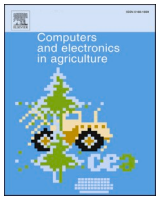
Computers and Electronics in Agriculture 230 (2025) 109802

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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



Original papers

Soil water content prediction across seasons using random forest based on precipitation-related data

Pei-Yuan Chen [a,*], Chien-Chih Chen [b,c], Chu Kang [a], Jia-Wei Liu [d], Yi-Heng Li [d]

[a] Graduate Institute of Hydrologic and Oceanic Sciences, National Central University, Taoyuan 320317, Taiwan, ROC
[b] Department of Earth Sciences, National Central University, Taoyuan 320317, Taiwan, ROC
[c] Earthquake-Disaster & Risk Evaluation and Management Center, National Central University, Taoyuan 320317, Taiwan, ROC
[d] Green Energy & Environment Research Laboratories, Industrial Technology Research Institute, Hsinchu 310401, Taiwan, ROC

ARTICLE INFO

Keywords:
Soil Water Content
Random Forest
Accumulated Precipitation
Rainy Season
Evaluation Indices

ABSTRACT

Predicting the soil water content (SWC) is crucial to prepare for and mitigate risks during dry periods, particularly before droughts. It also ensures effective water management and precise irrigation of agricultural land. Few studies have focused solely on the use of precipitation data to predict SWC. This study aimed to predict the hourly SWC for one or two days in advance at depths of 10 and 20 cm below the surface of agricultural land in Taichung, Taiwan, using Random Forest (RF), which has demonstrated promising results in previous studies. The model used hourly precipitation data from January 19, 2022, to April 18, 2023. Seven sets of strategically selected cumulative rainfall days were incorporated to balance the computational load and prediction accuracy. Based on the mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE) curves, a cumulative rainfall of 6–8 days was optimal for RF prediction at 10 and 20 cm SWC. The RF model demonstrated reasonable performance, with MAE of 0.6 % and 1.0 %, R^2 values of 0.5 and 0.9, MAPE of 25.2 % and 5.1 %, and RMSE of 2.4 % and 2.0 % for the 10 and 20 cm SWC predictions, respectively. The RF model performed well during dry periods but showed less accuracy during the rainy season, as indicated by the MAPE for the entire period, compared with the rainy season alone. This discrepancy may be due to the unusually high SWC in response to storm events. In conclusion, this study provides insights for improving SWC prediction accuracy across different seasons and practical guidelines for employing RF models in SWC prediction.

1. Introduction

Soil water content (SWC) is essential in the water cycle and significantly influences water management and agricultural production. Predicting SWC is essential to prepare for and mitigate risks during dry periods. This is particularly important before droughts as it allows for careful water management and precise irrigation of crops on agricultural land (Rani et al., 2022).

Various factors influence the SWC in the water cycle, including precipitation, evapotranspiration, infiltration, runoff, and percolation. Water balance in a hydrological model incorporating these mechanisms in each layer is one method used to predict water movement in the soil. These parametric methods require soil and surface parameters and verification (Dubois et al., 2021). The models used to simulate SWC include WOFOST, CERES, SWAP, SWAT, SWIM, and HYDRUS. The choice of model depends on the simulation purpose, spatial resolution, model complexity, and available parameter information (Eitzinger et al., 2004; Holsten et al., 2009; Neitsch et al., 2011; Tan et al., 2014).

Physical hydrological models may have high accuracy in SWC estimation but usually require significant input parameters and current meteo-hydrological variables. Therefore, machine learning (ML) methods have been applied to directly estimate SWC in cases of limited or missing input information (Karandish & Šimůnek, 2016; Dubois et al., 2021). At each current step, these methods use input information such as evaporation, air temperature, growing degree days, crop coefficient, water deficit, and irrigation depth. Nie et al. (2018) demonstrated that meteorological factors are the most influential in SWC prediction, followed by topographic factors, which are more important than soil attribute factors. Their results showed better performance of Support Vector Machine (SVM) than Random Forest (RF) and back-propagation

* Corresponding author.
E-mail addresses: pychen@ncu.edu.tw (P.-Y. Chen), chienchih.chen@g.ncu.edu.tw (C.-C. Chen), 110626006@cc.ncu.edu.tw (C. Kang), geljwl@itri.org.tw (J.-W. Liu), HenryLi@itri.org.tw (Y.-H. Li).

https://doi.org/10.1016/j.compag.2024.109802
Received 2 September 2024; Received in revised form 4 December 2024; Accepted 6 December 2024
Available online 26 December 2024

neural networks (BPNNs). Earlier days and the current day's SWC are also used to estimate future SWC (Prakash et al., 2018; Emami et al., 2024). Multiple linear regression (MLR) is superior to support vector regression and recurrent neural networks in predicting SWC for 1, 2, and 7 days in advance using the previous SWC (Prakash et al., 2018). Among the tree-based ML models, random trees performed better than REPTree and M5P in predicting SWC the next day based on SWC from three days prior (Emami et al., 2024). Surface and root-zone SWC and climate and landscape data were used in the Extreme Gradient Boosting (XGBoost) machine learning algorithm to predict multilayer SWC across the US (Karthikeyan and Mishra, 2021). Yu et al. (2022) used gridded meteorological data and SWC as inputs for the proposed ResBiLSTM to improve the precision of SWC predictions at various depths. Input variables are essential for ML performance; however, data availability may limit the application of these methods. Land/soil temperature, which is usually not readily available at a small scale, is used as an input variable in addition to meteorological variables when using a promising decision tree or a Deep Learning Regression Network (DLRN) to estimate SWC (Cai et al., 2019; Pekel, 2020). Data fusion of the normalized difference vegetation index (NDVI), surface albedo, and land surface temperature (LST) combined with RF provides reasonable estimates of spatiotemporally continuous soil moisture (Abowarda et al., 2021). Soil temperature was measured, and 13 other variables, including relative humidity, temperature, total radiation, and evapotranspiration, were used as inputs to the four ML methods by Nath et al. (2024). Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) outperformed CNN, LSTM, and MLR in their study.

The performances of RF, SVM, and deep learning methods such as neural networks are usually compared with varying results. SVM provides stable and unique soil moisture estimation for one of the regional stations with less computational burden than artificial neural networks (ANN), based on a model trained using past and current soil moisture and atmospheric data from neighboring measurements (Gill et al., 2006). Eight variables, including meteorological factors, potential evapotranspiration (ET), and SWC, were input to a CNN-LSTM to estimate daily SWC and ET using CNN, SVM, and RF (Alibabaei et al., 2021). Moreover, the parameters of the hydrological model were derived using the ML method. For example, Tramblay and Quintana Seguí (2022) used RF to estimate the maximum water-holding capacity of the Iberian Peninsula as input for the Soil Moisture Accounting Model (SMA) based on altitude, temperature, precipitation, evapotranspiration, and land use.

Satellite data, with limited penetration to the soil surface, are used in addition to local data for the current spatial estimation of SWC. RF has shown better results than SVM in estimating SWC based on GNSS-R Soil Moisture Retrieval (Jia et al., 2020). RF has also shown promising results in estimating soil moisture using subtractive clustering (SBC) and an adaptive neuro-fuzzy inference system (ANFIS) with data from dual polarimetric Sentinel-1 radar backscatter (Chaudhary et al., 2022). An unmanned aerial vehicle (UAV) is an option for resolutions higher than a few meters provided by satellite (Tmušić et al., 2020). Guan et al. (2022) found that RF could predict volumetric water content and electrical conductivity. It performed exceptionally well when crop types and degrees of drainage in soybean and corn fields were relatively homogeneous and for dry soil with low variability, considering the UAV-acquired vegetation indices and multispectral data. Multispectral data based on UAV help predict the SWC 5 cm below the surface when input into the RF regression, which outperforms the Elastic Net, General Linear Model, and Robust Linear Model (Bertalan et al., 2022). RF also performed comparably to a process-based model in an 18 cm SWC nowcasting using vegetation indices from sites in the Netherlands, except for extremely dry or wet cases with fewer samples for learning (Carranza et al., 2021). In addition, visible and near-infrared spectroscopy of fresh soil samples were used as inputs to estimate SWC. The least-squares SVM overperformed the cubists and two other multivariate regression methods (Morellos et al., 2016). Liu et al. (2023) used 32 characteristic parameters derived from images of in-field soil samples as inputs for four ML models; however, this was time-consuming. Among these methods, Gaussian process regression (GPR) is better than partial least squares regression (PLSR), random forest (RF), or support vector machine regression (SVMR).

However, prediction is more applicable when the input datasets exclude the current data. Wu et al. (2007) found similar results for SVM and ANN considering performance, which were used to predict soil moisture five days later, given the average and five-day soil moisture. Using a small test dataset from three sites and three dates, Gorthi and Dou (2011) showed that the decision-tree-based model M5 performs comparably or even better in estimating daily surface soil moisture according to meteorological, soil, and vegetation indices than the neural network-based Multilayer Perceptron Network. Vyas and Bandyopadhyay (2020) proposed a semi-supervised machine learning method based on a graph neural network to capture the spatiotemporal correlation of soil moisture at stations across Spain and the US on a daily or 15-day scale, addressing the common problem of missing ground truth SWC. Prasad et al. (2018) proposed a model based on ANN to forecast monthly soil moisture, which outperformed other models, including RF. In contrast, RF and neural networks exhibited comparable performances of soil water potential estimations in potato fields (Dubois et al., 2021).

According to a literature review, RF is applicable to SWC prediction and is better than the Classification and Regression Tree (CART) for overfitting, accuracy, and efficiency when dealing with complex datasets (Hastie et al., 2009; Rani et al., 2022). Moreover, RF is relatively easy to understand and interpret compared to other black-box models, such as ANN and SVM (Rani et al., 2022). However, few studies have used only data derived from precipitation to predict SWC. Therefore, this study aimed to predict the hourly SWC at 10 and 20 cm below the surface of agricultural land using RF based on commonly available hourly precipitation data. The novelty of this study is that SWC prediction is based merely on readily available precipitation data compared to other potential variables influencing SWC. This broadens the application of predicting SWC at locations with only rainfall data, enhancing agricultural and water management and early preparation.

2. Methodology

This study used RF to predict the hourly SWC at depths of 10 and 20 cm below the surface of agricultural land. The target variable was SWC, and the input variables included precipitation-related data and daily hours. A previous study proposed that RF could automatically retrieve useful input information from complex input sets (Wu et al., 2024). Therefore, various precipitation-related data, including hourly precipitation (hourly P.) and accumulated precipitation (Accum. P.), 24-h-ago precipitation (24-ago P.), and no-rain hours, together with daily hours, imply the influence of evapotranspiration on SWC.

Considering the different possible combinations of accumulated precipitation, this study examined the results of different cumulative rainfall events per day. The importance of the different input variables is further discussed. To determine the model's sensitivity to data length and characteristics, the optimal Accum. P. and the performance of the RF during the rainy season was compared with the entire period. Moreover, validation and prediction errors were used to justify the selection of Accum. P. and addressed under/overfitting issues. A flowchart of this process is shown in Fig. 1. The following sections describe the model/data and indices used to evaluate the performance of the model.

2.1. Data

This study used data from farmlands at the Taiwan Agricultural Research Institute (ARI) in Central Taiwan, where detailed SWC and precipitation data are available. The soil at a 0–30 cm depth comprised tiny particles of silty or sandy loam. Newly built hydrological and meteorological stations were located at different sites, as shown in

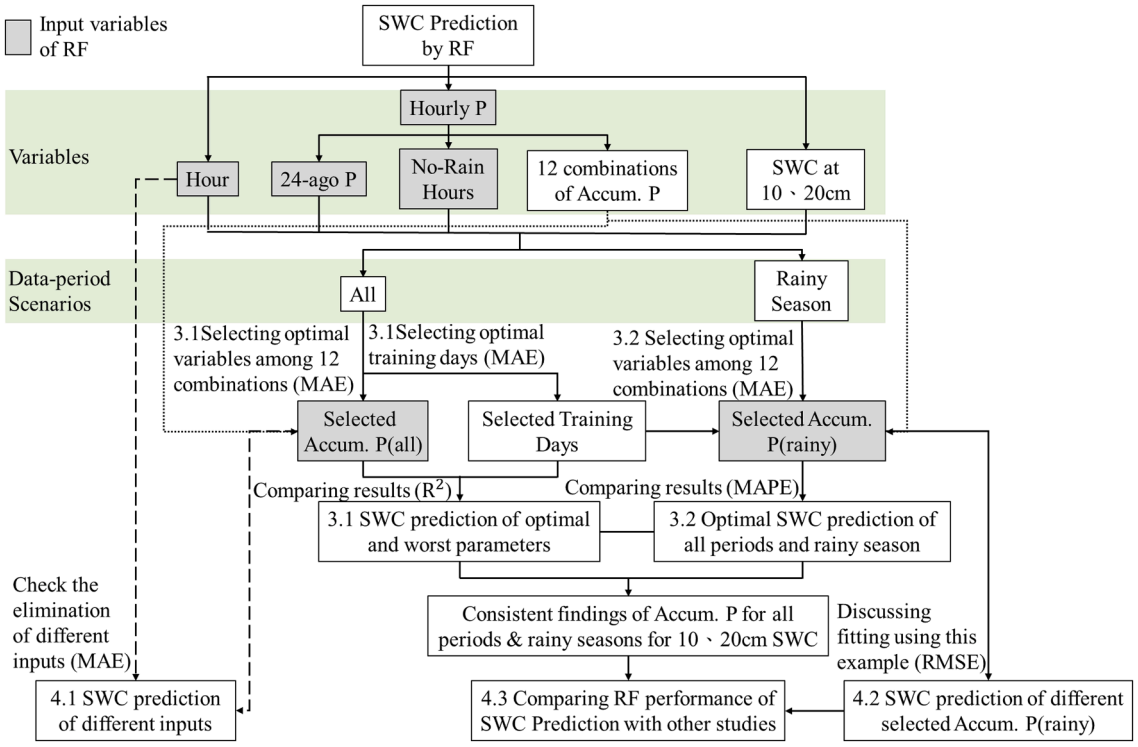


Fig. 1. Flowchart showing the influence of parameters, inputs (marked ones), and rainy seasons on soil water content (SWC) prediction by Random Forest (RF) analyzed in this study.

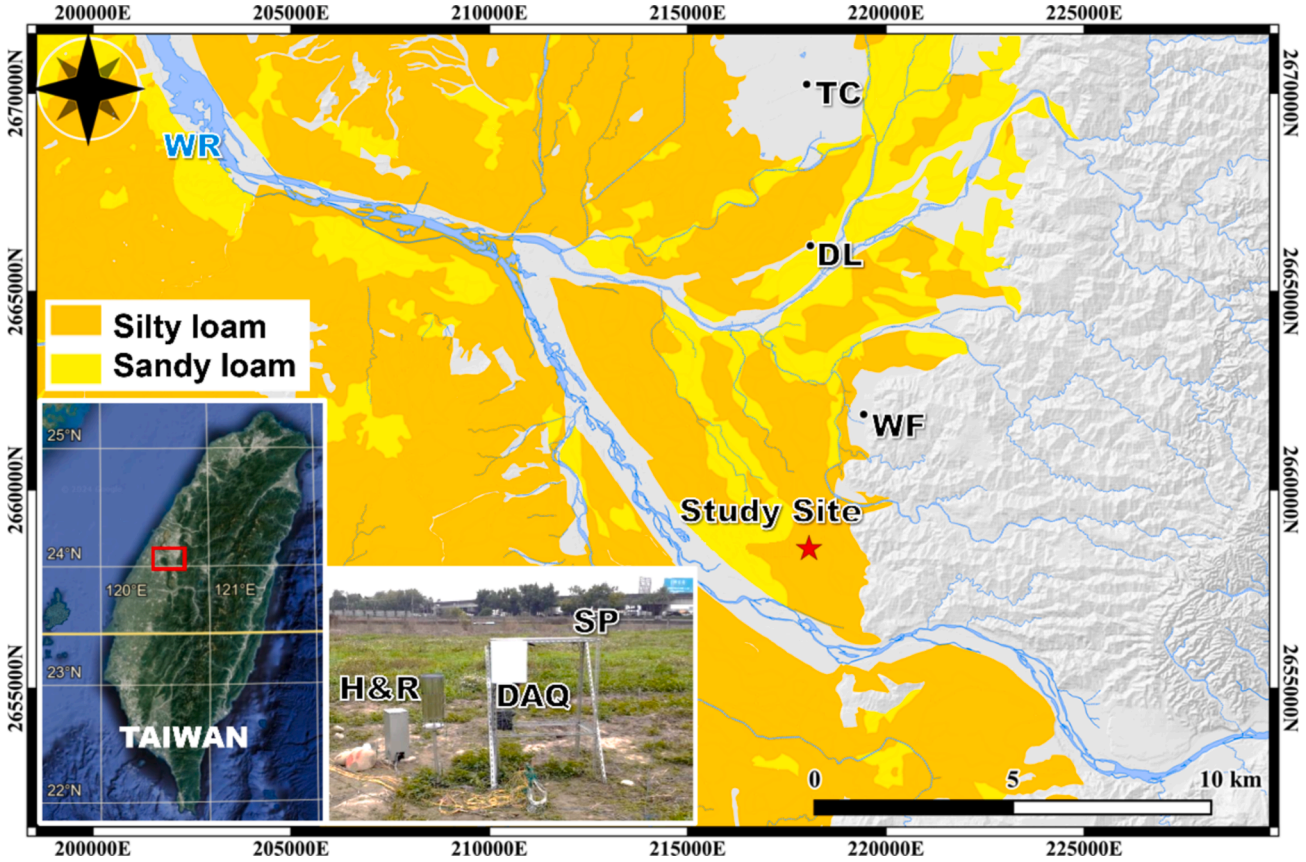


Fig. 2. Location of the study site and observation system, including hydrological measurements (SWC) and rain gauge (H&R), Data Acquisition (DAQ), and Solar Panel (SP). WR, Wu River; TC, Taichung; DL, Dali; WF, Wufeng.

Fig. 2. We used an RS-102D rain gauge made by Ogasawara, and volumetric SWC was measured using the SDI-12 from Sentek EnviroSCAN technology with an accuracy of ± 0.003 % Vol. An experienced team ensured minimum soil disturbance during EnviroSCAN installation, and the SWC data were verified with data measured from a nearby 2725ARL Jet-Fill Tensiometer (Supplementary Material).

Data from January 19, 2022, to April 18, 2023, were recorded at 10-minute intervals and transformed into hourly data for this study. For example, the average soil data recorded at 07:10, 07:20, 07:30, 07:40, 07:50, and 08:00 h were used as the data at 08:00 h. Most of the precipitation occurs from May to September, with the highest rainfall reaching 80 mm/h. The SWC at 10 cm depth changed more intensively than that at 20 cm depth, likely due to a more significant influence from meteorological factors. The range of SWC was 10– 47 %, as shown in Fig. 3.

2.2. Random Forest and input features

This study used the RF Model to predict SWC, which trains multiple trees to reduce the risk of overfitting in the decision tree method. RF is used owing to its practical advantages and strong performance, particularly in scenarios with limited data where interpretability is essential (Rani et al., 2022). The MATLAB toolbox for Statistics and Machine Learning, which features a convenient Graphical User Interface and Integrated Development Environment, was used for the RF cross-validation and prediction. A Bagging algorithm, which employs the bootstrap resampling method, was adopted to develop several tree classifiers (Sutton, 2005). The tree classifier proposed by Breiman et al. (1984) is based on the Gini Index (Sutton, 2005). Specifically, a classifier is based on a subset of new samples obtained from the original training samples. By repeating the resampling process several times, multiple classifiers are ensembled and used to predict the target variable based on the testing data of the input features. In this study, randomly selected features were used to identify the feature category with a small Gini Index representing data purity within each category and develop the Classification and Regression Trees (CART), for example, the classifiers (Wu et al., 2024). The optimal number of training days was examined to ensure sufficient training data for accurate prediction. Optimal hyperparameters, such as leaf size and learning cycle, were manually identified. Considering that the prediction errors increased with increasing prediction lengths and that the surface 10 cm SWC responds more directly to precipitation than the 20 cm SWC, the prediction periods were set to two days for the 10 cm SWC and one day for the 20 cm SWC. These settings were adjustable for prediction and decision-making purposes.

The input variables encompass various precipitation-related data, including hourly rainfall and combinations of accumulated rainfall, dry hours, and daily hours, ensuring a comprehensive consideration of factors influencing SWC. Precipitation is the direct input of soil water, and accumulated precipitation is the residence time of soil water caused by previous rainfall. In contrast, dry and daily hours help capture the influence of evapotranspiration, a possible loss of SWC dominated by solar radiation and SWC. Daily hours also show the possible percolation loss of surface SWC after rainfall and seasonal/diurnal effects on SWC.

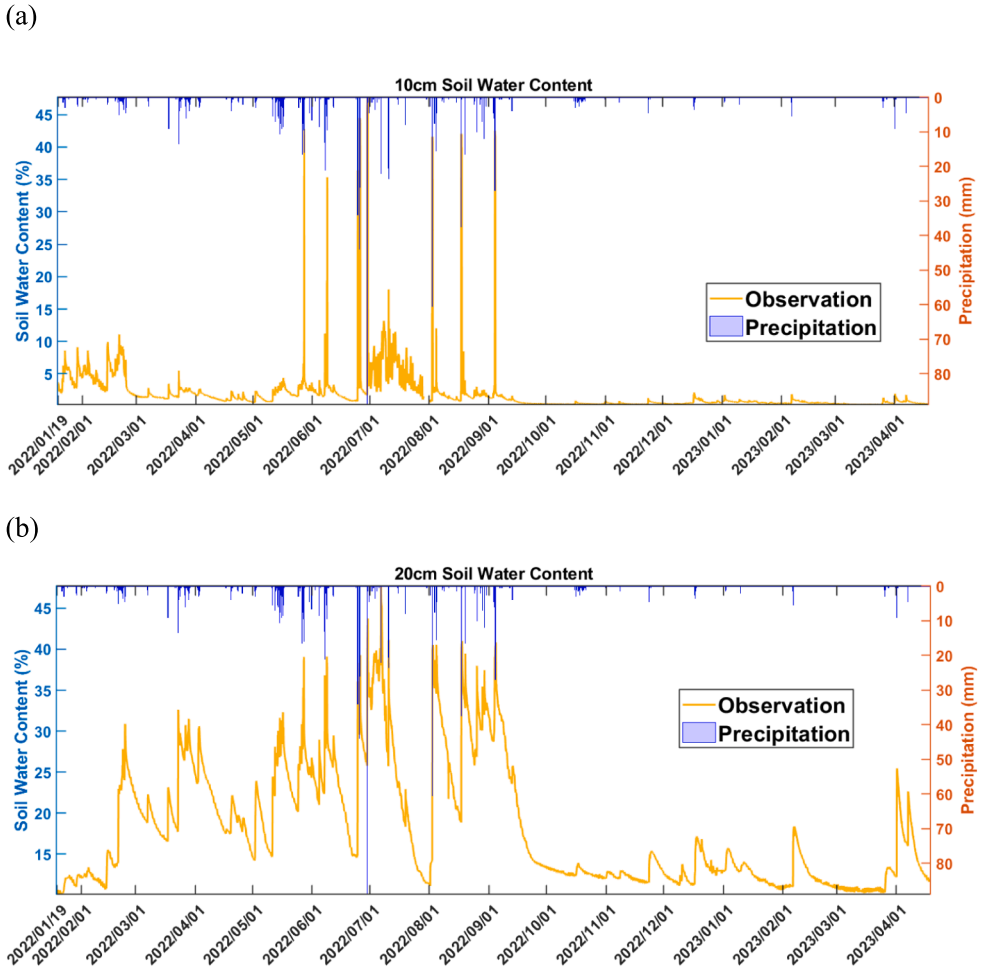


Fig. 3. (a) 10 and (b) 20 cm soil water content observations and precipitation at the study site at the Agricultural Research Institute from 19 January 2022 to 18 April 2023.

However, the selection of the accumulated precipitation remains a challenge. Twelve combinations, each containing seven variables of Accum. P., were tested in this study, and the largest Accum. P. ranged from 12 to 432 h (18 days) for the 12 combinations, as shown in Fig. 4. The selection of the seven variables of Accum P. is subjective, and its influence is discussed in Section 4.1.

2.3. Model performance evaluation indices

This study used four commonly used indices to evaluate the performance of RF in predicting SWC from different perspectives to gain further insight into the RF results. The indices used were the mean absolute error (MAE), mean absolute percentage error (MAPE), root mean square error (RMSE), and coefficient of determination (R^2).

The MAE of all the predictions was calculated as an average value to represent the overall performance during the prediction period. For specific cases, as shown in Fig. 1, the MAE of each prediction window (one or two days) was also presented using box plots rather than an average value. The MAE ranged from zero to infinity, with values approximately zero indicating precise predictions with minor errors. However, the MAE was influenced by the absolute values of the data.

The MAPE was also applied to compare the results for the entire period and the rainy season to assess the percentage of error rather than the absolute value. Lower MAPE values indicate more minor errors, but the bias was magnified when the absolute value was approximately zero. According to Lewis (1982), a MAPE within 20–50 % is a reasonable forecast, whereas values smaller than 10 % are highly accurate.

The validation and prediction RMSE were calculated using the MATLAB toolbox, which helps clarify the underfitting and overfitting challenges in this study. Smaller RMSE values, approximately zero, are favorable. The coefficient of determination (R^2) shows how well the points can be described by the linear regression model found in this study, ranging from 0 to 1 for the scatter plot of SWC observations and predictions. The equations of MAE, MAPE, RMSE, and R^2 were obtained from Chai and Draxler (2014) and Chicco et al. (2021).

3. Results

3.1. Influence of cumulative rainfall by days and training data length

The Accum. P. reflects the water entering the soil in the past, and its amount depends on the cumulative rainfall per day. The influence of the daily cumulative rainfall on SWC can be attributed to precipitation patterns, surface/soil properties, and other meteorological factors that dominate evapotranspiration. Extensive training data may improve RF performance owing to ergodicity (Bucci et al., 2020), where the model learns abundant materials or experiences. However, trade-offs concerning data collection costs and side effects such as overfitting may exist.

We first manually searched for optimal leaf size and learning cycle settings. Subsequently, various combinations of cumulative rainfall per day, as shown in Fig. 4, and durations of training data from 2 to 90 days were tested to identify the optimal training duration and Accum. P. variables. The leaf size represents the minimum number of data samples in each leaf node in a decision tree. Large leaves form a simple tree and may lead to underfitting, whereas small leaves with more complicated trees may cause overfitting. The learning cycle is the number of iterations for updating the weights during tree training. A small learning cycle may accelerate model convergence but carries the risk of underfitting.

The MAE of 12 cumulative rainfall per day is shown in Fig. 5, considering a leaf size of 10, a learning cycle of 500, and the best training days of 20 and 10 days for 10 and 20 cm SWC, respectively. The best cumulative rainfall per day was selected based on the lowest MAE, which was eight days for the 10 cm SWC. The curve stabilized after 6–8 days for the 20 cm SWC when observing the trends rather than the detailed values. This is consistent with a previous study in Taiwan, in which the MAE of shallow SWC prediction was reduced when considering past meteorological information for up to 192h (8 days) ago (Wu et al., 2024). Because the 10 cm SWC had much smaller values than the 20 cm SWC, the average MAE values of the 10 cm SWC were also smaller than those of the 20 cm SWC.

Fig. 6 shows the impact of changing the training data length on the MAE with the optimal cumulative rainfall per day. The training duration was increased from 2 to 10 days incrementally and up to 90 days in 10-day increments. The best training durations were selected based on the lowest points, 20 and 10 days, for the 10 and 20 cm SWC, respectively. Long training durations did not necessarily result in small MAEs. This may be because the training data must cover approximately 1.5–3 times (10 or 20 days), each with 6–8 days, for the RF model to learn the SWC fluctuations caused by local meteo-hydrological processes. The periodicity issue was also discussed in a previous study on deep learning (Zhang et al., 2020). Earlier information is a minor disturbance rather than helpful in predicting SWC.

The SWC predictions and errors for the optimal variable parameters are shown in Fig. 7. The MAE of 10 and 20 cm SWC were 0.58 and 0.90, respectively. RF typically performed reasonably well in predicting SWC, with an MAE smaller than 10 %, except for extreme rainfall-induced SWC spikes during mid-2022. The limited learning sample of dramatic SWC changes during extreme rainfall events inevitably leads to

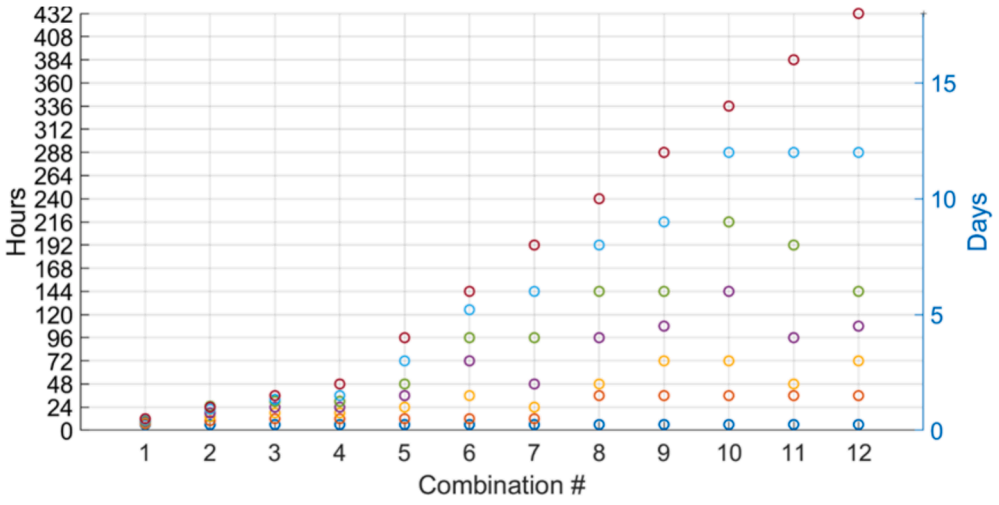


Fig. 4. Twelve combinations of cumulative rainfall by days, each containing seven variables of accumulated precipitation (Accum. P.).

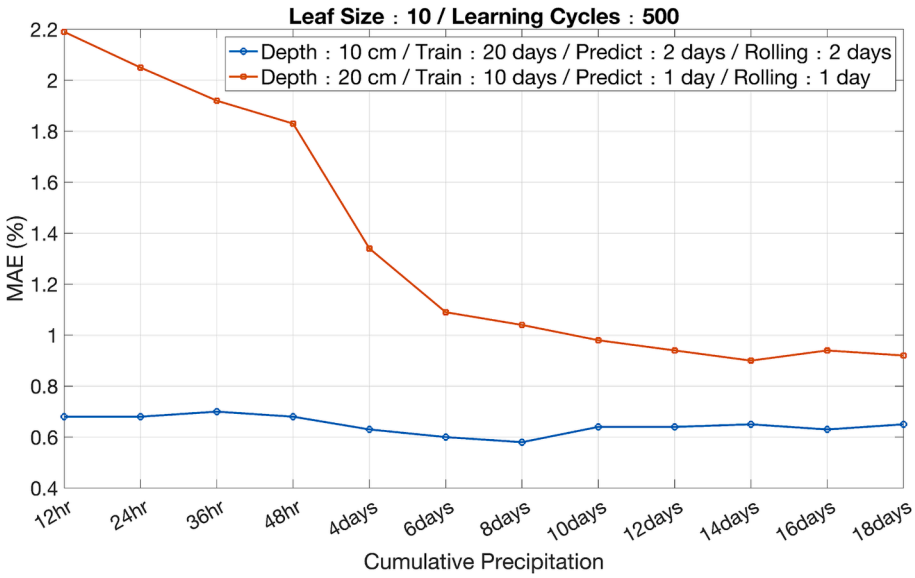


Fig. 5. Average mean absolute error (MAE) of 12 sets of cumulative rainfall by days for 10 and 20 cm soil water content for entire periods. The labels of horizontal axes indicate the longest cumulative days in every combination in Fig. 4.

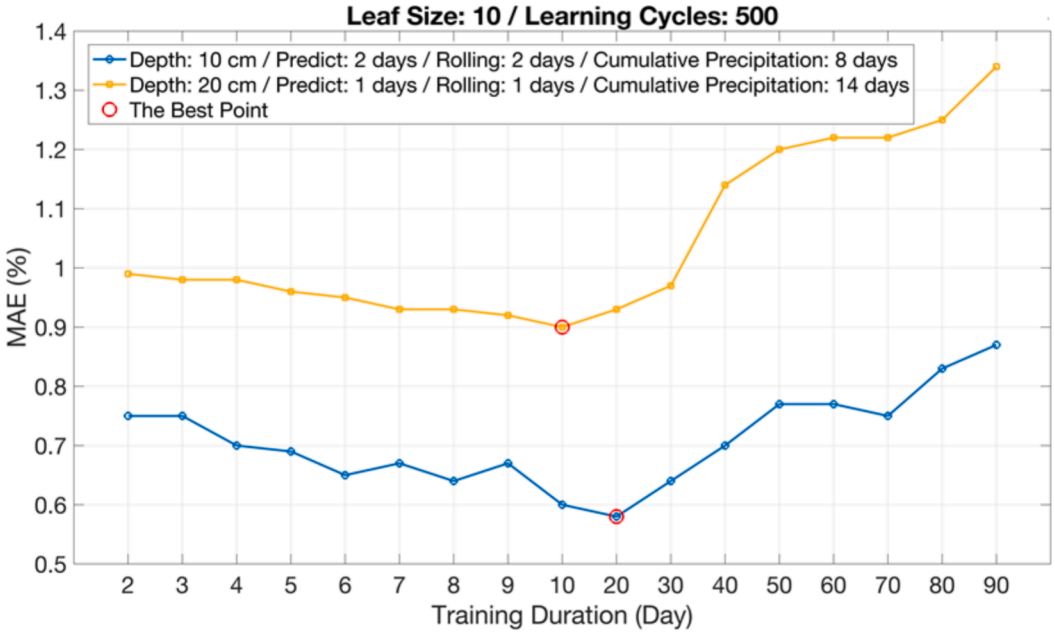


Fig. 6. Average mean absolute error (MAE) of 10 and 20 cm soil water content (SWC) using Random Forest (RF) based on different training days for entire periods.

significant prediction errors. The results for the rainy season are discussed in Section 3.2.

In addition to the results based on the optimal parameters, the results for the worst parameters are presented in Table 1. Importantly, the manually identified optimal and worst parameters refer to the training duration and rolling/predicting periods and not the input variables (cumulative rainfall by days). Changing the leaf size from 10 to 1000 and the learning cycle from 100 to 1000 yielded the same results; therefore, they remained constant. The R^2 of the results under the optimal setting showed an improvement of more than 0.2 compared to the worst setting for both 10 and 20 cm SWC. The predicted and observed 10 and 20 cm SWC were also compared in scatter plots for the optimal and worst parameters in Figs. 8 and 9, respectively. Evidently, under the optimal parameters, the predictions were closer to the observations based on the 1:1 line than to the worst parameters, highlighting the impact of the proper training duration and rolling/predicting periods. For the 10 cm SWC, this effect was particularly noticeable in the zoomed-in 0 %–10 % range of SWC, which occurred most frequently (Fig. 8(b) and (d)).

3.2. SWC prediction in the rainy season

The SWC prediction during the rainy season was less accurate than that during the dry periods (Fig. 7). Therefore, this study further evaluated the average prediction error during the rainy season, which lasted for five months, from April 16 to September 16, 2022. The most sensitive variable, Accum. P., was also tested to quantify the improvement in prediction when this variable was adjusted.

Initially, MAE was examined; however, the observed SWC influenced the values (Fig. 10). Therefore, MAPE, which shows the relative error,

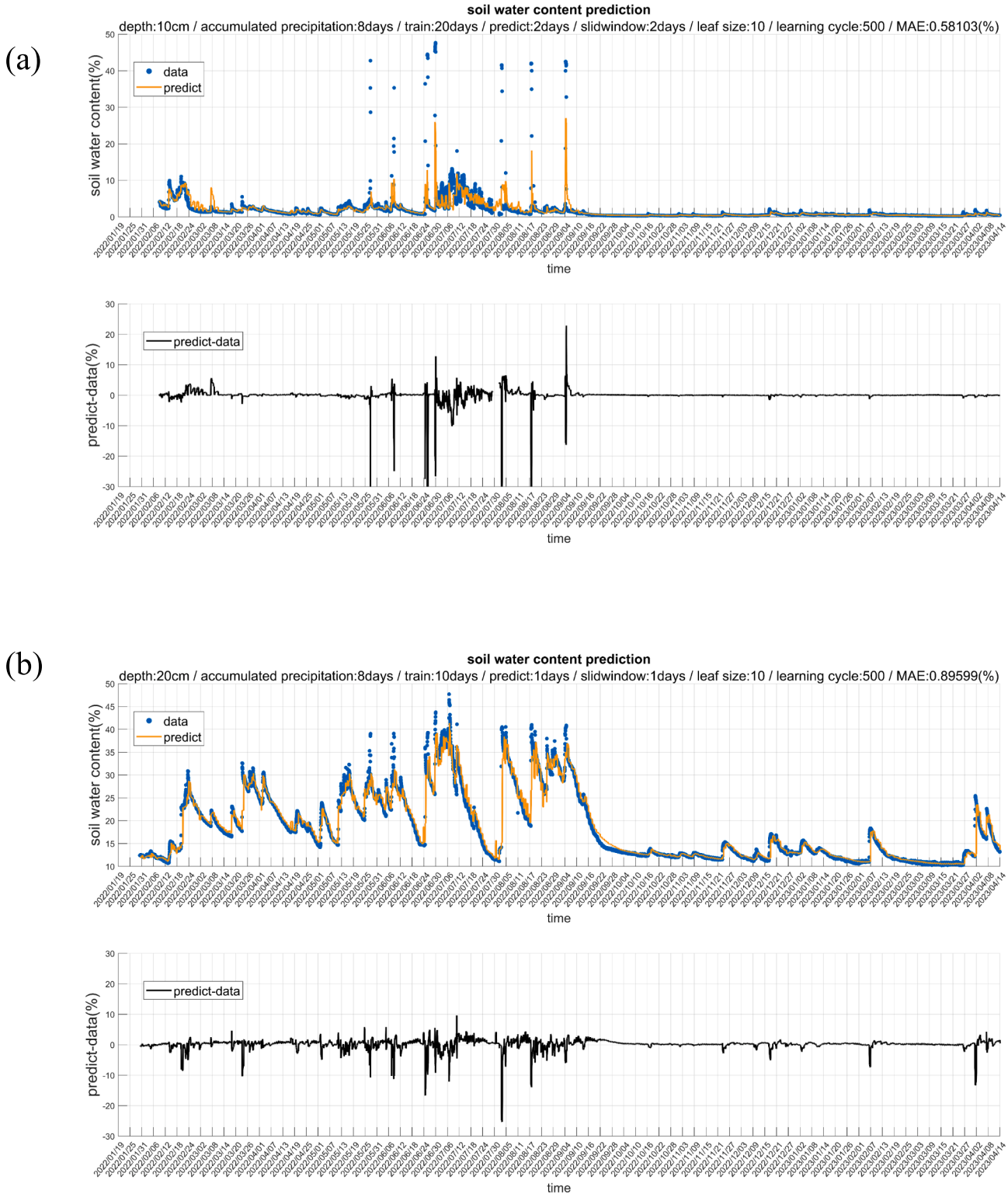


Fig. 7. (a) 10 and (b) 20 cm soil water content predictions and error time series under optimal parameter settings.

was used to compare predictions for the entire period and rainy season under different Accum. P. settings (Fig. 11). Lewis (1982) stated that MAPE values greater than 50 % are categorized as inaccurate forecasting. Consequently, the 10 cm SWC predicted by RF during the rainy season was considered unreliable compared to the other curves in Fig. 11. This was because of a few extreme SWC values/outliers in response to storm events, which were insufficient for the RF model to learn. Although the predicted SWCs were high, they were noticeably lower than the observed SWCs.

Based on the previous discussion, the curves stabilized again after 6 to 8 days, disregarding the 1 % difference in the MAPE. These results were influenced by precipitation, infiltration, and evapotranspiration

Table 1
Optimal and worst simulation settings and performance (R^2).

		10 cm SWC	20 cm SWC
Optimal setting ←→Worst setting	Training duration (days)	20 ←→90	10 ←→90
	Rolling/predicting periods(days)	2 ←→13	2 ←→14
Optimal results ←→Worst results	Performance (R^2)	0.52 ←→0.30	0.95 ←→0.69
Unchanged setting	Leaf size	10	
	Learning Cycle(times)	500	
	Cumulative rainfall by days (days)	8	14

under local weather conditions and soil properties. The performance of RF reached a MAPE of 5.1 %, corresponding to an MAE of 1.0 % and 6.6 % for 20 cm SWC prediction during the entire period and the rainy season, respectively. For a 10 cm SWC, a MAPE of 25.2 % was a reasonable forecast for the entire period, corresponding to an MAE of 0.6 %. For the small values of 10 cm SWC observed most of the time during the entire period, the small MAE and large MAPE demonstrated the effects of using different evaluation indices.

4. Discussion

4.1. Influence of input variables on predicting errors

Fig. 12 shows the MAE of each prediction window, highlighting the prediction errors associated with different input variables. To assess the impact of each variable, one type of input variable was excluded at a time to determine whether the MAE worsened compared to the all-variable case. Red crosses depict outliers. These results underscore the

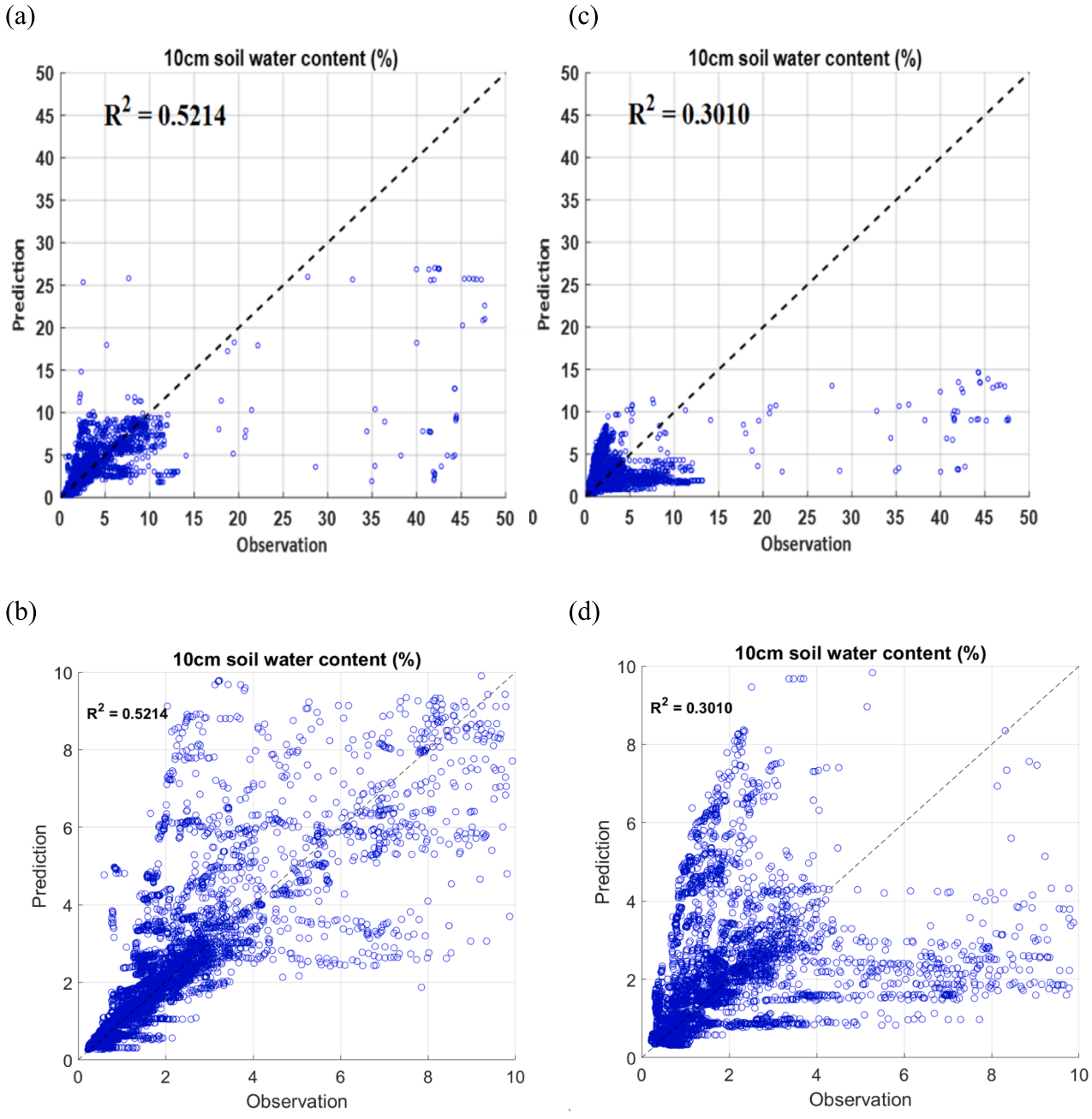


Fig. 8. Scatter plot of predicted and observed soil water content at 10 cm below the surface under the (a)(b) best and (c)(d) worst parameter settings with a scale of (a)(c) 0 %–50 % and (b)(d) 0 %–10 %.

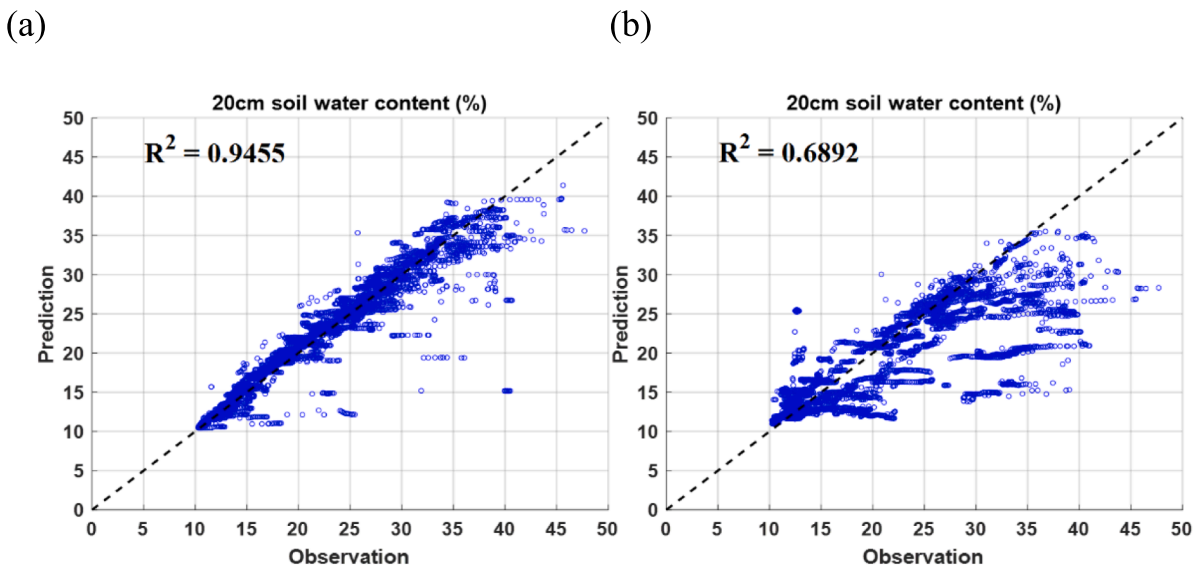


Fig. 9. Scatter plot of predicted and observed soil water content at 20 cm below the surface under the (a) best and (b) worst parameter settings.

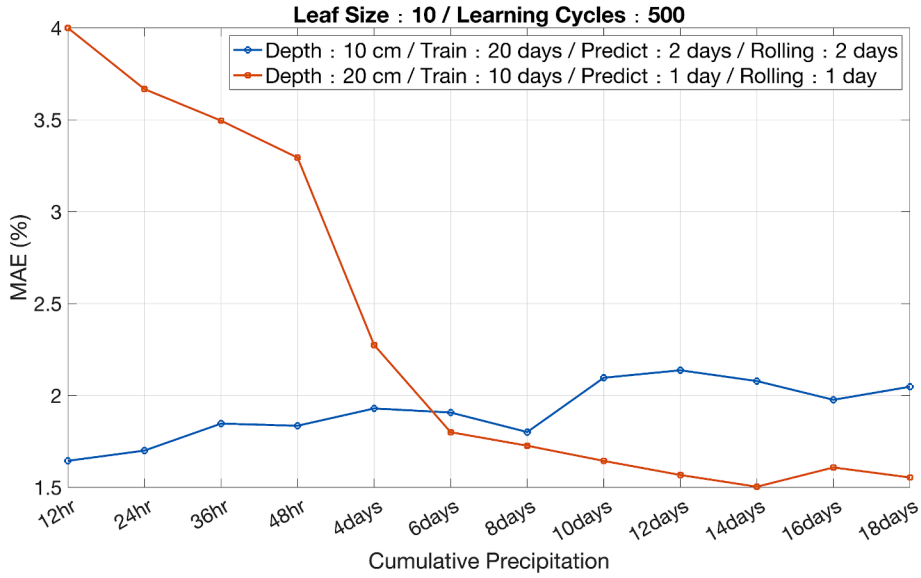


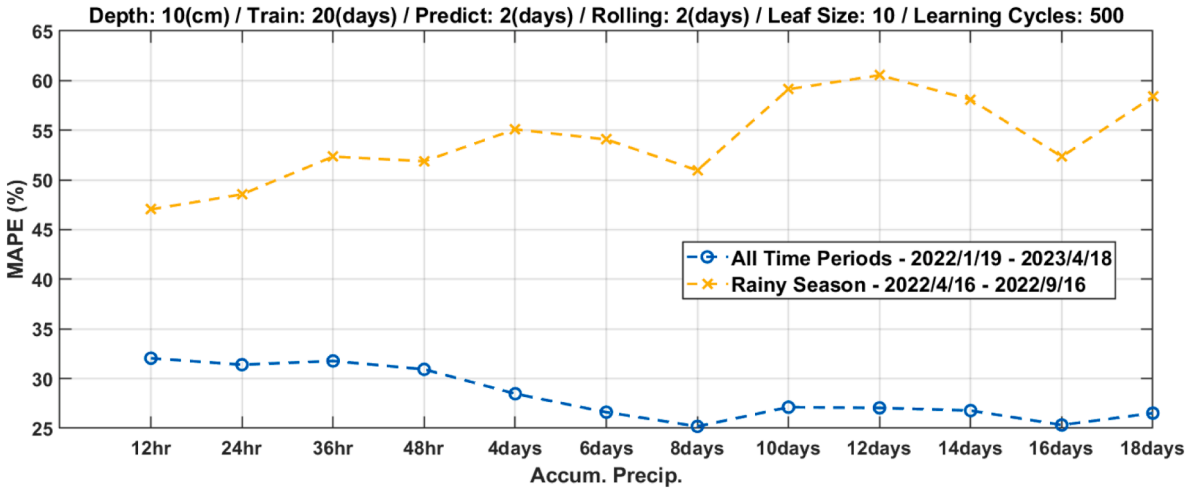
Fig. 10. Average mean absolute error (MAE) of 12 sets of cumulative rainfall by days for 10 and 20 cm soil water content for the rainy season. The labels of horizontal axes indicate the longest cumulative days in every combination in Fig. 4.

importance of using Accum. P. in SWC prediction, and excluding other variables had minimal impact. Because the RF automatically retrieves valuable information, all variables influencing SWC can be used as inputs to maintain a variety of characteristic variables (Wu et al., 2024). Here, Accum. P. refers to the 10th combination in Fig. 4, which consists of seven cumulative rainfall events per day.

This study further examined the influence of different accumulated rainfall inputs on MAE using the 20 cm SWC prediction as an example. Three cases were considered: seven sets of strategically selected cumulative rainfall per day, seven sets of randomly selected cumulative rainfall per day, and 16 sets of cumulative rainfall per day, as shown in Fig. 13. This was to verify the effectiveness of using seven sets of strategically selected cumulative rainfall per day rather than an entire set. Each point in Fig. 13 represents the cumulative rainfall appended to that day. For example, the last red point in Fig. 13 represents the prediction error of the seven sets of strategically selected cumulative rainfall per day, including 0.25, 1.5, 3, 6, 9, 12, and 14 days. Note that the 10th combination in Fig. 4, up to 14 days, is used as an example instead of the optimal result of up to 8 days identified earlier.

As precipitation accumulates over extended periods, the prediction error converges because of the availability of sufficient information for the RF model to learn. The appended accumulated rainfall time for all 16 sets of cumulative rainfall per day (in blue) had the highest MAE among the three cases when accumulated for more than four days. However, this case had the highest computational load. However, the seven sets of strategically selected cumulative rainfall per day showed comparable errors to the entire 16 sets of cumulative rainfall per day, making them more efficient in computation. Moreover, the seven strategically selected and seven randomly selected cumulative rainfall per day achieved predictions as good as the 16 sets of cumulative rainfall per day when accumulated for more than 10 days. This indicates that when accumulated for 13 or 14 days, any of the seven selected sets of cumulative rainfall per day have an insignificant impact.

(a)



(b)

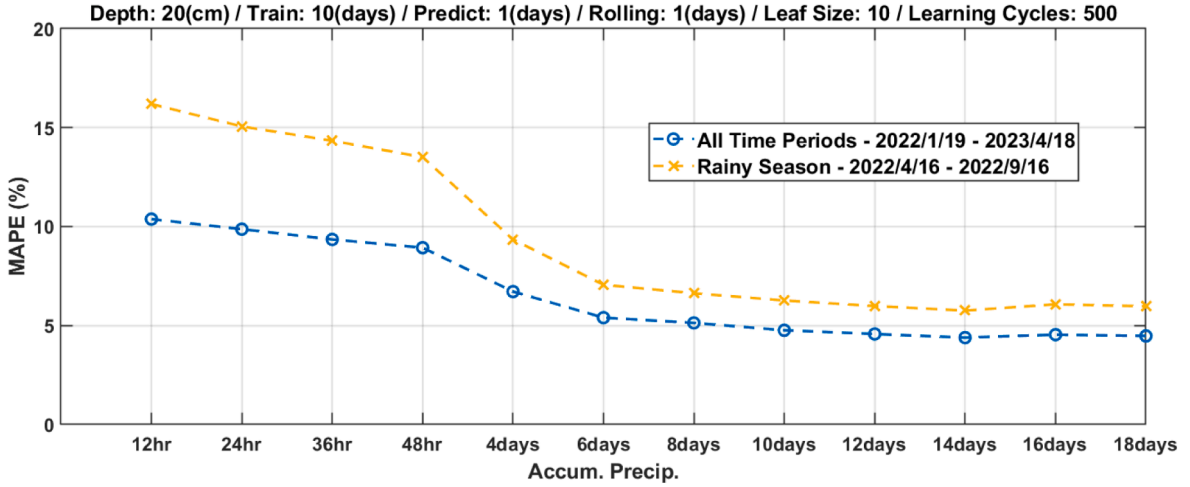


Fig. 11. Average mean absolute percentage error (MAPE) of 12 sets of cumulative rainfall by days for (a) 10 and (b) 20 cm soil water content for the whole period and the rainy season. The labels of horizontal axes indicate the longest cumulative days in every combination in Fig. 4.

4.2. Discussion on Under/Overfitting

Data from the rainy season were used to discuss the underfitting and overfitting issues concerning the 12 combinations of cumulative rainfall per day. The training days, leaf size, and learning cycles were the same as those used for the entire period predictions. The RMSE was used for convenience, utilizing the MATLAB toolbox command to show the validation error based on the cross-validation. The validation and prediction errors are shown in the upper panels of Fig. 14(a) and (b), respectively, and the differences between them are shown in the lower panels.

For the 10 cm SWC, the smallest RMSE of the validation error was found when the cumulative rainfall per day reached eight days. Both the validation and prediction errors increased for more extended cumulative rainfall per day, which may imply that the Accum. P. earlier than eight days may provide abortive information for predicting future SWC. Furthermore, the prediction errors were slightly larger than the validation errors, indicating the onset of overfitting. The differences between the validation and prediction errors were plotted, and a quadratic curve was fitted. The minimum difference was approximately eight days, showing that both errors had similar values despite the two error curves crossing the upper panel.

For the 20 cm SWC, the validation error, prediction error, and the difference between them started to show convergence for the cumulative rainfall again by eight days, which is the 6th combination shown in Fig. 4. The results were consistent with those for 10 cm SWC and those found based on the MAPE calculated for the entire period for 10 and 20 cm SWC, the rainy season for 20 cm SWC, and MAE for the entire period for both depths. The performance of the RF reached RMSE values of 2.4 % and 2.0 % for the 10 and 20 cm SWC predictions, respectively.

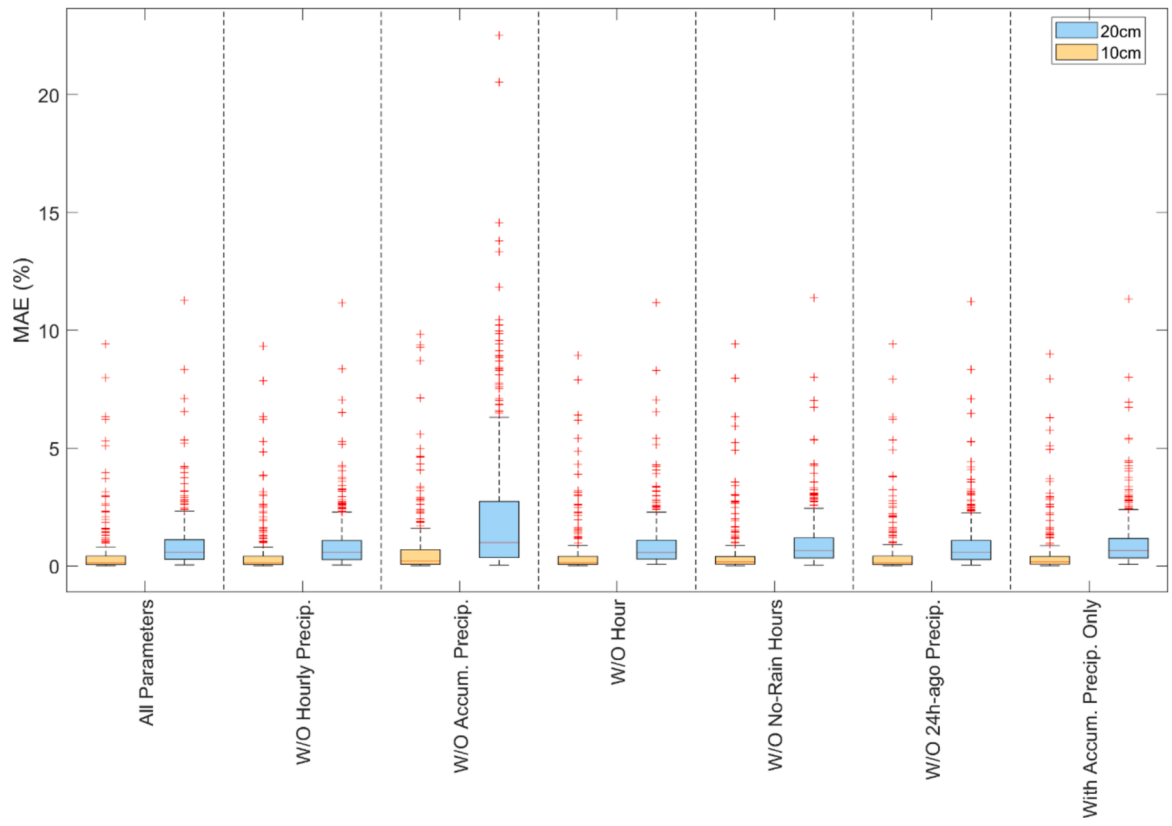


Fig. 12. Mean absolute error (MAE) of each soil water content prediction window for different input variables, where hourly Precip. is the hourly precipitation, Accum. Precip. is the accumulated precipitation, 24 h-ago Precip. is the 24-hour-ago precipitation.

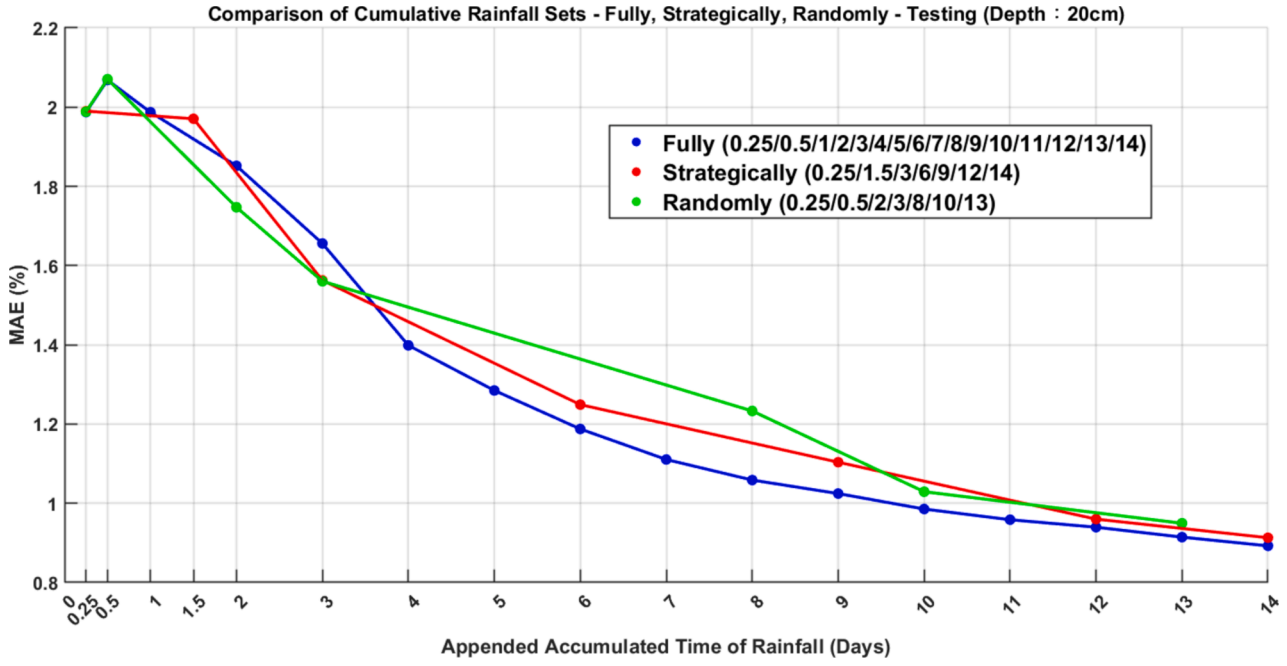


Fig. 13. Average mean absolute error (MAE) of the seven sets of strategically selected cumulative rainfall by days, the seven sets of randomly selected cumulative rainfall by days, and the entire 16 sets of cumulative rainfall by days up to 14 days under different appended accumulated times of rainfall.

4.3. Performances of Random Forest in estimating SWC

Carranza et al. (2021) explored the potential of machine learning (ML) for nowcasting root zone soil moisture (RZSM) at various spatial and temporal scales, primarily for agricultural and hydrological applications. Their study compared ML predictions to those from process-based models, highlighting the flexibility of RF and fewer assumptions regarding the underlying soil processes. Carranza et al. (2021) used RF

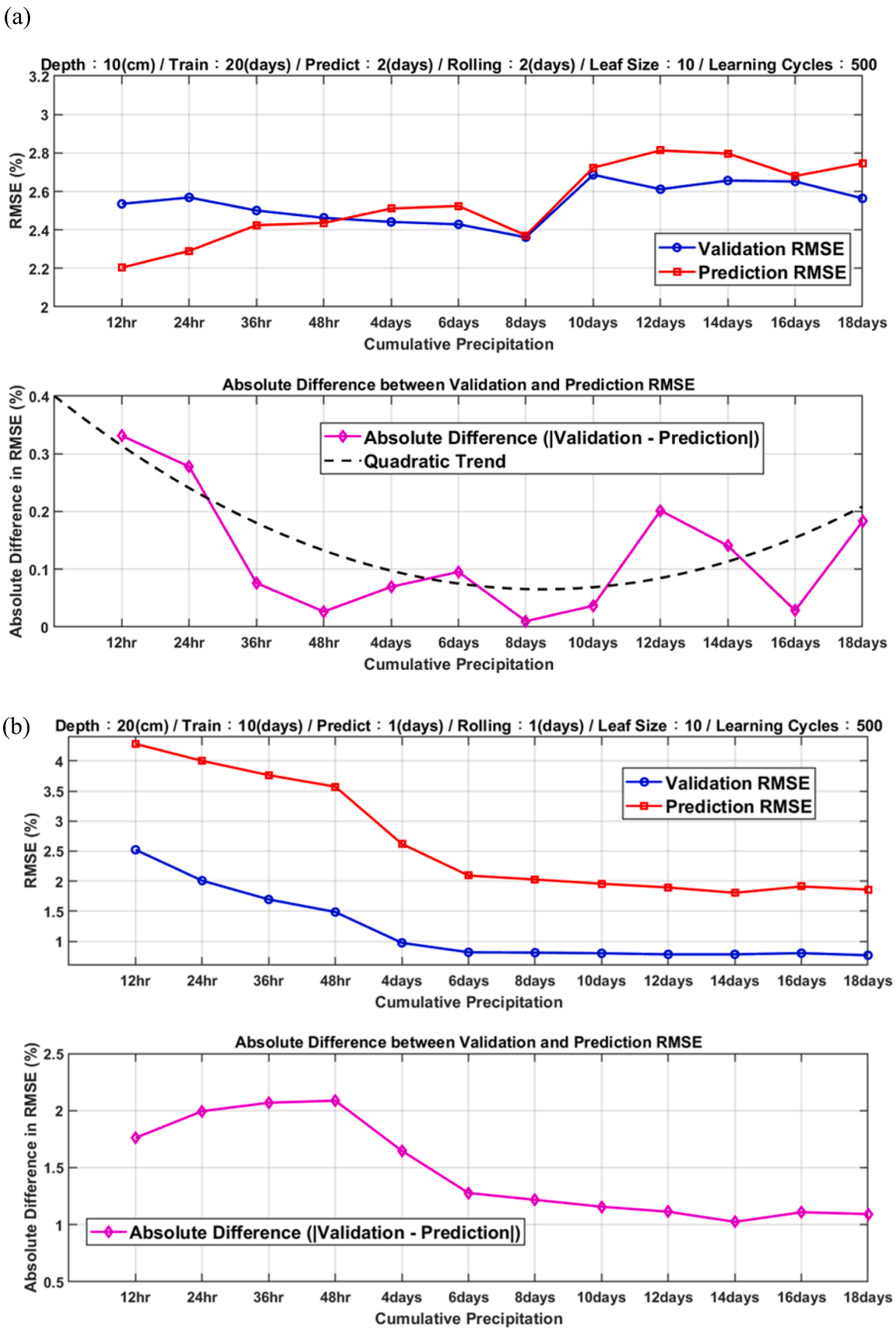


Fig. 14. Validation and prediction errors (upper panel) and differences (lower panel) in the root mean square error (RMSE) of (a) 10 and (b) 20 cm soil water content predictions under 12 combinations of cumulative rainfall by days. The quadratic curve is also shown in the lower panel of (a) the 10 cm soil water content. The labels of horizontal axes represent the longest cumulative days in every combination in Fig. 4.

for daily RZSM interpolation and extrapolation and compared the performance of RF with that of process-based models. They identified the potential of RF in data-poor areas, although they noted challenges with extreme condition prediction owing to infrequent sampling. The R^2 of RF and HRDRUS 1D model predictions were 0.75–0.97 and 0.73–0.84, while the RMSE of the two methods were 0.98 %–6.21 % and 2.23–3.53 %, respectively.

However, this study particularly focused on using precipitation data to predict SWC at two distinct depths (10 and 20 cm) in a paddy field in Taiwan. Our forecasting results showed an R^2 and RMSE of 0.5–0.9 and 2.0–2.4 %. Unlike Carranza et al. (2021), this study highlighted short-term (1–2 day) hourly predictions and focused on cumulative rainfall to achieve a balanced accuracy and computational efficiency performance. Our study optimized RF performance by adjusting the model parameters (leaf size, training duration, and cumulative rainy days). It thoroughly evaluated cumulative rainfall and its effect on prediction accuracy by examining three different cumulative rainfall cases to assess the computational efficiency and prediction quality. Unlike many other studies (Gorthi and Dou, 2011; Prasad et al., 2018; Pekel, 2020; Dubois et al., 2021; Wu et al., 2024) that have used a wide range of input variables, this study focused on precipitation-related data. Our study makes a novel contribution by focusing solely on precipitation data for SWC prediction, thereby reducing the need for other environmental and soil-related inputs. This approach broadens the applicability of RF to regions where only precipitation data are accessible, thereby making it an efficient tool for SWC prediction in various agricultural settings.

Bertalan et al. (2022) focused on spatial SWC heterogeneity and the effectiveness of UAV-based cameras combined with ML for high-resolution SWC mapping, whereas our study underscored a simple precipitation-only RF model for SWC prediction. Bertalan et al. (2022) provided high-accuracy models useful in precision agriculture with data-rich imaging tools and showed that RF is the most effective ML model for their data type, with an R^2 of 0.97. This study provides a practical solution tailored to data-sparse environments, enhancing seasonal SWC prediction and aiding irrigation planning. Both approaches illustrate the utility of the RF for SWC prediction, although in contexts that vary significantly in terms of data availability and application.

Other studies that predicted SWC based on RF were also compared with the results obtained in our study. Guan et al. (2022) obtained an MAE of approximately 1 % using a UAV-acquired vegetation index and multispectral data. In contrast, the prediction was slightly more accurate for the entire period but less accurate during the rainy season than theirs. The performance of image-based RF prediction may vary with soil types, achieving an R^2 of 0.64–0.93 and RMSE of 1.98–5.01 % (Liu et al., 2023). The RMSE of the RF prediction based on satellite data was also small, reaching 2 % in the study by Jia et al. (2020) and 3.3–7.5 % in the study by Chaudhary et al., (2022).

For studies using SWC and other meteorological variables as predictors, the performance of RF and the best model CNN-LSTM achieved an R^2 of 0.88–0.92 and 0.95–0.98, respectively (Alibabaei et al., 2021). Spatial SWC prediction based on meteorology, topography, and soil properties by Nie et al.(2016) showed R^2 of up to 0.46 and 0.73 and RMSE values of 10.76 % and 7.52 % for RF and SVM, respectively. Moreover, studies forecasting SWC rather than only nowcasting were compared with a 1–2-day lead time in this study. Prasad et al. (2018) showed hydrometeorological-variable-derived monthly SWC predictions with an RMSE of 0.1–5.2 % and 0.1–1.9 %, MAE of 0.1–3.5 % and 0.1–1.4 %, MAPE of 0.8–21.98 % and 0.63–9.85 % based on RF and ANN-based models, respectively. Their RF performance was comparable to our results, showing an RMSE of 2.0–2.4 %, MAE of 0.6–6.6 %, and MAPE of 5.1–25.2 %. As in our study, predicting SWC one day in advance led to an R^2 above 0.92, regardless of the method used (RF, SVM, or NN) (Dubois et al., 2021). In addition, their results for shallower soil depths had pronounced errors compared to deeper depths, owing to more significant water dynamics near the surface. This is similar to the cases of our 10 cm SWC prediction and the RF model's performance during the rainy season, with sudden increases in SWC caused by extreme rainfall that are more difficult to predict than the progressive drying process. Another reason for the 10 cm SWC errors is that non-precipitation (NP) variables directly influence the surface more than the 20 cm SWC, which was not used as an input in this study. This argument is supported by Wu et al. (2024), who found that the summary contribution of NP parameters, including wind speed, solar radiation, humidity, pressure, and air temperature, is crucial for good SWC estimates.

Based on this comparison, physical-based or deep learning models may provide more accurate SWC nowcasting or forecasting than RF. Furthermore, increasing the input information favors SWC forecasting. However, our study showed a comparable and reasonable performance of RF in SWC prediction based on easily obtained accumulated rainfall information. The results showed that the RF performs better in cases where the few-hour SWC variations were within 30 %, which is essential for agricultural drought and water resource management (Rani et al., 2022). This is also the case in Taiwan. Limitations in the accuracy of our methods exist, but this study made an innovative attempt and achieved a breakthrough. In the future, studies are suggested to improve parameter optimization using Grid Search or combining manual and automatic models and forecasting time to enhance accuracy and applicability. Future studies could explore other approaches, such as deep learning, when there is an abundant dataset available.

The methodology of machine learning itself is based on profound knowledge. However, this study focuses on the application of machine learning. The SWC prediction in this study is based merely on precipitation data, which is easily available compared to other potential variables influencing SWC. RF algorithm is selected in this study for SWC prediction due to its practical advantages and strong performance, especially in scenarios with limited data where interpretability is important. This study has demonstrated that RF can effectively handle non-linear relationships without requiring complex temporal sequence modeling. As a preliminary result of comparing the predictive performance of different machine learning algorithms, a Long Short-Term Memory (LSTM) network model was tested but found its performance, with an MAE of 1.5814 % under exactly the same conditions as Fig. 7(b), was not as strong as the results achieved with RF. While LSTM networks are powerful tools for time-series data, they typically require large datasets to avoid overfitting and perform optimally. The relatively modest size and seasonal nature of the dataset likely limited LSTM's ability to generalize effectively. The primary objective of this study is to develop a straightforward, accessible model for SWC prediction that would be applicable in data-limited scenarios. RF's robustness and ease of interpretability made it an ideal choice for this purpose. However, future studies could explore LSTM and other deep-learning approaches as more extensive datasets become available. That would even allow a more benchmarking comparison, helping to identify the specific conditions under which each approach may be most advantageous for SWC prediction.

5. Conclusions

This study employed RF models to predict SWC at depths of 10 and 20 cm beneath the surface of a fallow paddy field at the Taiwan ARI in central Taiwan from January 19, 2022, to April 18, 2023, by inputting various precipitation-related data, dry hours, and daily hours. The optimal settings for leaf size and learning cycles were determined manually, followed by assessments of training durations and different combinations of cumulative rainfall per day to refine the parameters for precise SWC prediction. Accum. P., identified as the most sensitive variable, was specifically tested to evaluate its impact on prediction accuracy. The findings revealed that excluding other variables has a minimal effect because RF algorithms inherently extract pertinent information from complex input datasets. The optimal cumulative rainfall per day was approximately eight days for both the 10 cm and 20 cm SWC

predictions. The results also indicated that SWC predictions were less accurate during the rainy season than during the dry periods, underscoring the challenges in SWC prediction under varying weather conditions.

Moreover, analyzing the effect of various accumulated rainfall inputs indicated that the seven strategically selected sets yielded MAE errors comparable to those of all 16 sets, providing computational efficiency and avoiding potential overfitting. In contrast, the validation and prediction RMSE increased with more extended cumulative rainfall by day, signaling the onset of overfitting. Consistent with prior findings, the optimal cumulative rainfall per day was approximately eight. In conclusion, this study demonstrated the efficacy of RF models in predicting SWC solely using precipitation-related data, establishing optimal parameters for cumulative rainfall by day and training duration. This focused approach simplifies the model and achieves reasonable performance comparable to that of previous studies, enhancing its applicability in scenarios with limited data availability. Future studies are suggested to improve parameter optimization and forecasting time. Additionally, deep learning can be used to incorporate more available input data, thereby enhancing accuracy and applicability. This study highlights the role of accumulated precipitation as an input variable and provides insights for enhancing the accuracy of SWC prediction across different seasons. It provides practical guidelines for employing RF models in SWC prediction.

Data Availability Statement: The authors will supply the relevant data in response to reasonable requests.

CRediT authorship contribution statement

Pei-Yuan Chen: Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Investigation, Funding acquisition, Conceptualization. **Chien-Chih Chen:** Writing – review & editing, Visualization, Software, Methodology, Investigation, Funding acquisition. **Chu Kang:** Writing – original draft, Visualization, Software. **Jia-Wei Liu:** Writing – review & editing, Methodology, Formal analysis, Data curation, Conceptualization. **Yi-Heng Li:** Writing – review & editing, Supervision, Resources, Project administration.

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.

Acknowledgments

This work was partially supported by Taiwan's National Science and Technology Council (NSTC) under Contract No. NSTC 113-2621-M-008-005, NSTC 112-2313-B-008-001, MOST 111-2634-F-008-001, and MOST 110-2634-F-008-008.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2024.109802.

Data availability

Data will be made available on request.

References

Abowarda, A.S., Bai, L., Zhang, C., Long, D., Li, X., Huang, Q., Sun, Z., 2021. Generating surface soil moisture at 30 m spatial resolution using both data fusion and machine learning toward better water resources management at the field scale. Remote Sens. Environ. 255, 112301.

Alibabaei, K., Gaspar, P.D., Lima, T.M., 2021. Modeling soil water content and reference evapotranspiration from climate data using deep learning method. Appl. Sci. 11 (11), 5029.

Bertalan, L., Holb, I., Pataki, A., Négyesi, G., Szabó, G., Szalóki, A.K., Szabó, S., 2022. UAV-based multispectral and thermal cameras to predict soil water content–a machine learning approach. Comput. Electron. Agric. 200, 107262.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Cart. Classification and regression trees.

Bucci, M.A., Semeraro, O., Chibbaro, S., Allauzen, A., Mathelin, L., 2020. When deep learning meets ergodic theory. In 73rd Annual APS/DFD Meeting.

Cai, Y., Zheng, W., Zhang, X., Zhangzhong, L., Xue, X., 2019. Research on soil moisture prediction model based on deep learning. PLoS One 14 (4), e0214508.

Carranza, C., Nolet, C., Pezij, M., van der Ploeg, M., 2021. Root zone soil moisture estimation with Random Forest. J. Hydrol. 593, 125840.

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7 (3), 1247–1250.

Chaudhary, S.K., Srivastava, P.K., Gupta, D.K., Kumar, P., Prasad, R., Pandey, D.K., Gupta, M., 2022. Machine learning algorithms for soil moisture estimation using Sentinel-1: model development and implementation. Adv. Space Res. 69 (4), 1799–1812.

Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput. Sci. 7, e623.

Dubois, A., Teytaud, F., Verel, S., 2021. Short-term soil moisture forecasts for potato crop farming: a machine learning approach. Comput. Electron. Agric. 180, 105902.

Eitzinger, J., Trnka, M., Hösch, J., Žalud, Z., Dubrovský, M., 2004. Comparison of CERES, WOFOST and SWAP models in simulating soil water content during growing season under different soil conditions. Ecol. Model. 171 (3), 223–246.

Emami, S., Rezaverdinejad, V., Dehghanisanij, H., Emami, H., Elbeltagi, A., 2024. Data mining predictive algorithms for estimating soil water content. Soft. Comput. 28 (6), 4915–4931.

Gill, M.K., Asefa, T., Kemblowski, M.W., McKee, M., 2006. Soil moisture prediction using support vector machines 1. JAWRA J. American Water Resour. Association 42 (4), 1033–1046.

Gorthi, S., Dou, H. (2011). Prediction models for the estimation of soil moisture content. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (Vol. 54808, pp. 945-953).

Guan, Y., Grote, K., Schott, J., Leverett, K., 2022. Prediction of soil water content and electrical conductivity using random forest methods with UAV multispectral and ground-coupled geophysical data. Remote Sensing 14 (4), 1023.

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction Vol. 2, 1–758.

Holsten, A., Vetter, T., Vohland, K., Krysanova, V., 2009. Impact of climate change on soil moisture dynamics in Brandenburg with a focus on nature conservation areas. Ecol. Model. 220 (17), 2076–2087.

Jia, Y., Jin, S., Savi, P., Yan, Q., Li, W., 2020. Modeling and theoretical analysis of GNSS-R soil moisture retrieval based on the random forest and support vector machine learning approach. Remote Sens. (Basel) 12 (22), 3679.

Karandish, F., Šimůnek, J., 2016. A comparison of numerical and machine-learning modeling of soil water content with limited input data. J. Hydrol. 543, 892–909.

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.

Lewis, C. D. (1982). Industrial and business forecasting methods: A practical guide to exponential smoothing and curve fitting. (No Title).

Liu, G., Tian, S., Xu, G., Zhang, C., Cai, M., 2023. Combination of effective color information and machine learning for rapid prediction of soil water content. J. Rock Mech. Geotech. Eng. 15 (9), 2441–2457.

Morellos, A., Pantazi, X.E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., Mouazen, A.M., 2016. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. Biosyst. Eng. 152, 104–116.

Nath, K., Nayak, P.C., Kasiviswanathan, K.S., 2024. Soil volumetric water content prediction using unique hybrid deep learning algorithm. Neural Comput. & Applic. 1–23.

Neitsch, S.L., Arnold, J.G., Kiniry, J.R., Williams, J.R., 2011. Soil and water assessment tool theoretical documentation version 2009. Texas Water Resources Institute.

Nie, H., Yang, L., Li, X., Ren, L., Xu, J., Feng, Y. (2018). Spatial prediction of soil moisture content in winter wheat based on machine learning model. In 2018 26th *International Conference on Geoinformatics* (pp. 1-6). IEEE.

Pekel, E., 2020. Estimation of soil moisture using decision tree regression. Theor. Appl. Climatol. 139 (3), 1111–1119.

Prakash, S., Sharma, A., Sahu, S. S. (2018, April). Soil moisture prediction using machine learning. In 2018 *Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 1-6). IEEE.

Prasad, R., Deo, R.C., Li, Y., Maraseni, T., 2018. Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydro-meteorological predictors. Soil Tillage Res. 181, 63–81.

Rani, A., Kumar, N., Kumar, J., Sinha, N.K., 2022. Machine learning for soil moisture assessment. In: *Deep Learning for Sustainable Agriculture*. Academic Press, pp. 143–168.

Sutton, C.D., 2005. Classification and regression trees, bagging, and boosting. Handbook of Statist. 24, 303–329.

Tan, X., Shao, D., Liu, H., 2014. Simulating soil water regime in lowland paddy fields under different water managements using HYDRUS-1D. Agric. Water Manag. 132, 69–78.

Tmušić, G., Manfreda, S., Aasen, H., James, M.R., Gonçalves, G., Ben-Dor, E., McCabe, M. F., 2020. Current practices in UAS-based environmental monitoring. Remote Sensing 12 (6), 1001.

Tramblay, Y., Quintana Seguí, P., 2022. Estimating soil moisture conditions for drought monitoring with random forests and a simple soil moisture accounting scheme. Nat. Hazards Earth Syst. Sci. 22 (4), 1325–1334.

Vyas, A., Bandyopadhyay, S. (2020). Dynamic structure learning through graph neural network for forecasting soil moisture in precision agriculture. *arXiv preprint arXiv: 2012.03506*.

Wu, T.H., Chen, P.Y., Chen, C.C., Chung, M.J., Ye, Z.K., Li, M.H., 2024. Classification and Regression Tree (CART)-based estimation of soil water content based on meteorological inputs and explorations of hydrodynamics behind. Agric. Water Manag. 299, 108869.

Wu, W., Wang, X., Xie, D., Liu, H. (2008). Soil water content forecasting by support vector machine in purple hilly region. In Computer And Computing Technologies In Agriculture, Volume I: First IFIP TC 12 *International Conference on Computer and Computing Technologies in Agriculture (CCTA 2007)*, Wuyishan, China, August 18-20, 2007 1 (pp. 223-230). Springer US.

Yu, J., Tang, S., Zhangzhong, L., Zheng, W., Wang, L., Wong, A., Xu, L., 2020. A deep learning approach for multi-depth soil water content prediction in summer maize growth period. IEEE Access 8, 199097–199110.

Zhang, H., Lu, H., Nayak, A., 2020. Periodic time series data analysis by deep learning methodology. IEEE Access 8, 223078–223088.