrothermal processes and meteorological data, as well as coupled processes of soil water and heat transport. Moreover, employing SHapley Additive exPlanations (SHAP) analysis, we identify the most influential factors for predicting heat flux in shallow soils. This research demonstrates that deep learning models are a viable and valuable tool for simulating soil hydrothermal processes in arid and semi-arid regions.

Keywords: soil hydrothermal processes; soil water content; soil temperature; soil heat flux; deep learning; Informer–LSTM; SHAP (SHapley Additive exPlanations)



Academic Editors: Ilie Bodale and Vasile Stoleru

Received: 12 September 2025

Revised: 8 October 2025

Accepted: 14 October 2025

Published: 22 October 2025

Citation: Li, N.; Sun, X.; Wang, P.; Wang, W.; Ma, Z. Quantifying Field Soil Moisture, Temperature, and Heat Flux Using an Informer–LSTM Deep Learning Model. *Agronomy* **2025**, *15*, 2453. <https://doi.org/10.3390/agronomy15112453>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The vadose zone is the crucial pathway linking between atmospheric water, surface water, and groundwater, and it is characterized by complex hydrothermal transformation processes [1–3]. It governs numerous physical processes at the terrestrial surface, including the balance of mass and energy between the ground and the atmosphere and

fundamental biological processes such as plant growth [4]. In unsaturated soils near the land–atmosphere interface, the coupled movement of soil water and heat plays a critical role in the hydrological cycle. Consequently, accurate modeling of hydrothermal transport in the vadose zone is indispensable to improve our understanding of regional water balance, a subject that has received significant scholarly attention [5].

Soil water and heat transport dynamics are closely related to the atmosphere above and the groundwater below, thus they are influenced by local meteorological conditions such as solar radiation, air temperature, and the depth of the groundwater, which can significantly affect capillary rise and drainage. In traditional hydrological process models, these environmental factors are typically incorporated into boundary condition calculations and then combined with the water and heat governing equations, like the Richards equation and soil thermal transport equation, to simulate and predict soil water and heat processes [6–12]. However, in such methods, the rationality of establishing a complete coupling model, the accuracy of the model’s constitutive properties, particularly the hydrothermal characteristics function, and the correctness of the surface water and heat flux estimations will greatly affect the accuracy and predictive ability of the simulation results [13]. Therefore, modeling the soil hydrothermal dynamics is challenging because such processes are highly nonlinear and the boundary conditions are highly variable [14].

With the rapid development of artificial intelligence technologies in recent years, deep learning methods have been increasingly utilized to address the modeling of hydrology processes [15–20]. Researchers have explored the potential of deep learning methods for various vadose zone applications [18] including the characterization of soil hydraulic properties and surrogate models for control equations. We note that various methods based on long short-term memory (LSTM), convolutional neural networks (CNNs), Gated Recirculation Unit (GRU), and Transformer branches are used to estimate and predict the soil moisture [14,21–28] and soil temperature [29–35]. We also noticed that several studies have successfully applied deep learning methods to estimate soil heat flux under various environmental and spatial contexts. For example, Bonsoms and Boulet [36] estimated regional soil heat flux in South America by training neural networks and random forest models with field measurements of net radiation and surface temperature, as well as satellite data. Zheng and Jia [37] compared the soil heat flux estimated by empirical equations and machine learning models. In a separate study, Cross and Drewry [38] applied ensemble machine learning models to evaluate the effectiveness of meteorological and remote sensing datasets to estimate soil heat flux with different predictor variables.

Nevertheless, most deep learning applications for simulating soil water and heat processes focus on large-scale predictions of surface moisture, temperature, and heat flux. Therefore, there is a notable gap in research focused on simulating these processes at different depths within soil profiles [23,26,39]. To address this gap, this study proposes a hybrid deep learning model Informer–LSTM and applies it to model soil water and heat dynamics in field experiments. Our model is built on the Transformer architecture [40], which is well suited for processing long data sequences. However, traditional Transformers struggle with time-series forecasting due to high computational demands. To address this, we used the Informer [41], a Transformer variant specifically designed for long sequence forecasting. We then integrated a Long Short-Term Memory (LSTM) network [42], a recurrent neural network (RNN) [43] that excels at capturing short-term information, into the Informer. By combining these two architectures, our hybrid model leverages the strengths of both, allowing it to efficiently capture both long-range dependencies and short-term patterns.

We utilized field measurements of soil hydrothermal dynamics from two soil profiles with different groundwater levels in Ordos, Inner Mongolia, China, to train and validate our model. Our hybrid Informer–LSTM model was developed to simulate three key

hydrothermal variables: soil water content, soil temperature, and soil heat flux, which are all crucial variables in environmental and agricultural sciences. We investigated several input scenarios to understand their impact on model performance. To predict the soil water content and temperature, we used three types of input data: meteorological data (e.g., radiation and air temperature), groundwater levels, which are both crucial indicators of soil water and heat processes [2,44], and soil moisture and temperature data from the same depth (as we also explored the feasibility of using one variable to predict the other). Simultaneously, for each variable to be modeled, we explore how different combinations of input data impact the model's effectiveness. Additionally, we explored the use of deep learning to predict soil heat flux, which effectively couples energy transfer processes at the soil surface with those in the soil [45]. This model used soil moisture, soil temperature, and meteorological data as inputs. Furthermore, to analyze the model's predictions for soil heat flux, we used SHAP (SHapley Additive exPlanations) [46], a method designed to interpret deep learning models. This allowed us to illustrate how the Informer-LSTM model weighs the influence of different input data when making its predictions.

This paper is organized as follows: Section 2 first gives a brief overview of the climate and hydrological conditions in the study area, as well as the meteorological and field experimental data. Then, we briefly introduce the methods including the random forest method, which is used to analyze the importance of nine meteorological factors affecting soil water and heat processes, and the hybrid deep learning model Informer-LSTM; Section 3 presents the estimation results for soil water content, temperature, and heat flux; finally, Section 4 discusses and summarizes the simulation results.

2. Materials and Methods

2.1. Study Area and Field Data

2.1.1. Study Area

The study area is located in the Ordos Basin, specifically north of Baiyun Mountain in Inner Mongolia, China (Figure 1a). This region, situated in northwest China, has an elevation ranging from approximately 1000 to 1400 m. The area receives an average annual precipitation of approximately 340.3 mm, with maximum daily rainfall reaching up to 149.9 mm, while evaporation is intense, averaging about 2634.2 mm per year. Due to the combination of scarce precipitation, high evaporation, and limited surface water resources, groundwater serves as the critical, and sometimes the only, water source for vegetation survival and socioeconomic development.

The field experiments were presented in [47] in detail, thus only a summary is presented here. The in situ experiments are conducted in the National Meteorological Station in Ordos, Inner Mongolia (108°43' E, 37°51' N), at an elevation of approximately 1210 m. The station has a long-term average temperature of 8 °C, with a maximum of 36.7 °C and a minimum of −34.3 °C. Meteorological data was collected hourly at this station. We illustrated the precipitation, potential evaporation, and soil surface temperature based on the data, as shown in Figure 1b. In this field experiment, several cylindrical columns with different initial groundwater depths were installed. Each column has an inner diameter of 0.6 m and is sealed at the bottom to ensure full contact with the soil matrix, and they are installed parallel to the ground so that the top surface is directly exposed to the atmosphere. The columns are uniformly filled with Maowusu sandy soil, and a PVR observation tube is centrally installed in each column to monitor the depth of the groundwater table. Soil profile data (including soil water content, temperature, and groundwater level) is collected every 5 min. To demonstrate the effectiveness of deep learning models in simulating water and heat dynamics from shallow to deep soil layers, we selected soil columns with maximum depths of 1.2 and 3.0 m. Figure 1c shows the soil columns.

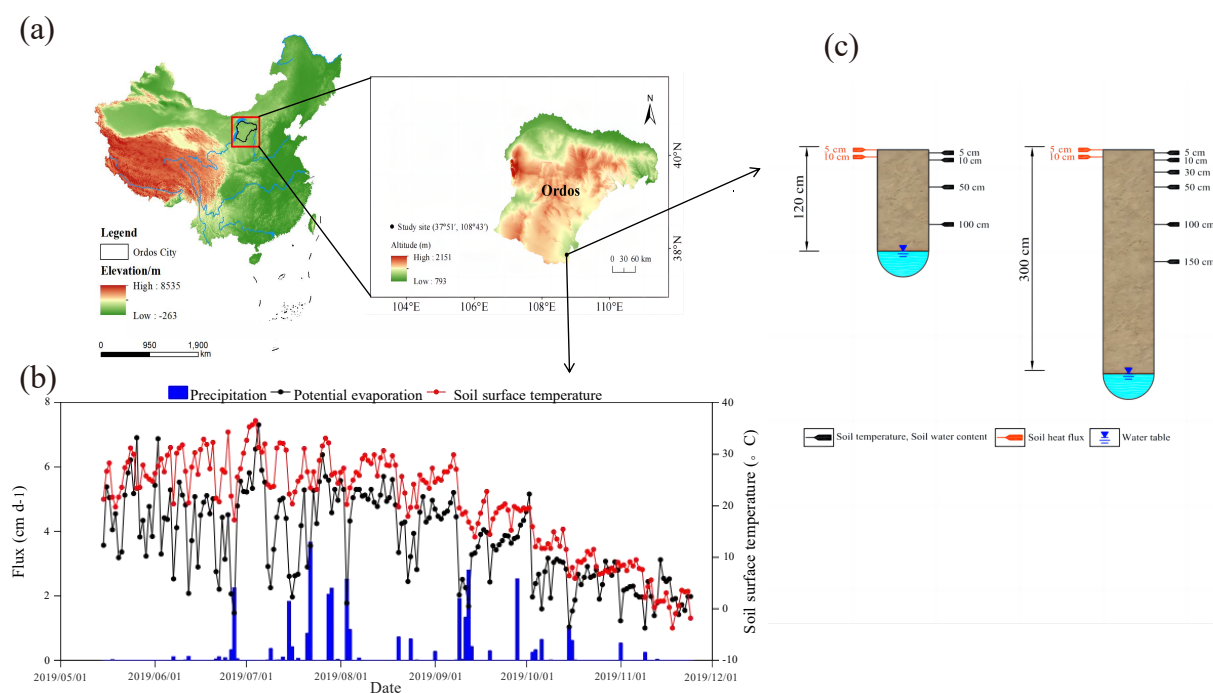


Figure 1. (a) Location of the study area. (b) Meteorological characteristics. (c) Soil profile diagram, with soil water content and temperature sensors installed at various depths: 5 cm, 10 cm, 50 cm, 100 cm for the 1.2 m soil column; 5 cm, 10 cm, 30 cm, 50 cm, 100 cm, and 150 cm for the 3.0 m soil column; and soil heat flux sensors placed at 5 and 10 cm for both columns.

2.1.2. Meteorological Data

This paper collected meteorological data over 167 days (from 11 May 2019, to 25 October 2019), including longwave radiation (LR) (W/m^2), shortwave radiation (SR) (W/m^2), average air pressure (AP) (hPa), average wind speed (WS) (m/s), ground temperature (GT) ($^{\circ}\text{C}$), air temperature (AT) ($^{\circ}\text{C}$), relative humidity (RH) (%), sunshine duration (SD) (hour), and precipitation (P) (mm). Table 1 shows the statistical information of these meteorological factors at a 1 h scale from 11 May 2019 to 25 October 2019. It respectively statistically analyzes the maximum (Max), minimum (Min), mean (Mean), standard deviation (SD), and coefficient of variation (CV). The data showed a high degree of dispersion for longwave radiation, with a CV of 1.41, the highest among all factors. This is further highlighted by the large difference between its maximum and minimum values, which was $1201.27 \text{ W}/\text{m}^2$. In contrast, average air pressure had the lowest CV at 0.04, indicating very little fluctuation in the data.

Table 1. Statistical characteristics of meteorological factors.

Meteorological Factor	Unit	Max	Min	Mean	SD	CV
LR (longwave radiation)	W/m^2	1185.43	−15.84	223.77	315.82	1.41
SR (shortwave radiation)	W/m^2	3.95	−165.74	−59.85	37.20	0.62
WS (average wind speed)	m/s	23.20	0.00	2.49	1.59	0.64
AP (average air pressure)	hPa	892.30	71.10	874.41	31.05	0.04
AT (air temperature)	$^{\circ}\text{C}$	34.80	−6.70	18.15	7.24	0.40
GT (ground temperature)	$^{\circ}\text{C}$	67.60	−5.30	23.67	13.90	0.59
RH (relative humidity)	%	99.00	6.00	62.93	26.50	0.42
SD (sunshine duration)	hour	1.00	0	0.36	0.45	0.23
P (precipitation)	mm	1.94	0	0.0078	0.0656	8.33

2.2. Method

2.2.1. Random Forest

Random forest (RF) is a powerful ensemble learning model that improves prediction accuracy and stability by combining multiple decision trees. Its core architecture is based on multiple cart trees, which are independent and diverse, thereby effectively preventing the overfitting issues commonly associated with traditional single decision trees [48–50]. The key to RF lies in the way it introduces “randomness”. Specifically, during the training phase, each decision tree is constructed independently using different sample sets and feature subsets. This process typically employs bootstrap sampling, where a certain proportion of data is randomly selected from the original dataset as input for a specific tree. Additionally, when splitting nodes, the random forest considers only a subset of candidate features rather than all features, further enhancing diversity. It randomly selects features for branching, making it run very quickly. This paper uses the Gini index for importance scoring:

$$\text{Gini index} = 1 - \sum_{i=1}^M (g_i)^2 \quad (1)$$

where M is the number of classes, and g_i is the probability of a data point belonging to class i . The Gini index is also a common way to measure feature importance in an RF. The importance of a feature is measured by how much it reduces the Gini index on average. The more the reduction, the more important the feature is considered. This is an efficient metric because the values are calculated automatically as a byproduct of the tree construction process. We will use the RF method to analyze the importance of nine meteorological factors affecting soil water and heat dynamics, thereby selecting more effective input data.

The random forest workflow includes three main steps: (1) preprocessing: missing and abnormal values were removed, and all meteorological and soil hydrothermal measurements were temporally aligned and standardized; (2) feature selection: feature importance was evaluated with the Gini index; (3) training, validation, and testing split: the dataset was divided into 70% for training, 15% for validation, and 15% for testing. The training set was used for parameter learning, the validation set was used for hyperparameter tuning and to avoid overfitting, and the testing set was used for independent evaluation of predictive performance.

2.2.2. Informer–LSTM

In this section, we introduce the coupling scheme for the two models: Informer and LSTM. The output of the Informer neural network is a hidden higher-dimensional feature, which is fed into an LSTM layer to generate predictions for soil moisture, temperature, and heat flux. The introduction of LSTM enhances the original Informer network’s ability to integrate sequential information. Figure 2 illustrates the conceptual diagram of the Informer–LSTM model. And, we present the process of running the hybrid model in pseudocode format in Table 2.

Informer is a supervised learning model based on attention mechanisms, primarily designed for long sequence time-sequence forecasting tasks [41]. It is an efficient Transformer-based model, introducing key innovative modifications to the self-attention mechanism to handle long time series data. Specifically, it incorporates a ProbSparse Self-Attention mechanism that dramatically reduces computational complexity by calculating attention scores for only the most key parts of a sequence. Additionally, Informer employs a distilling operation to compress long sequences into shorter, key representations, further reducing the computational load and enhancing prediction performance. Owing to its advantages in processing long sequences, the Informer model is particularly suitable for scenarios requiring long-term forecasting.

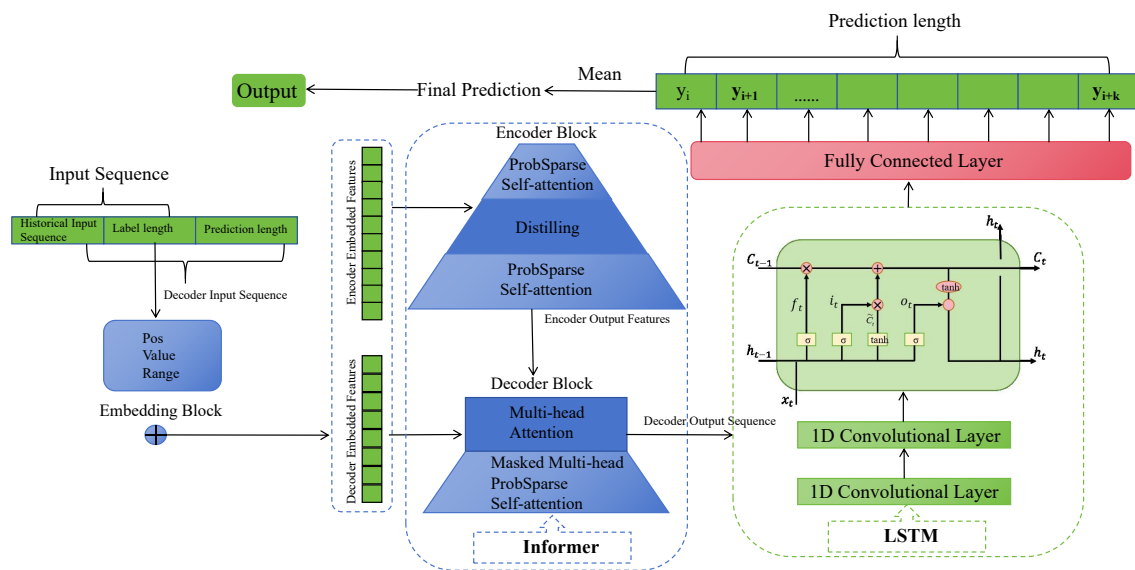


Figure 2. Conceptual diagram of the Informer-LSTM model.

Table 2. Algorithm: Informer-LSTM hybrid model for soil hydrothermal prediction.

Step 1: Input and Initialization Historical input sequence X_{his}, label sequence X_{label}, and prediction length L_p. Goal: Predict soil variables Y_{pred} (soil temperature, soil moisture, or soil heat flux).
Step 2: Embedding Stage Embed input and label sequences with positional and value encoding: $X_{enc} \leftarrow \text{Embedding}(X_{his}, \text{PositionalEncoding}, \text{ValueRange})$ $X_{dec} \leftarrow \text{Embedding}(X_{label}, \text{PositionalEncoding}, \text{ValueRange})$
Step 3: Informer Encoder For each encoder layer: • Apply ProbSparse Self-Attention: $X_{enc} \leftarrow \text{ProbSparse_SelfAttention}(X_{enc})$ • Apply Distilling to shorten long sequences: $X_{enc} \leftarrow \text{Distilling}(X_{enc})$.
Step 4: Informer Decoder For each decoder layer: • Apply Masked ProbSparse Self-Attention (MPSA): $X_{dec} \leftarrow \text{MPSA}(X_{dec})$ • Apply Multi-Head Attention (MHA) using encoder output: $X_{dec} \leftarrow \text{MHA}(X_{dec}, X_{enc})$ Output encoded-decoded sequence: $H_{informer} \leftarrow \text{Decoder_Output}(X_{dec})$
Step 5: LSTM Integration Feed Informer output into LSTM pipeline: $H_{conv} \leftarrow \text{Conv1D}(H_{informer})$ $H_{seq} \leftarrow \text{LSTM}(H_{conv})$ $Y_{seq} \leftarrow \text{FullyConnected}(H_{seq})$
Step 6: Post-processing and Output Aggregate prediction results along feature dimension: $Y_{pred} \leftarrow \text{Mean}(Y_{seq}, \text{features})$ Return predicted soil variables Y_{pred}

The structure of Informer is shown in the middle of Figure 2. The Encoder primarily receives large-scale long series inputs. Here, the ProbSparse Self-Attention replaces the conventional self-attention mechanism, and the trapezoidal component represents the extraction process of self-attention, which significantly reduces the size of the network. The multi-layer stacking further enhances the robustness of the model. Meanwhile, the Decoder receives long series inputs with the target elements padded with zero. This approach

computes the attention-weighted components of the feature map, and the model then outputs these elements in a fast, generated format.

LSTM [42] is an optimized version of Recurrent Neural Networks (RNNs). Although conventional RNNs can handle time-series data, their performance tends to deteriorate when processing long sequences with long-term dependencies due to the “vanishing gradient” problem. To overcome this limitation, LSTM was developed by incorporating additional components such as a forget gate, an input gate, and an output gate, which enable it to handle long sequences effectively. LSTMs use a memory cell to control what information is kept or discarded. This is managed by three gates: the forget gate (f_t), which decides what to discard from the previous state; the input gate (i_t), which determines what new information to add; and the output gate (o_t), which controls the information passed to the next state. The overall architecture is illustrated in the right part of Figure 2.

In fact, we initially attempted to model soil hydrothermal dynamics using Informer and LSTM models separately. Our results showed that the LSTM model’s performance was not satisfactory, while the Informer model yielded better results. Specifically, the LSTM exhibited poor generalization capabilities when simulating soil water content and temperature, whereas the Informer yielded significantly better results, though its generalization performance for soil temperature simulation remained suboptimal. However, combining both models significantly improved the simulation’s effectiveness. Therefore, we chose to use a hybrid Informer–LSTM model for this study.

2.2.3. Architecture of the Deep Learning

This study uses the Windows 11 operating system, with Python 3.10 and PyCharm 2023.3.2 as the development platform. The loss function is obtained by calculating the MSE between the model’s simulated values and the observed values. The optimizer is Adam, and the activation function is GELU. Cross-validation is applied to prevent overfitting, and five sets of numerical simulation experiments are conducted to take the average to obtain more accurate results. The dataset is split, with the first 80% as the training set and the remaining 20% as the test set. Specifically, data from 11 May to 23 September, a total of 135 days, is used for training, and data from 24 September to 25 October, a total of 32 days, is used for testing. Table 3 shows the hyperparameters of the Informer–LSTM model were primarily determined through a combination of Grid Search and empirical tuning. The specific process is detailed below:

Step 1. Initial Screening and Empirical Setting

Based on previous research and the model’s convergence characteristics, initial candidate values were first set within a reasonable range. For example, the learning rate was set to 110^{-3} , the dropout rate to 0.05, and the batch size to 64.

Step 2. Validation Set Optimization

The best hyperparameter combination was selected using the minimization of the Validation Loss criterion. The final determined hyperparameters are listed in Table 3.

Step 3. Tuning and Training Strategy

Three strategies including early stopping, learning rate scheduling, and cross-validation were adopted during the training process to improve model stability and generalization performance.

Step 4. Final Verification

The model was verified for finality in an independent test set (which represents 20% of the total data). Evaluation metrics included RMSE, MAE, and R^2 . To mitigate the influence of random factors, each set of experiments was run five times and the average result was taken.

Table 3. Hyperparameters for the Informer–LSTM model.

Hyperparameter	Value
Sequence length	3
Label length	3
Dropout rate	0.05
Training epochs	20
Patience	3
Learning rate	0.00064
Batch size	64
Number of attention heads	Increase by one with the number of feature inputs
Attention factor	5
Model dimension	512

2.2.4. Deep Learning Interpretability: SHapley Additive exPlanations (SHAP)

SHAP is a method for interpreting the results of deep learning models. It uses Shapley values, a concept from game theory, to assign an “importance” value to each feature for a specific prediction. This helps to make “black-box” models more transparent by showing exactly how each feature contributed to a specific outcome. The core concept of SHAP originates from the Shapley value in game theory. In cooperative games, the Shapley value is used to calculate the contribution of each participant to the final outcome. SHAP applies this concept to machine learning by treating the model as a game, where features are the players and the model output is the final result. The Shapley value assigns a contribution score to each feature, indicating its influence on the model’s output.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (2)$$

where S is a subset of features, N is the set of all features, and $f(S \cup \{i\})$ represents the model prediction after adding feature i to subset S ; $f(S)$ represents the prediction of the subset S .

The SHAP value has linearity, meaning that the model output can be decomposed into the SHAP values for each feature, as follows:

$$f(x) = \phi_0 + \sum_{i=1}^N \phi_i \quad (3)$$

where ϕ_0 is the baseline value (the model’s output when there are no features), and ϕ_i is the contribution value of feature i .

2.3. Metrics

In order to evaluate the performance of the deep learning model, four metrics are used to quantify the difference between prediction results and observations. These include the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the coefficient of determination (R^2). These metrics evaluate the model’s performance from different perspectives. MAE measures the average absolute difference between predicted and observed values, with a lower MAE meaning higher accuracy. The RMSE is more sensitive to large errors, so a lower RMSE indicates a more robust model. MAPE gives a percentage-based error, which is useful for comparing performance across different datasets. R^2 shows the model’s ability to explain the variability

in observed data, with a value closer to 1 indicating a better fit. These evaluation metrics are as follows:

$$\text{MAE} = \frac{\sum_{i=1}^n |x_i - y_i|}{n} \quad (4)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (x_i - y_i)^2}{n}} \quad (5)$$

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{x_i - y_i}{x_i} \right| \times 100 \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (x_i - y_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (7)$$

where n is the total number of samples, x_i is the i th observed value, $\bar{x}$ is the mean of the observed values, y_i is the i th simulated value, and $\bar{y}$ is the mean of the simulated values.

3. Results and Discussions

In this section, we first present the results of our model's training and simulation of soil water content and temperature. Based on these simulated results, we then predict the soil heat flux and compare it with the measured values.

3.1. Impact of Meteorological Factors on Soil Temperature and Moisture

This paper discusses the potential of meteorological factors in simulating and predicting soil water and heat processes. We first present an analysis of the importance of the nine meteorological factors using the RF method and exclude factors with extremely low importance. The RF model is implemented using the sklearn library, with the number of decision trees being 100 and the random state seed being 42, ensuring that the randomness of the model is consistent each time the code is run. The feature importance scores are calculated using the Gini index (Equation (1)).

Here, we use the results of feature importance score for the soil column of 3.0 m to demonstrate the importance of various meteorological factors for soil moisture content and soil temperature. Figure 3 shows the importance analysis of meteorological factors on soil temperature. In shallow soil including 5 cm and 10 cm, air temperature (AT) is the most dominant factor, with its importance score significantly higher than all other meteorological elements. As depth increases to the middle soil layer (20 cm, 50 cm), AT remains the primary influence. However, the importance of other factors, such as RH and AP, gradually increases, indicating their growing influence on soil temperature at greater depths. This partially aligns with the conclusions drawn in [29,35]. Alizamir et al. [29] found AT is the most highly correlated factor with soil temperature, especially for the shallow depths (5 cm, 10 cm, and 50 cm), while Asadzadeh et al. [35] established that at shallow depths (5–20 cm), surface infrared temperature is the primary influencing factor, whereas at deeper levels (50–100 cm), air temperature becomes the dominant factor.

Similarly, Figure 4 shows the importance analysis of meteorological factors on soil water content. As documented in [14], the relationship between meteorological data and soil water content is highly variable, site-specific, and non-linear. As can be seen, meteorological elements influence soil water content differently than they do soil temperature. For soil water content and temperature, the consistent phenomenon is that across all depths, sunshine duration and precipitation consistently show the lowest importance. It has also been reported that precipitation has a negligible effect on daily soil temperature across different soil depths [35]. Based on this finding, these two factors will be excluded from

subsequent model simulations. That is, when meteorological data is used as input for the Informer–LSTM model, the first seven elements listed in Table 1 will be selected.

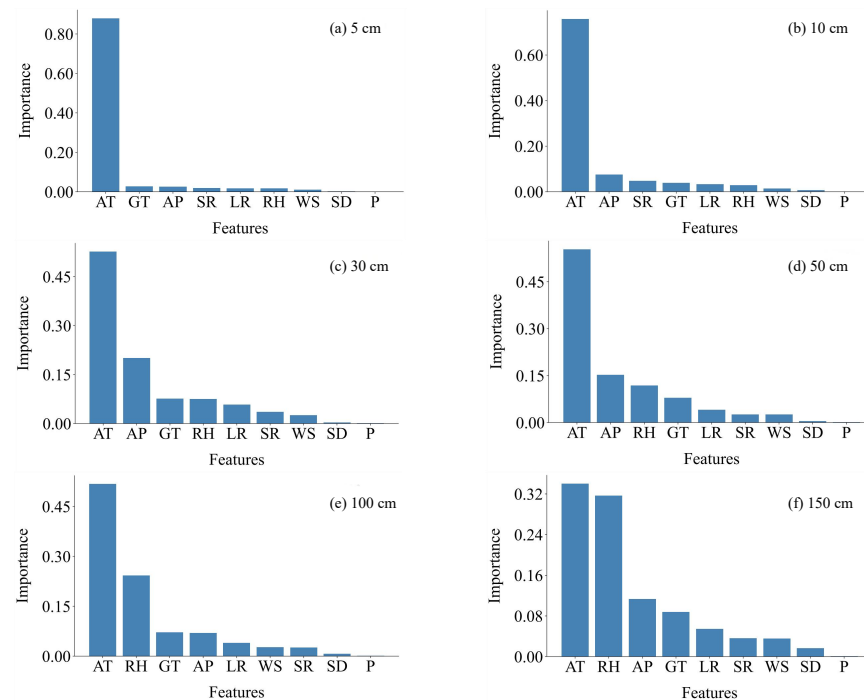


Figure 3. Score plot of multivariate meteorological elements influencing soil temperature at different depth in the soil profile.

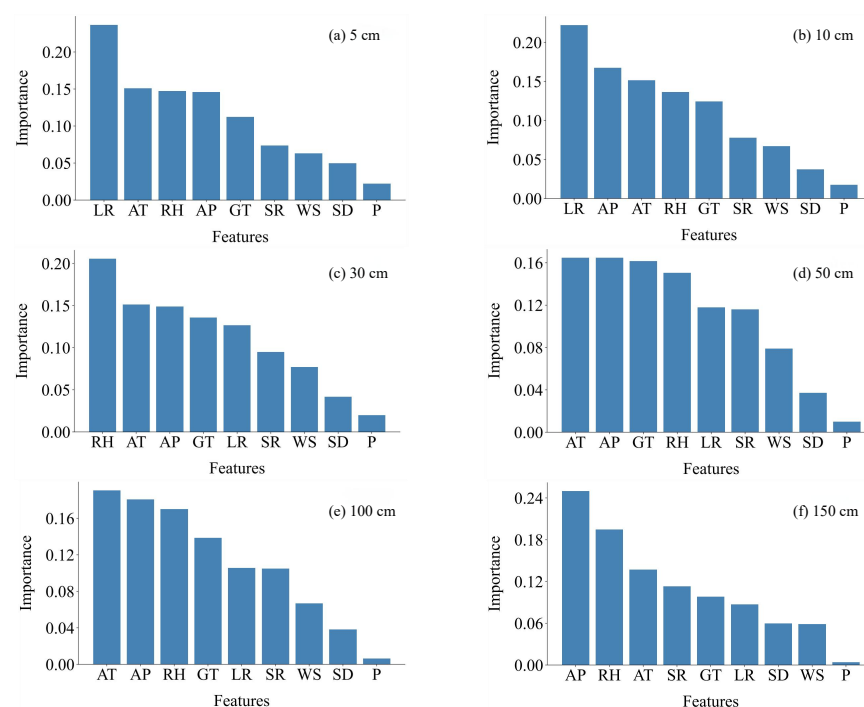


Figure 4. Score plot of multivariate meteorological elements influencing soil water content at different depth in the soil profile.

3.2. Simulation of Water Content at Different Soil Depths Using Informer–LSTM

In this section, soil water content at various depths of the soil columns is analyzed and predicted by using the Informer–LSTM model with different combinations of input features. The measurements from soil columns with depths of 1.2 and 3.0 m are used for

model training and testing. Our training period is from 11 May 2019 to 23 September 2019, while the testing period is from 24 September 2019 to 25 October 2019. For clarity, we have only provided the results from the test period here. Figure 5 (for the 1.2 m soil column) and Figure 6 (for the 3.0 m soil column) show the comparison between the predicted and observed soil water content at various depths using three types of input including meteorological data (including the first seven meteorological factors listed in Table 1), soil temperature (at the same depth as the predicted water content), and groundwater level. Overall, the Informer–LSTM model successfully captures the general trends of soil moisture throughout the test period at all depths, regardless of the input features used. Nevertheless, we observe that soil moisture content near the surface exhibits greater sensitivity to meteorological data, whilst that closer to the subsoil layer responds more readily to groundwater levels. This is particularly evident in using only meteorological data as input: the model performed well at shallow depths but struggled to accurately predict moisture content at deeper levels (100 and 150 cm) in the 3.0 m soil column. This is likely because shallow soil is more directly influenced by surface conditions, while deeper layers are less affected by these surface variations. The figure also indicates that simulation results using soil temperature as input generally perform slightly less favorably. This may stem from the time lag in soil temperature’s response to moisture content. Furthermore, predicting soil water content using soil temperature from the same layer may have resulted in the loss of its correlation with the adjacent layer.

Actually, in addition to using meteorological data, soil temperature, and groundwater level as separate inputs, we also simulated and predicted moisture content using any two of these three types of data in combination, as well as all three types of data. That is, there are a total of seven input scenarios. However, for the sake of clarity and readability of the graphical display, we only show the simulation results when three types of data are used as inputs separately. Surprisingly, introducing multiple input features did not improve the performance of the deep learning model and even worsened the simulation results.

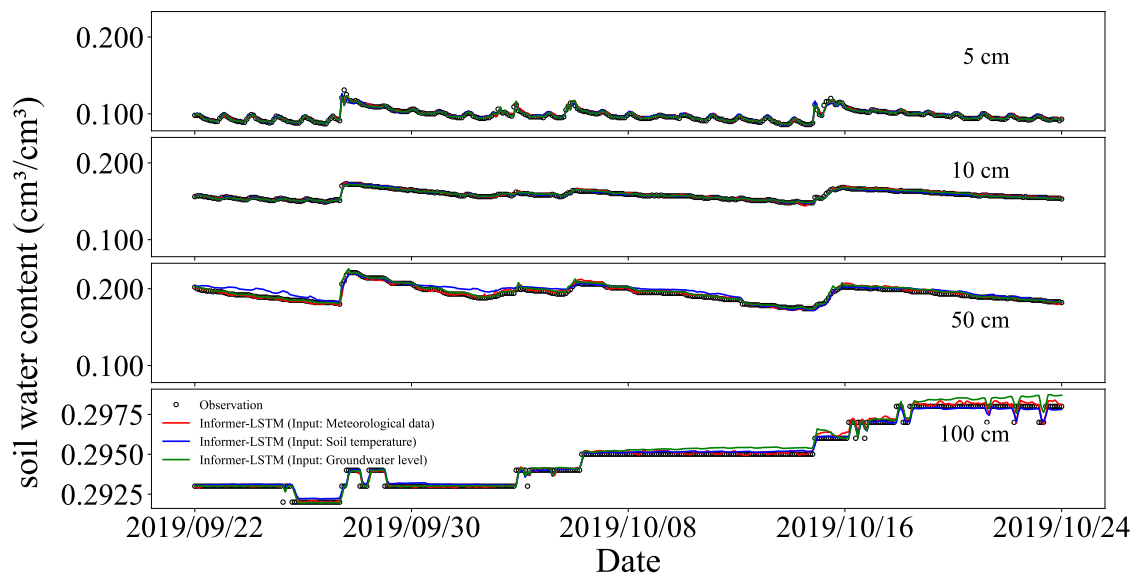


Figure 5. Comparative analysis of simulated and measured soil water content in the 1.2 m soil column under different input features.

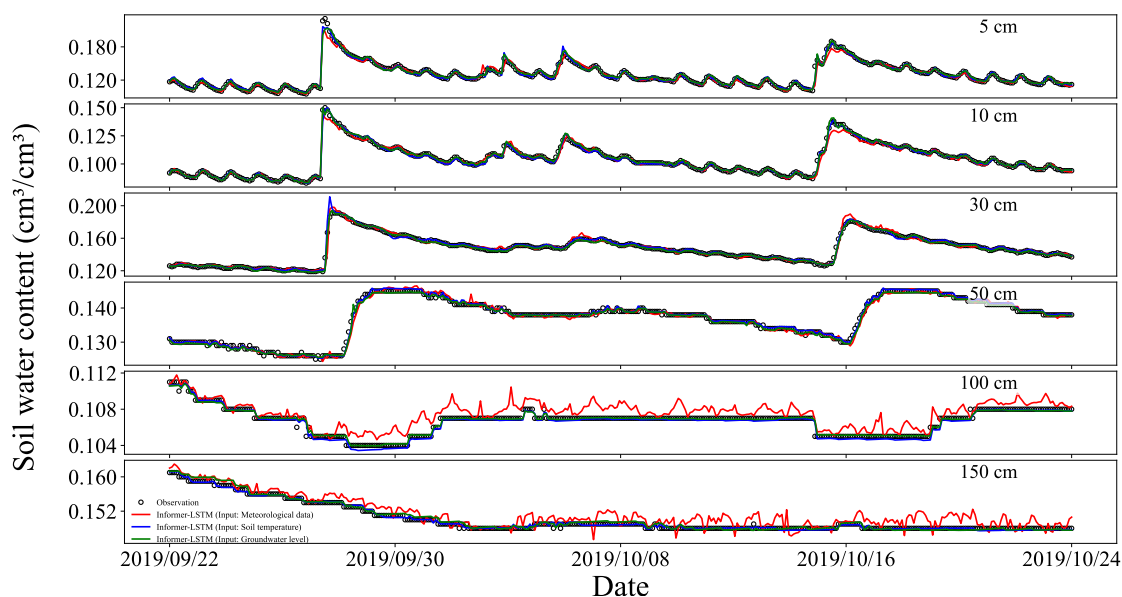


Figure 6. Comparative analysis of simulated and measured soil water content in the 3.0 m soil column under different input features.

We quantitatively evaluated the Informer–LSTM model’s performance for predicting soil water content using seven different input feature combinations. Table 4 summarizes R^2 , RMSE, MAPE, and MAE for soil water content predictions in the 3.0 m soil column. The statistical indicators for the 1.2 m soil column show a similar pattern to those of the 3.0 m. Therefore, we only present the results for the 3.0 m column in this paper. Overall, models that use a single type of data (e.g., soil temperature, meteorological data, or groundwater levels) as input achieved higher simulation accuracy. In both shallow and deep layers, the best performance was achieved using soil temperature and groundwater level factors, followed by meteorological data. At shallow (5 and 10 cm) and middle (30 and 50 cm) depths, the differences in performance corresponding to the three types of input were small. However, the differences significantly increase at deeper depths, that is, the model fit was significantly worse when using only meteorological data as input. And, the performance of Informer–LSTM using meteorological data as input decreased, while using groundwater levels and water content as inputs improved with increasing depth. In any case, we can observe that soil water content exhibits strong correlations with meteorological factors, soil temperature, and groundwater levels. This aligns with the findings of [22], which concluded that the fluctuation of SWC is attributed to changes in environmental factors such as soil composition, meteorological data, and groundwater, resulting in a complex non-linear relationship that the deep learning model is designed to fit. Moreover, using multiple data inputs does not significantly improve model performance; in fact, it led to markedly poorer simulation results at deeper soil layers, such as at the 100 cm and 150 cm depths within the 3.0 m soil column. This is likely due to the introduction of more data errors, particularly in time series data such as soil water and thermal dynamics, which exhibit pronounced random noise.

Table 4. The simulation accuracy of Informer–LSTM for soil water content in the 3.0 m soil column under different input features. MFC: meteorological factor combination (including the first seven meteorological elements listed in Table 1), ST: soil temperature, GWL: groundwater level.

Soil Depth (cm)	Inputs	Evaluation Indicators			
		R^2 (—)	RMSE (cm ³ /cm ³)	MAPE (%)	MAE (cm ³ /cm ³)
5	MFC	0.8513	0.0086	0.0224	0.0033
	ST	0.8759	0.0078	0.0190	0.0026
	GWL	0.8732	0.0079	0.0183	0.0026
	ST, GWL	0.8553	0.0084	0.0259	0.0036
	MFC, ST	0.8381	0.0089	0.0251	0.0037
	MFC, GWL	0.8474	0.0086	0.0233	0.0034
	MFC, GWL, ST	0.8452	0.0087	0.0251	0.0036
10	MFC	0.9097	0.0039	0.0132	0.0015
	ST	0.9252	0.0036	0.0111	0.0012
	GWL	0.9242	0.0036	0.0115	0.0013
	ST, GWL	0.9052	0.0040	0.0162	0.0018
	MFC, ST	0.8914	0.0043	0.0174	0.0020
	MFC, GWL	0.9089	0.0039	0.0140	0.0016
	MFC, GWL, ST	0.9020	0.0041	0.0157	0.0018
30	MFC	0.9574	0.0034	0.0112	0.0017
	ST	0.9628	0.0032	0.0089	0.0014
	GWL	0.9649	0.0031	0.0076	0.0012
	ST, GWL	0.8954	0.0053	0.0235	0.0036
	MFC, ST	0.9427	0.0039	0.0156	0.0024
	MFC, GWL	0.9511	0.0036	0.0122	0.0019
	MFC, GWL, ST	0.9322	0.0043	0.0169	0.0025
50	MFC	0.9636	0.0015	0.0086	0.0012
	ST	0.9826	0.0008	0.0038	0.0005
	GWL	0.9833	0.0007	0.0037	0.0005
	ST, GWL	0.8869	0.0019	0.0118	0.0016
	MFC, ST	0.8874	0.0026	0.0148	0.0021
	MFC, GWL	0.9524	0.0013	0.0072	0.0010
	MFC, GWL, ST	0.9256	0.0017	0.0100	0.0014
100	MFC	0.8072	0.0006	0.0045	0.0005
	ST	0.9434	0.0003	0.0021	0.0002
	GWL	0.9465	0.0003	0.0021	0.0002
	ST, GWL	−2.9085	0.0028	0.0239	0.0026
	MFC, ST	−1.6135	0.0023	0.0185	0.0020
	MFC, GWL	−0.9983	0.0020	0.0171	0.0018
	MFC, GWL, ST	−0.9381	0.0020	0.0162	0.0017
150	MFC	0.8658	0.0013	0.0065	0.0010
	ST	0.9645	0.0006	0.0038	0.0006
	GWL	0.9785	0.0005	0.0025	0.0004
	ST, GWL	−4.6870	0.0082	0.0534	0.0080
	MFC, ST	−0.8933	0.0047	0.0285	0.0043
	MFC, GWL	0.7561	0.0017	0.0098	0.0015
	MFC, GWL, ST	−0.8505	0.0047	0.0292	0.0044

3.3. Simulation of Soil Temperature at Different Depths Using Informer–LSTM

This section analyzes the Informer–LSTM’s performance in simulating and predicting soil temperature of the soil profile. We examine how different combinations of input features affect the model’s accuracy. Figures 7 and 8 demonstrates the model’s performance by comparing predicted and observed soil temperatures at various soil depths. Similarly, seven different input scenarios were tested, including three scenarios using each data type individually, three scenarios using combinations of two data types, and one scenario using all three data types. For simplicity and clarity, Figures 7 and 8 only display the simulation results for the three scenarios where each data type was used as the sole input.

The figures show simulation results within the testing period only. Overall, the model—with various combinations of input features—accurately captures the trend of soil temperature variations before mid-October, but significant deviations occur after mid-October. Specifically, for both 1.2 m and 3.0 m soil columns, within 50 cm with soil temperature exhibiting significant fluctuations, the model failed to predict lower soil temperature values after 15th October. While at deeper soil, such as 100 cm in both columns and 150 cm in the 3.0 m column, the simulated values were consistently higher than the observations after 3 October. Additionally, the model appears to perform less well in predicting shallow soil temperatures than in predicting deeper temperatures. This could be due to the greater variability in soil temperatures in shallow soil posing challenges for prediction, whereas soil temperatures gradually stabilize in deeper layers.

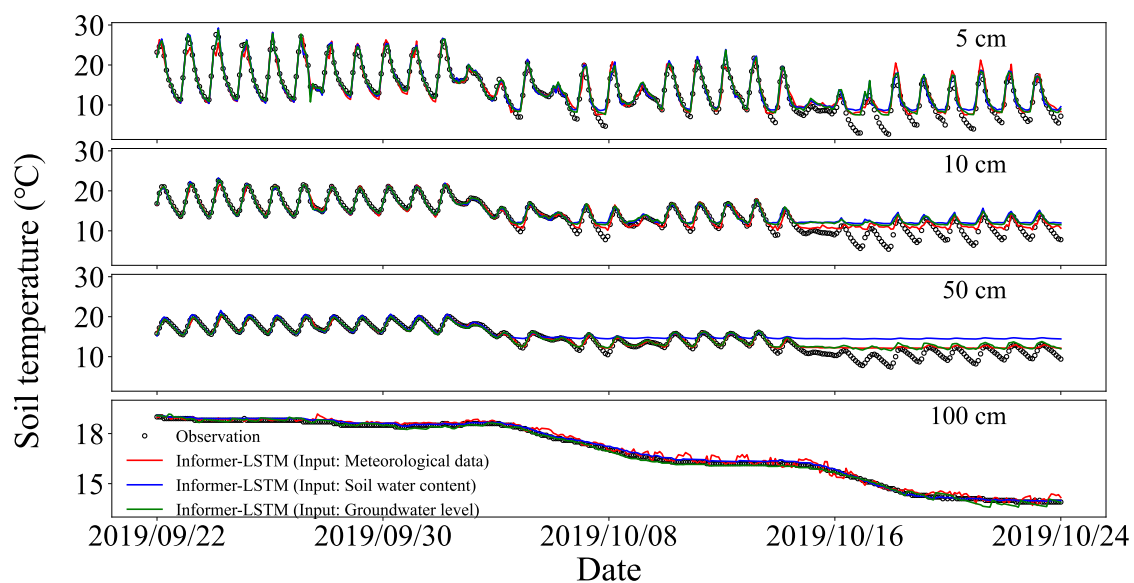


Figure 7. Comparative analysis of simulated and measured soil temperature in the 1.2 m soil column under different input features.

To present a more comprehensive set of simulation results for soil temperature by Informer–LSTM, we have listed the model simulation metrics corresponding to different input conditions in Table 5, which summarizes the R^2 , RMSE, MAPE, and MAE values when using seven different input feature combinations at each depth. Here, due to space constraints, we present only the simulation accuracy for soil temperature in the 3.0 m soil column. Overall, as depth increases, the model’s simulation results gradually improve, regardless of the input conditions. Similarly to the water content simulations, incorporating multiple types of input data does not significantly enhance the simulation results and even reduces accuracy at times, particularly in deeper soil layers (100 cm and 150 cm depths). Additionally, we can observe that meteorological data as input is advantageous

for predicting upper soil temperatures, offering greater benefits than soil water content and groundwater level data.

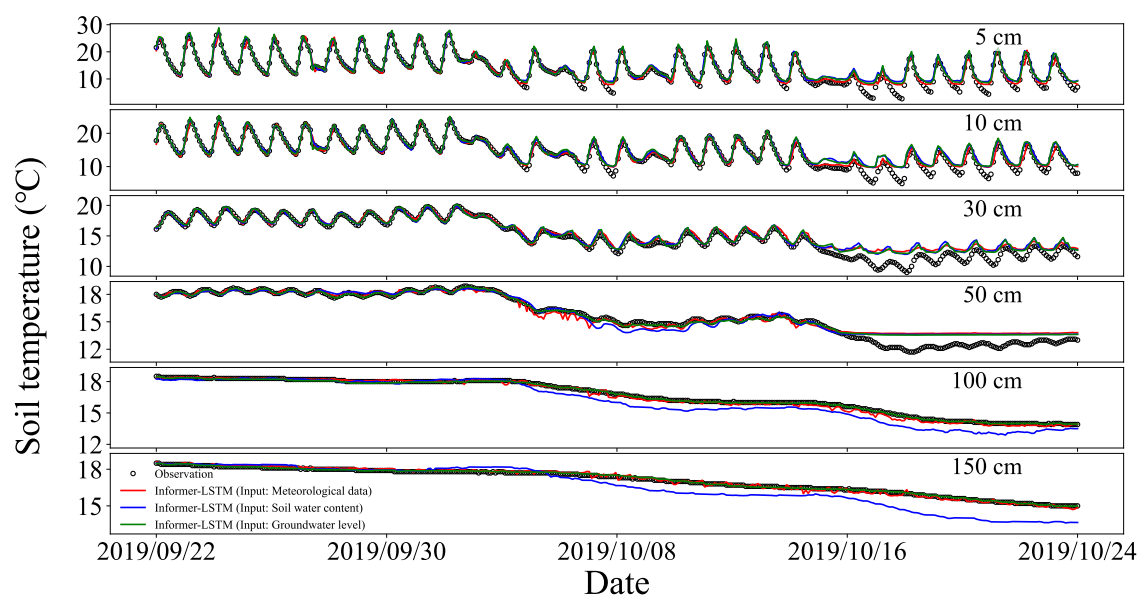


Figure 8. Comparative analysis of simulated and measured soil temperature in the 3.0 m soil column under different input features.

Table 5. Simulation accuracy of the Informer–LSTM model for soil temperature prediction of the 3.0 m soil column under different input features. MFC: meteorological factor combination, SWC: soil water content, GWL: groundwater level.

Soil Depth (cm)	Input	Evaluation Indicators			
		R^2 (—)	RMSE (°C)	MAPE (%)	MAE (°C)
5	MFC	0.9030	1.6531	12.41	1.1508
	SWC	0.8252	2.2191	16.80	1.5711
	GWL	0.8112	2.3060	17.35	1.7013
	SWC, GWL	0.8134	2.2930	16.13	1.5967
	MFC, SWC	0.9109	1.5836	11.66	1.1197
	MFC, GWL	0.8932	1.7342	13.05	1.2335
	MFC, GWL, SM	0.9114	1.5797	11.64	1.1622
10	MFC	0.9197	1.2072	7.830	0.8173
	SWC	0.8316	1.7482	11.65	1.269
	GWL	0.838	1.7145	10.98	1.2149
	SWC, GWL	0.8075	1.8696	12.75	1.3680
	MFC, SWC	0.9287	1.1373	7.340	0.7712
	MFC, GWL	0.9049	1.3136	8.440	0.8451
	MFC, GWL, SM	0.9253	1.1645	7.700	0.8162
30	MFC	0.8841	0.9682	5.050	0.5986
	SWC	0.8819	0.9772	5.320	0.6456
	GWL	0.88	0.9849	5.460	0.6697
	SWC, GWL	0.8230	1.1965	6.790	0.8438
	MFC, SWC	0.8806	0.9825	5.240	0.6212
	MFC, GWL	0.9056	0.8737	4.400	0.5157
	MFC, GWL, SM	0.8386	1.1422	6.330	0.758

Table 5. *Cont.*

Soil Depth (cm)	Input	Evaluation Indicators			
		R^2 (—)	RMSE (°C)	MAPE (%)	MAE (°C)
50	MFC	0.9003	0.7291	3.540	0.4712
	SWC	0.9542	0.4943	2.500	0.3479
	GWL	0.9338	0.594	2.750	0.3628
	SWC, GWL	0.9164	0.6680	3.520	0.4817
	MFC, SWC	0.912	0.6851	3.150	0.4149
	MFC, GWL	0.8971	0.7408	3.650	0.486
	MFC, GWL, SM	0.9166	0.6669	3.300	0.4421
100	MFC	0.9885	0.1672	0.800	0.1278
	SWC	0.9952	0.1276	0.630	0.0984
	GWL	0.9984	0.0618	0.300	0.05
	SWC, GWL	0.7517	0.7754	4.190	0.6840
	MFC, SWC	0.9951	0.1078	0.500	0.0810
	MFC, GWL	0.8441	0.6142	2.800	0.4284
	MFC, GWL, SM	0.9687	0.275	1.380	0.2141
150	MFC	0.9819	0.1364	0.670	0.1133
	SWC	0.9912	0.1451	0.770	0.1245
	GWL	0.9961	0.0629	0.290	0.0493
	SWC, GWL	−0.8793	1.3916	7.360	1.2255
	MFC, SWC	0.7924	0.4625	2.500	0.4256
	MFC, GWL	0.9461	0.2356	1.160	0.1973
	MFC, GWL, SM	0.9533	0.2191	1.080	0.1811

3.4. Simulation of Soil Heat Flux at Different Depths Using Informer–LSTM

In the previous section, we demonstrated the performance of the hybrid deep learning model Informer–LSTM in simulating soil water content and temperature. This section explores its potential for simulating soil heat flux using different inputs. Soil heat flux is directly related to soil temperature and water content because these factors determine how efficiently heat is stored and transferred through the soil. Therefore, we have selected soil water content and temperature as inputs for our model, along with standard meteorological factors. We will compare three scenarios to predict soil heat flux at shallow depths (5 cm and 10 cm): meteorological factors as the sole input, moisture content and temperature as combined inputs, all three factors—meteorological data, moisture content, and temperature—as combined inputs.

Figures 9 and 10 compared the Informer–LSTM’s predicted soil heat flux values against observed data at different soil depths with various input configurations. As shown in the figures, for both soil columns, regardless of the input conditions, the model performed exceptionally well throughout the entire testing period, fully capturing the dynamic changes in the surface soil heat flux. And using multiple types of input data did not appear to improve the model’s results.

A quantitative analysis of the performance of Informer–LSTM for soil heat flux predictions is presented in Table 6. The results are consistent with Figures 9 and 10. The model performed best when using only meteorological data, with the highest R^2 and the lowest RMSE, MAPE, and MAE. When using only soil moisture and temperature as inputs, the model performed worst, but only slightly worse. The simulation results with all the three data types as inputs fall between the two extremes. This results will be further interpreted by the SHAP analysis in the next subsection.

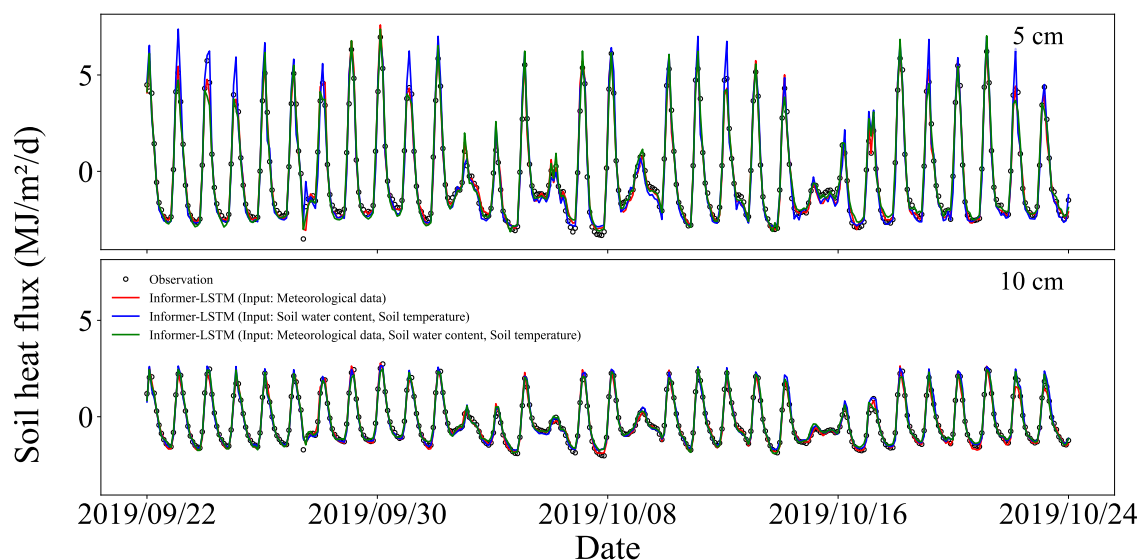


Figure 9. Comparisons between the simulated soil heat flux and observed data in the 1.2 m soil column under different input features.

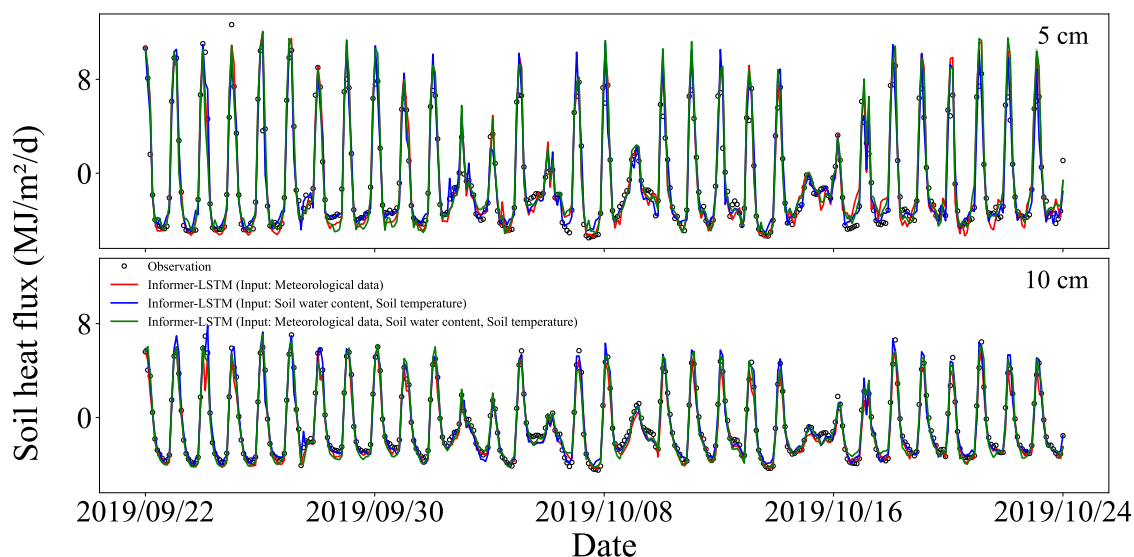


Figure 10. Comparisons between the simulated soil heat flux and observed data in the 3.0 m soil column under different input features.

Table 6. The simulation accuracy of Informer-LSTM for prediction of soil heat flux at top soils (5 and 10 cm) in 1.2 m and 3.0 m soil columns.

Soil Column	Soil Depth (cm)	Inputs	Evaluation Indicators			
			R^2 (—)	RMSE (MJ/m ² /d)	MAPE (%)	MAE (MJ/m ² /d)
1.2 m	5	MFC	0.9160	8.4537	2.7621	5.3601
		SM, ST	0.8892	9.7082	6.4036	6.4940
		MFC, SM, ST	0.9152	8.4940	3.0813	5.4764
	10	MFC	0.9537	2.9748	1.0605	2.0450
		SM, ST	0.9180	3.9578	1.6930	2.5730
		MFC, SM, ST	0.9237	8.0563	2.2297	5.2154

Table 6. Cont.

Soil Column	Soil Depth (cm)	Inputs	Evaluation Indicators			
			R^2 (—)	RMSE (MJ/m ² /d)	MAPE (%)	MAE (MJ/m ² /d)
1.2 m	5	MFC	0.7985	22.2681	0.7862	15.2318
		SM, ST	0.7244	26.0425	0.8984	18.8354
		MFC, SM, ST	0.7835	23.0791	0.8236	15.6490
	10	MFC	0.8976	10.8302	0.5522	7.1323
		SM, ST	0.8649	12.4403	0.6949	8.1077
		MFC, SM, ST	0.8823	11.6096	0.6358	7.8982

3.5. Model Explainability

To enhance the interpretability of our deep learning model for simulating soil hydrothermal transport, we applied the SHAP (SHapley Additive exPlanations) method. Using soil heat flux simulation as a case study, we demonstrate how this method can explain the relationship between a model's inputs and its outputs. By analyzing the Informer-LSTM model, which uses a combination of soil temperature, soil moisture, and meteorological data, SHAP allows us to precisely quantify the influence of each input variable on the model's outputs. This provides a deep understanding of the model's decision-making process.

Figure 11 reveals the most influential features for soil heat flux in the 1.2 m soil column. Because the results from the 3.0 m column are similar, we have only included the 1.2 m data for analysis. The figure indicates that the long-wave radiation, short-wave radiation, and relative humidity consistently have a significant impact on the model's predictions, identifying them as key features. In contrast, soil-related variables (like soil moisture and temperature) and other meteorological factors (such as atmospheric pressure and relative humidity) generally have a less pronounced effect on the output.

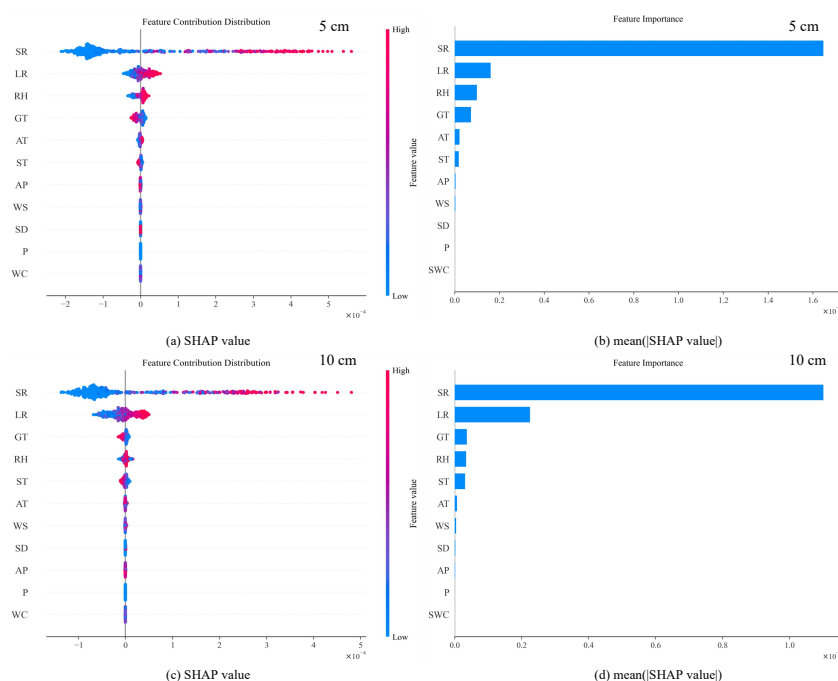


Figure 11. Distribution of SHAP values for soil heat flux at different soil depths under varying input conditions.

4. Conclusions

This study proposes a hybrid deep learning model Informer–LSTM for simulating soil water and heat transport processes. Soil water content, soil temperature, and soil heat flux at various depths of soil profiles with different groundwater levels are simulated and predicted. We evaluate the performance of the model under various input combinations from multi-source data, including meteorological factors, groundwater level, and soil hydrothermal dynamics. We also used SHAP analysis to identify and quantify the influence of each input feature on the model’s predictions, providing insight into the model’s decision-making process. Our Informer–LSTM model accurately simulated key soil hydrothermal variables—including soil water content, temperature, and heat flux—within the soil profiles.

We found that using meteorological data or groundwater levels resulted in high-precision predictions for soil water and temperature. Specifically, meteorological variables were better for predicting conditions in the upper soil layers, which are directly influenced by atmospheric conditions. Conversely, groundwater levels were more effective for predictions in deeper layers, as they are a primary driver of water and heat dynamics far below the surface. Additionally, we observed a strong two-way predictive relationship between soil water content and temperature. The accuracy of these predictions was comparable to those using groundwater levels, with performance improving with increasing depth. This highlights the close link between soil water and thermal processes.

For both soil water content and temperature modeling, using a single input data type typically led to better results than combining multiple factors. Introducing a variety of data, especially for predictions in deeper soil layers, did not improve accuracy and often degraded performance. This is likely because the model overfit to noise or irrelevant data. And a key limitation of the soil temperature model was its poor performance during the latter half of the testing period, showing high error. This suggests the model has limited generalization capability, likely because our simulation period did not cover a full annual cycle, including a transition from lower to higher temperatures. Future research should use a complete annual dataset to more thoroughly verify the model’s stability and applicability over a longer period.

In addition to using meteorological and groundwater data, we also verified the Informer–LSTM model’s ability to simulate soil heat flux at top soils (5 and 10 cm) in both columns. We found that using meteorological data alone consistently produced the best results. To understand why, we used SHAP (SHapley Additive exPlanations) to analyze the relationship between the model’s predictions (soil heat flux) and its various inputs. The SHAP analysis revealed that longwave radiation, shortwave radiation, and relative humidity were the most influential factors in the model’s predictions. Conversely, soil moisture and soil temperature had a minimal impact on the model’s output in most cases.

This study introduces a data-driven model to successfully simulate soil water and heat transport. However, the lack of a full or multiple full annual datasets could restrict the model’s ability to capture inter-annual variability and seasonal transitions, raising questions about the generalizability of the model, especially for extreme climatic conditions. Furthermore, we did observe in our simulations that deep learning methods exhibit limited generalization capability when modeling soil water content and temperature; incorporating physical constraints may prove an effective approach. Therefore, further research shall endeavor to collect longer-term time series data and explore integrating physical processes during both frozen and unfrozen periods with deep learning methodologies, thereby validating the model’s applicability across extended time sequences.

Author Contributions: Conceptualization, N.L. and W.W.; methodology, X.S. and N.L.; software, X.S. and P.W.; validation, N.L. and P.W.; investigation, N.L. and W.W.; resources, W.W. and Z.M.; data curation, W.W. and Z.M.; Formal analysis, X.S. and Z.M.; writing—original draft preparation, N.L. and X.S.; writing—review and editing, N.L. and P.W.; visualization, X.S. and P.W.; supervision, N.L. and W.W.; project administration, N.L.; funding acquisition, N.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key R&D Program of China (Grant No. 2023YFC3709900), the National Natural Science Foundation of China (Grant No. 42272280), and the Fundamental Research Funds for the Central Universities (Grant No. 2024SHEEAR001).

Data Availability Statement: The data presented in this study are openly available in GitHub at <https://github.com/yangtiansci-mailbox/Data-on-soil-water-content-and-temperature.git> (accessed on 12 August 2025), reference number 1044667892.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Hopmans, J.; Van Genuchten, M. Vadose zone: Hydrological process. In *Encyclopedia of Soils in the Environment*; Elsevier: Oxford, UK, 2005; pp. 209–216.
- Arora, B.; Dwivedi, D.; Faybishenko, B.; Jana, R.B.; Wainwright, H.M. Understanding and predicting vadose zone processes. *Rev. Mineral. Geochem.* **2019**, *85*, 303–328. [\[CrossRef\]](#)
- Vereecken, H.; Amelung, W.; Bauke, S.L.; Boga, H.; Brüggemann, N.; Montzka, C.; Vanderborght, J.; Bechtold, M.; Blöschl, G.; Carminati, A.; et al. Soil hydrology in the Earth system. *Nat. Rev. Earth Environ.* **2022**, *3*, 573–587. [\[CrossRef\]](#)
- Heinse, R.; Link, T.E. Vadose zone processes: A compendium for teaching interdisciplinary modeling. *J. Contemp. Water Res. Educ.* **2013**, *152*, 22–31. [\[CrossRef\]](#)
- Vanderborght, J.; Fetzer, T.; Mosthaf, K.; Smits, K.M.; Helmig, R. Heat and water transport in soils and across the soil-atmosphere interface: 1. Theory and different model concepts. *Water Resour. Res.* **2017**, *53*, 1057–1079. [\[CrossRef\]](#)
- Saito, H.; Šimuněk, J.; Mohanty, B.P. Numerical analysis of coupled water, vapor, and heat transport in the vadose zone. *Vadose Zone J.* **2006**, *5*, 784–800. [\[CrossRef\]](#)
- Garcia Gonzalez, R.; Verhoef, A.; Luigi Vidale, P.; Braud, I. Incorporation of water vapor transfer in the JULES land surface model: Implications for key soil variables and land surface fluxes. *Water Resour. Res.* **2012**, *48*, W05538. [\[CrossRef\]](#)
- Deb, S.K.; Shukla, M.K.; Sharma, P.; Mexal, J.G. Coupled liquid water, water vapor, and heat transport simulations in an unsaturated zone of a sandy loam field. *Soil Sci.* **2011**, *176*, 387–398. [\[CrossRef\]](#)
- Scharnagl, B.; Vrugt, J.; Vereecken, H.; Herbst, M. Inverse modelling of in situ soil water dynamics: Investigating the effect of different prior distributions of the soil hydraulic parameters. *Hydrol. Earth Syst. Sci.* **2011**, *15*, 3043–3059. [\[CrossRef\]](#)
- Zhou, H.; Zhao, W.Z. Evolution of soil-water states in the vadose zone of a desert soil after an extreme rainfall event and its impact on the ecosystem. *Hydrogeol. J.* **2021**, *29*, 2127–2147. [\[CrossRef\]](#)
- Lu, T.; Zheng, C.; Zhou, B.; Wu, J.; Wang, X.; Zhao, Y.; Liu, X.; Yuan, W. Response of Liquid Water and Vapor Flow to Rainfall Events in Sandy Soil of Arid and Semi-Arid Regions. *Agronomy* **2023**, *13*, 2424. [\[CrossRef\]](#)
- Gao, W.; Liu, X.; Zheng, C.; Lu, Y.; He, J.; He, Y. Comparison of the soil water, vapor, and heat dynamics between summer maize and bare fields in arid and semi-arid areas. *Agronomy* **2023**, *13*, 1171. [\[CrossRef\]](#)
- Zhang, M.; Wen, Z.; Xue, K.; Chen, L.; Li, D. A coupled model for liquid water, water vapor and heat transport of saturated–unsaturated soil in cold regions: Model formulation and verification. *Environ. Earth Sci.* **2016**, *75*, 701. [\[CrossRef\]](#)
- Wang, Y.; Shi, L.; Hu, Y.; Hu, X.; Song, W.; Wang, L. A comprehensive study of deep learning for soil moisture prediction. *Hydrol. Earth Syst. Sci.* **2023**, *28*, 917–943. [\[CrossRef\]](#)
- Mohammadi, B. A review on the applications of machine learning for runoff modeling. *Sustain. Water Resour. Manag.* **2021**, *7*, 98. [\[CrossRef\]](#)
- Sit, M.; Demiray, B.Z.; Xiang, Z.; Ewing, G.J.; Sermet, Y.; Demir, I. A comprehensive review of deep learning applications in hydrology and water resources. *Water Sci. Technol.* **2020**, *82*, 2635–2670. [\[CrossRef\]](#) [\[PubMed\]](#)
- Xu, T.; Liang, F. Machine learning for hydrologic sciences: An introductory overview. *Wiley Interdiscip. Rev. Water* **2021**, *8*, e1533. [\[CrossRef\]](#)
- Li, X.; Nieber, J.L.; Kumar, V. Machine learning applications in vadose zone hydrology: A review. *Vadose Zone J.* **2024**, *23*, e20361. [\[CrossRef\]](#)
- Minasny, B.; Bandai, T.; Ghezzehei, T.A.; Huang, Y.C.; Ma, Y.; McBratney, A.B.; Ng, W.; Norouzi, S.; Padarian, J.; Sharififar, A.; et al. Soil Science-Informed Machine Learning. *Geoderma* **2024**, *452*, 117094. [\[CrossRef\]](#)

20. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686–707. [\[CrossRef\]](#)
21. Fang, K.; Pan, M.; Shen, C. The value of SMAP for long-term soil moisture estimation with the help of deep learning. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2221–2233. [\[CrossRef\]](#)
22. Yu, J.; Tang, S.; Zhangzhong, L.; Zheng, W.; Wang, L.; Wong, A.; Xu, L. A deep learning approach for multi-depth soil water content prediction in summer maize growth period. *IEEE Access* **2020**, *8*, 199097–199110. [\[CrossRef\]](#)
23. Li, Q.; Shi, G.; Shangguan, W.; Nourani, V.; Li, J.; Li, L.; Huang, F.; Zhang, Y.; Wang, C.; Wang, D.; et al. A 1 km daily soil moisture dataset over China using in situ measurement and machine learning. *Earth Syst. Sci. Data* **2022**, *14*, 5267–5286. [\[CrossRef\]](#)
24. Alibabaei, K.; Gaspar, P.D.; Lima, T.M. Modeling soil water content and reference evapotranspiration from climate data using deep learning method. *Appl. Sci.* **2021**, *11*, 5029. [\[CrossRef\]](#)
25. Orth, R. Global soil moisture data derived through machine learning trained with in-situ measurements. *Sci. Data* **2021**, *8*, 170. [\[CrossRef\]](#)
26. Wang, Y.; Wang, W.; Ma, Z.; Zhao, M.; Li, W.; Hou, X.; Li, J.; Ye, F.; Ma, W. A deep learning approach based on physical constraints for predicting soil moisture in unsaturated zones. *Water Resour. Res.* **2023**, *59*, e2023WR035194. [\[CrossRef\]](#)
27. Kara, A.; Pekel, E.; Ozcetin, E.; Yildiz, G.B. Genetic algorithm optimized a deep learning method with attention mechanism for soil moisture prediction. *Neural Comput. Appl.* **2024**, *36*, 1761–1772. [\[CrossRef\]](#)
28. Bakhshian, S.; Zarepakzad, N.; Nevermann, H.; Hohenegger, C.; Or, D.; Shokri, N. Field-scale soil moisture dynamics predicted by deep learning. *Adv. Water Resour.* **2025**, *201*, 104976. [\[CrossRef\]](#)
29. Alizamir, M.; Kisi, O.; Ahmed, A.N.; Mert, C.; Fai, C.M.; Kim, S.; Kim, N.W.; El-Shafie, A. Advanced machine learning model for better prediction accuracy of soil temperature at different depths. *PLoS ONE* **2020**, *15*, e0231055. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Farhangmehr, V.; Cobo, J.H.; Mohammadian, A.; Payeur, P.; Shirkhani, H.; Imanian, H. A convolutional neural network model for soil temperature prediction under ordinary and hot weather conditions: Comparison with a multilayer perceptron model. *Sustainability* **2023**, *15*, 7897. [\[CrossRef\]](#)
31. Hou, J.; Wang, Y.; Zhou, J.; Tian, Q. Prediction of hourly air temperature based on CNN–LSTM. *Geomat. Nat. Hazards Risk* **2022**, *13*, 1962–1986. [\[CrossRef\]](#)
32. Imanian, H.; Mohammadian, A.; Farhangmehr, V.; Payeur, P.; Goodarzi, D.; Hiedra Cobo, J.; Shirkhani, H. A comparative analysis of deep learning models for soil temperature prediction in cold climates. *Theor. Appl. Climatol.* **2024**, *155*, 2571–2587. [\[CrossRef\]](#)
33. Didem, G. Estimation of soil temperatures with machine learning algorithms—Giresun and Bayburt stations in Turkey. *Theor. Appl. Climatol.* **2022**, *147*, 109–125.
34. Taheri, M.; Schreiner, H.K.; Mohammadian, A.; Shirkhani, H.; Payeur, P.; Imanian, H.; Cobo, J.H. A review of machine learning approaches to soil temperature estimation. *Sustainability* **2023**, *15*, 7677. [\[CrossRef\]](#)
35. Asadzadeh, F.; Emami, S.; Elbeltagi, A.; Akiner, M.E.; Rezaverdinejad, V.; Taran, F.; Salem, A. Investigating the impact of meteorological parameters on daily soil temperature changes using machine learning models. *Sci. Rep.* **2025**, *15*, 19988. [\[CrossRef\]](#)
36. Bonsoms, J.; Boulet, G. Ensemble machine learning outperforms empirical equations for the ground heat flux estimation with remote sensing data. *Remote Sens.* **2022**, *14*, 1788. [\[CrossRef\]](#)
37. Zheng, C.; Jia, L. Evaluation of different methods for soil heat flux estimation at large scales using remote sensing observations. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 6081–6084.
38. Cross, J.F.; Drewry, D.T. Ensemble machine learning for interpretable soil heat flux estimation. *Ecol. Inform.* **2024**, *82*, 102697. [\[CrossRef\]](#)
39. Yu, J.; Zhang, X.; Xu, L.; Dong, J.; Zhangzhong, L. A hybrid CNN-GRU model for predicting soil moisture in maize root zone. *Agric. Water Manag.* **2021**, *245*, 106649. [\[CrossRef\]](#)
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
41. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence 2021, Virtual, 2–9 February 2021; pp. 11106–11115. [\[CrossRef\]](#)
42. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#)
43. Elman, J.L. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211. [\[CrossRef\]](#)
44. Maxwell, R.M.; Chow, F.K.; Kollet, S.J. The groundwater–land–surface–atmosphere connection: Soil moisture effects on the atmospheric boundary layer in fully-coupled simulations. *Adv. Water Resour.* **2007**, *30*, 2447–2466. [\[CrossRef\]](#)
45. Sauer, T.J.; Horton, R. Soil heat flux. *Micrometeorol. Agric. Syst.* **2005**, *47*, 131–154.
46. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.

47. Ma, W.W.; Zhang, Z. Assessing bare-soil evaporation from different water-table depths using lysimeters and a numerical model in the Ordos Basin. *China* **2019**, *27*, 2707–2718. [[CrossRef](#)]
48. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
49. Cutler, D.R.; Edwards, T.C., Jr.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. Random forests for classification in ecology. *Ecology* **2007**, *88*, 2783–2792. [[CrossRef](#)]
50. Salman, H.A.; Kalakech, A.; Steiti, A. Random forest algorithm overview. *Babylon. J. Mach. Learn.* **2024**, *2024*, 69–79. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.