RESEARCH ARTICLE https://doi.org/10.7235/HORT.20240001

Data Mining Approach to Predicting Soil Moisture Based on Meteorological Factors and Flow Rates

Su Hoon Choi[1†], Sang-Hyun Lee[2,3,4†], Ung Yang[2], and Min Soo Kim[5*]

[1]Department of Mathematics and Statistics, Chonnam National University, Gwangju 61186, Korea
[2]Asian Pear Research Institute, Chonnam National University, Gwangju 61186, Korea
[3]Department of Horticulture, College of Agriculture and Life Sciences, Chonnam National University, Gwangju 61186, Korea
[4]Interdisciplinary Program in IT-Bio Convergence System, Chonnam National University, Gwangju 61186, Korea
[5]Department of Statistics, Chonnam National University, Gwangju 61186, Korea

*Corresponding author: kimms@chonnam.ac.kr

†These authors contributed equally to this work.

Abstract

Accurate predictions of the soil moisture, a major limitation related to crop growth, are essential for effective irrigation planning, yield predictions, and proper water resource management. Current attempts to predict soil moisture levels have relied on historical soil moisture data; however, obtaining this type of data can be challenging for farmers engaged in outdoor cultivation. Therefore, this study aimed to predict current soil moisture contents based only on meteorological and flow rate data without previous soil moisture information. To predict the soil moisture, data mining approaches, in this case random forest (RF), support vector regression (SVR), and deep neural network (DNN), were employed. Through the Granger causality test, explanatory variables were determined at different time lags extending to three hours as model inputs for soil moisture predictions. The predictive performance of the models was found to be improved when all meteorological and flow rate data ranging from the previous to the current time could be used as opposed to only data from the current time. The results obtained for the test set showed that the best performance was achieved by the DNN model when it applied explanatory variables from "t" to "t-3" time points with a RMSE of 1.542 and a R^2 value of 0.580. These results can enable real-time soil moisture data monitoring for farmers who currently lack access to soil databases. The constructed model is expected to serve as a framework for future studies that consider various environmental factors, such as soil characteristics, topography, and vegetation patterns.

Additional key words: deep neural network, granger causality test, random forest, support vector regression, time lags

Received: June 21, 2023
Revised: September 8, 2023
Accepted: October 5, 2023

HORTICULTURAL SCIENCE and TECHNOLOGY
42(1):1-14, 2024
URL: http://www.hst-j.org

pISSN : 1226-8763
eISSN : 2465-8588

This work was supported by a grant from the Korea Institute of Planning & Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) through the Open Field Smart Agriculture Technology Short-Term Advancement Program, funded by the Ministry of Agriculture, Food and Rural Affairs (MAFRA) (No.322034-3), Republic of Korea.

Introduction

The content and distribution of soil moisture, which plays a critical role in ensuring optimal crop yields, can influence a wide range of processes, from soil microbial activity to nutrient cycling (Aerts, 1997) and groundwater recharging (Rushton et al., 2006). Soil moisture can be a significant limiting

factor for crop growth, considering the increasingly severe droughts in recent years and the anticipated future climate changes (Deutsch et al., 2010). Waterlogging stress causes hypoxia or anoxia in plant roots, reducing nutrient uptake, crop growth, and yields (Gomathi et al., 2015). Meanwhile, drought stress caused by delayed or insufficient irrigation during crop growth stages can reduce the potential yield and quality of crops (Ntukamazina et al., 2017; Shi et al., 2022). Therefore, accurate predictions of soil moisture conditions can aid in planning irrigation schedules, predicting yields, and in proper water resource management (Prasad et al., 2018; Kwon et al., 2020; Nam et al., 2023).

The properties of the soil and the topography are known to impact the soil moisture distribution considerably (Zhu and Lin, 2011). Soil moisture spatial patterns are affected by variations in certain soil parameters, such as the hydraulic conductivity (Garcia et al., 2014), soil texture, and soil depth (Xu et al., 2008; Takagi and Lin, 2011). Additionally, soil moisture levels vary spatially due to topographic variability (Grayson et al., 1997; Western et al., 1999), soil water redistribution (Williams et al., 2003), and vegetation patterns (Qiu et al., 2001; Hupet and Vanclooster, 2002).

Soil moisture predictions are closely related to meteorological factors, and prediction models can be divided into two categories: equation-based physical models and machine-learning models (Lamorski et al., 2013). To assess the soil moisture content, equation-based models rely on a variety of factors, including precipitation, runoff, and irrigation (Van Dam et al., 1997; Šimůnek and Van Genuchten, 2008; Neitsch et al., 2011). If field measurement data are available, no theoretical calibration is required because the model's parameters relate to the physical quantities. However, it is challenging to establish an ideal mathematical model capable of predicting soil moisture as doing so involves complex structural properties and meteorological factors.

Data collection has become easier since the development of wireless sensor networks and agricultural technologies; thus, data analysis has started to highlight the complex relationships among soil moisture and meteorological variables, which has prompted research on the use of machine-learning algorithms to predict soil moisture levels. Machine learning models have the advantage of being highly predictive while not requiring physical soil properties as model inputs (Lamorski et al., 2013). Various machine learning models have been employed to predict soil moisture. In particular, a support vector machine and a tree-based algorithm have been used to make soil moisture predictions with meteorological variables, such as the air temperature, relative humidity, and degree of insolation (Gill et al., 2006). Machine learning models have the disadvantage of being more difficult to interpret intuitively than equation-based models, though they have been shown to offer better predictive power.

Recently, there have been numerous attempts to predict soil moisture levels using neural network-based models due to training via algorithm optimization (Acharya et al., 2021). In a comparison with results from machine learning and neural network-based models, Cai et al. (2019) attempted to improve the results of Beijing's soil moisture prediction model through deep learning techniques. Additionally, Ji et al. (2017) improved the neural network activation function based on earlier work by Hou et al. (2016), who used artificial neural networks to predict soil moisture contents at various depths with multi-input meteorological data. Furthermore, soil moisture predictions are being attempted through hybrid models based on various neural network models, such as the convolution neural network and long short-term memory models (Yu et al., 2020, 2021). Gao et al. (2022) constructed a deep-LSTM model capable of rapidly analyzing large amounts of time-series data consisting of ten-minute units over two years. These studies have shown that soil moisture predictions using neural network−based models can be effective and feasible. Suitable data fitting and generalization capabilities have demonstrated high accuracy in predicting soil moisture values and can provide effective rationales for irrigation management systems (Filipović et al., 2022).

Numerous studies have demonstrated that soil moisture content levels from previous time lags significantly affect soil moisture predictions (as summarized in Table 1). Although most studies utilize soil moisture levels from previous time points to predict future soil moisture levels, it is challenging to obtain soil moisture data for many farms engaged in outdoor cultivation. The aim of this study was to develop a machine-learning algorithm, such as RF, SVR, and DNN, to predict the current soil moisture using only meteorological and flow rate data. Before fitting the models to the meteorological and flow rate data, a Granger causality test was performed to determine the time lag over which these variables affect the soil moisture. The predictive performance of each model, fitted with explanatory variables with various time lags, was assessed based on the RMSE and R^2 values. The overall process for soil moisture prediction and the explanatory variables used in the prediction model are well depicted in Figs. 1 and 2, respectively. A predictive model based only on meteorological and flow rate data will be useful to develop an integrated management system for efficient irrigation in outdoor cultivation systems, for which soil data are limited.

Table 1. Preliminary Studies of Soil Moisture Predictions using Machine Learning and Neural Network-Based Models

Authors	Inputs	Methods[z]			Unit of "t"
		Statistical model	Machine learning model	Neural network model	
Gill et al. (2006)	Air temperature, relative humidity, solar radiation, soil temperature, and soil moisture		SVM	ANN	days
Matei et al. (2017)	air temperature, soil temperature, precipitation, and soil moisture	MLR, LR	k-NN, SVM, FLM, DT, RF	NN	days
Adeyemi et al. (2018)	Wind speed, rainfall, air temperature, net radiation, relative humidity, and soil moisture			FFNN, LSTM	days
Cai et al. (2019)	Air temperature, air pressure, relative humidity, wind speed, surface temperature, precipitation, and soil moisture	MLR	SVM	ANN, AGNN, DNNR	days
Yu et al. (2020)	Air temperature, ground surface temperature, hours of sunshine, cumulative rainfall, wind speed, air humidity, and soil moisture		SVR, RF	MLP, DNNR, CNN-LSTM, ResBiLSTM	days
Acharya et al. (2021)	Air temperature, rainfall, wind direction, and soil moisture	MLR	CART, RF, BRT, SVR	ANN	days
Yu et al. (2021)	Air temperature, air humidity, rainfall, wind speed, net radiation, and soil moisture			CNN, GRU, Hybrid CNN-GRU	days
Gao et al. (2022)	Air temperature, relative humidity, wind speed, precipitation, and soil moisture			Deep-LSTM, ENN, GRNN	10 minutes
Filipović et al. (2022)	Air temperature, precipitation, vapor pressure deficit, and soil moisture	ARIMA	RF	LSTM	days
Present study	Air temperature, relative humidity, wind speed, air pressure, precipitation, and flow rate		RF, SVR	DNN	hours

[z]SVM: support vector machine; ANN: artificial neural network; MLR: multiple linear regression; LR: logistic regression; k-NN: k-nearest neighbors; FLM: fast large margin; DT: decision tree; RF: random forest; NN: neural network; SVR: support vector regression; FFNN: feed-forward neural network; LSTM: long short-term memory; AGNN: adaptive genetic neural network; DNNR: deep neural network regression; MLP: multilayer perceptron; CART: classification and regression trees; BRT: boosted regression trees; CNN: convolution neural network; GRU: gated recurrent unit; ENN: Elman neural network; GRNN: generalized regression neural network; ARIMA: auto-regressive integrated moving average.

Materials and Methods

Study Site and Data Collection

The study was carried out at a pear orchard in Wanggok-myeon, Naju-si, Korea (34°58'34.1"N 126°42'12.7"E; 25 – 30 m above sea level; 0.83ha). Data on soil moisture were obtained from a Hydra Probe II soil sensor (SDI-12, Stevens Water Monitoring System Inc., OR, USA) installed at a depth of 20 cm located 10 cm from a representative dripper and 1 m from the trunk. The soil was composed of silt loam (depth 0 – 100 cm) and clay loam (depth below 100 cm), which are appropriate for the growth of pear trees. The study site was also equipped with a drip irrigation system to keep soil moisture at the proper level during the growing season. Flow rate sensors were used to identify irrigation events and the amounts of water provided.

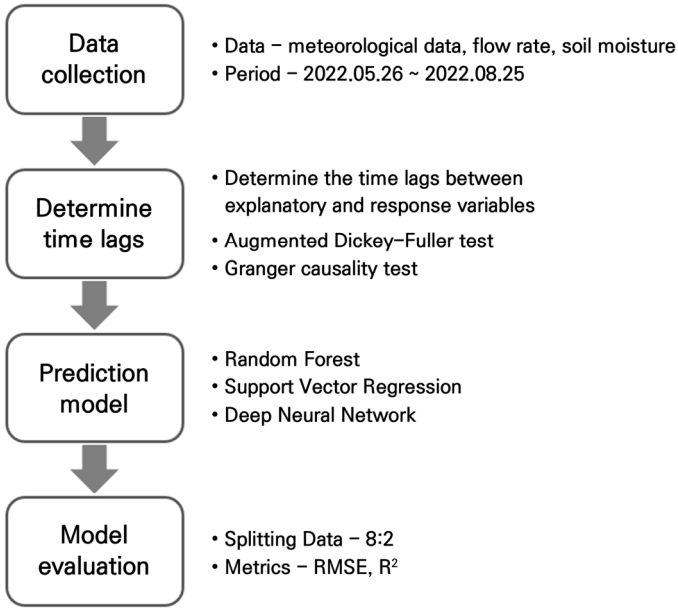

Fig. 1. Flowchart of the analysis process.

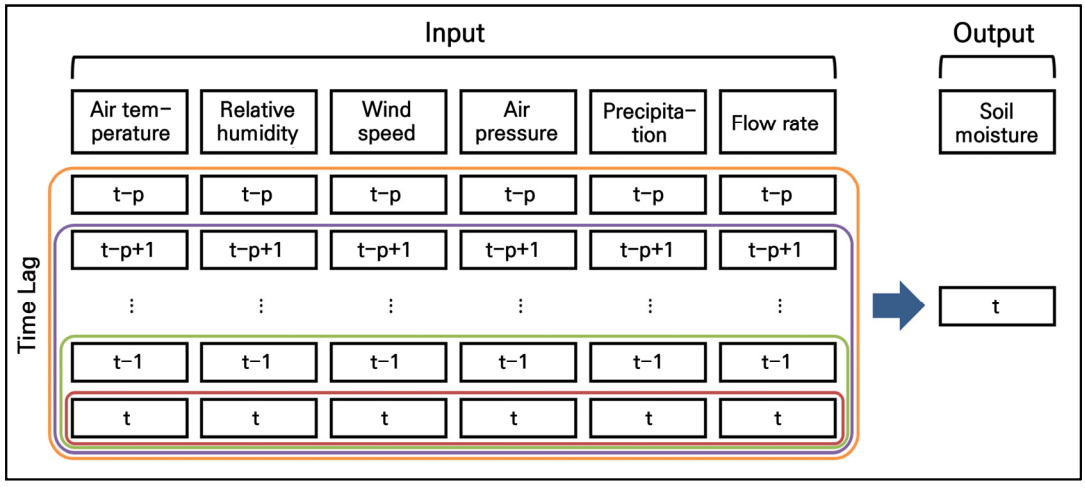

Fig. 2. Prediction model structure.

The data periods used in the analysis were chosen such that they coincided with the growing season for pear trees. The included data ranged from when the soil moisture sensor was installed to four weeks before harvest (from May 26 to August 25, 2022). This was due to the fact that irrigation was halted four weeks prior to harvest in order to reduce the amount of nitrogen that the pear trees absorbed and to raise the sugar content of the fruit. Meteorological data (air temperature, relative humidity, wind speed, air pressure, and precipitation) were collected hourly from an automatic weather station (AWS) located approximately 7 km southwest of the pear orchard.

Methodology

Identifying the soil moisture content in real time was important when determining the appropriate irrigation timing and amount of water required for crop growth. Before using the meteorological and flow rate data in the prediction model, the Granger causality test was utilized to determine the time lags over which these variables affect the soil moisture. Because a prerequisite of the Granger causality test is a stationary time series, an augmented Dickey−Fuller test was conducted to assess the stationarity of the data.

Augmented Dickey-Fuller Test

Dickey and Fuller (1979) devised their test to determine whether a time series is a stationary or a unit root process from the autoregressive model using the following assumptions (Dickey and Fuller, 1979):

$$H_0 : \alpha = 1,$$
$$H_1 : |\alpha| < 1,$$
$$Y_t = \alpha Y_{t-1} + \varepsilon_t.$$

Expanding this with respect to the constant term and the non-probabilistic trend, we have

$$Y_t = c + \beta t + \alpha Y_{t-1} + \phi \Delta Y_{t-1} + \varepsilon_t,$$

where c represents the level of the time-series data as a constant term and α represents the trend. Each parameter is estimated from the regression model to test how likely α is to have a unit root. If the time series satisfies the stationary criterion, the average approaches a constant value, meaning that the influence of Y_{t-1} is weakened.

The augmented Dickey−Fuller (ADF) test is an augmented and extended high-order regression version of the original Dickey−Fuller test, which added a difference term in the form of p-lag, representing a verification method that strengthens the power of the existing test in view of the autocorrelation of the error term by incorporating the tendency of the difference up to $t-p$ (Mushtaq, 2011):

$$Y_t = c + \beta t + \alpha Y_{t-1} + \sum_{j=1}^{p} \phi_j \Delta Y_{t-j} + \varepsilon_t.$$

Granger Causality Test

The Granger causality test works differently from exploratory statistics, whereby the confirmation statistics are based on data with a minimal prior assumption of causality (Bressler and Seth, 2011). Using the unique information of past lags in any time series, this test can determine the presence of a causal relationship based on the degree of statistical significance and can serve as a more helpful predictor than linear predictions when other time series cannot be used. However, while the cause acts to produce the result, it is clear that any current result could not have caused past events. Based on this logic, the Granger causality test evaluates the null hypothesis, whereby one variable does not help predict another (Granger, 1969).

$$H_0 : All\ b_j = 0 \text{ or } c_j = 0$$

$$H_1 : not\ H_0$$

To determine the random vector Y_t, a series of linear equation structures described in present and past terms can be expressed as follows:

$$A_0 Y_t = \sum_{j=1}^{m} A_j Y_{t-j} + \varepsilon_t$$

$$Y_t = \sum_{j=1}^{m} a_j Y_{t-j} + \sum_{j=1}^{m} b_j X_{t-j} + \varepsilon_t^{'},$$

$$X_t = \sum_{j=1}^{m} c_j Y_{t-j} + \sum_{j=1}^{m} d_j X_{t-j} + \varepsilon_t^{''},$$

where ε_t represents a white-noise random vector. Based on the Granger causality test results, the lag of the meteorological variables used to predict the soil moisture can be determined and fitted to the prediction model. Random forest (RF) and support vector regression (SVR) are machine learning techniques frequently used in previous studies. The deep neural network (DNN), a neural network-based model, was also used as an analysis model. All analyses utilized Python (version 3.8.13) alongside packages such as Statsmodels (version 0.13.2) and Scikit-learn (version 1.1.2).

Random Forest

RF is an ensemble technique that combines multiple decision trees in a machine-learning algorithm. It utilizes a set of tree-structure learners consisting of input vectors and random vectors (Breiman, 2001). Random vectors are considered independent and are identically distributed for each tree, where they are used in the following two steps. Firstly, the data used for the independent individual decision tree were randomly restored and extracted through the bagging method, which improved the accuracy by learning several versions of the training sets that had been randomly restored and extracted using bootstrap techniques from the entire training set (Breiman, 1996). Secondly, the optimization process is repeated to find the best segmentation for a subset of randomly selected predictors instead of all predictors during the segmenting of each node in an individual tree. Hence, the decision trees form a weak correlation with each other, resulting

in lower overall variance and fewer prediction errors and making the RF robust against the overfitting problem. The goal of RF is to find a prediction function that minimizes the expected value of the prediction loss from the output vector. The loss function is generally a square error loss. Unlike categorical classifications that predict classes with weightless voting, the prediction function averages individual learners as follows (Cutler et al., 2012):

$$f(x) = \frac{1}{J}\sum_{j=1}^{j} h_j(x)\,.$$

Support Vector Regression

A support vector machine (SVM) maps learning data in the input space to a high-dimensional feature space and configures a hyperplane separated by a maximum margin to distinguish clearly between different types of data (Cortes and Vapnik, 1995). SVR, using the same principles used with SVM for regression problems, determines a function that can approximate future values within the decision boundary.

To solve the nonlinear regression problem, we matched the input value to a high-dimensional feature space and identified a linear function associated with the resulting value (Lu et al., 2009). SVR learns by using symmetric loss functions to penalize high and low mis-estimations equally. Therefore, a minimum radius $(-\epsilon, \epsilon)$ is symmetrically formed around the estimated function, and within this range all errors smaller than the threshold are ignored. This reduced sensitivity to noisy inputs makes the ϵ-insensitive region more powerful for the model. A general SVR estimation equation is as follows:

$$f(x) = (w \cdot \phi(x)) + b,$$

where ϕ represents a non-linear transformation from the R^n-dimensional to the high-dimensional space; the goal of SVR is to find vectors w and scalar b that minimize the regression risk R_{reg}. The ϵ-insensitive loss function is used as the cost function Γ (Wu et al., 2004):

$$R_{reg}(f) = \frac{1}{2}\|w\|^2 + C\sum_{i=0}^{l}\Gamma(f(x_i)-y_i)\,,$$

$$\Gamma(f(x)-y) = \begin{cases} \text{if } |f(x)-y| \geq \epsilon, & |f(x)-y|-\epsilon \\ otherwise\,, & 0 \end{cases}.$$

Deep Neural Network

A DNN is a collection of neurons with multiple hidden layers. Previously, LeCun et al. (1989) introduced a network of three hidden layers trained using error backpropagation. The DNN concept has recently begun to attract attention, with one hidden layer presenting better results in fully connected networks with fewer feature maps. In order to build a deep layer that is not expressed as a linear function, a non-linear activation function is required; functions such as Sigmoid and ReLU are often used. Multilayer neurons are learned by jointly implementing complex non-linear mapping and applying the weights of each of the neurons (Montavon et al., 2018).

The DNN serves as an efficient and powerful method for dealing with large-scale regression problems (Lu et al., 2018). It has the advantage of being able to express input variables in a non-linear combination, whereby a larger amount of data corresponds to better performance by the method.

The progress to the j-th node of the h-th hidden layer is expressed as follows, referring to the number of nodes in each h-th hidden layer N_h and the weight vector of each node $w_{i,j}$ ($w_{i,0}$ is the bias):

$$x_j^h = \sum_{i=1}^{N_{h-1}} f^h(w_{i,j}^h \, x_i^{h-1} + w_{0,j}^h) .$$

The training data pass through the linear or non-linear transformation layer in the feed-forward neural network, which consists of several hidden layers, and through the backpropagation process, while the weight parameter is aptly approximated from the loss function and the mapping function is improved to enhance the prediction performance (Qi et al., 2020).

Fitting Models

The training and test sets were randomly divided according to a ratio of 8:2 for the entire dataset. In the hyperparameter tuning process, the optimal hyperparameters were determined and analyzed using the grid search method through five-fold cross-validation among the candidates presented in Suppl. Tables 1 to 3 for each model. The root mean square error (RMSE) and R-squared (R^2) were used as the metrics to compare the prediction results. RMSE is an evaluation index often used to indicate the difference between actual and predicted values; the closer to 0 it is, the more accurate the prediction also is. R^2 is referred to as the coefficient of determination and represents the proportion of the variance of the dependent variable described by the linear regression model (Lewis-Beck and Lewis-Beck, 2015). R^2 has a value between 0 and 1, and the closer it is to 1, the closer the regression line is to a perfect correlation. RMSE and R^2 are calculated as follows:

$$RMSE = \sqrt{MSE} = \sqrt{E((Y - \hat{Y})^2)} ,$$

$$R^2 = 1 - \frac{\sum (Y - \hat{Y})^2}{\sum (Y - \overline{Y})^2} .$$

Results

This section shows the results of determining the lag between the explanatory variables and soil moisture levels through the Granger causality test as well as the prediction results from the machine learning models (RF, SVR, and DNN).

Exploratory Data Analysis

Before performing the principal analysis, an exploratory data analysis process was conducted to summarize and visualize the primary data characteristics. First, up to 22 missing values for each variable were confirmed in the AWS

meteorological data. Given that the number of missing values represented only 1% of the total number of data values, 2,208, linear interpolation was used to replace the missing values with the averages of the adjacent values. Table 2 provides a summary of each variable after the missing values were replaced. The data graphs for all variables by time are presented in Suppl. Fig. 1.

Augmented Dickey-Fuller Test

Table 3 shows the results of the ADF test, which was conducted to confirm the stationary of the variables. It was found that all variables, with the exception of the air temperature, were stationary time series at the 5% significance level. If the time-series data do not satisfy stationarity, the Granger causality test should be performed after conversion to a stationary time series through mathematical manipulation such as differencing transformation. However, the transformation of non-stationary time series into stationary sequences may cause a loss of information from the raw data. The stationarity assumption of all variables was satisfied by the ADF test at the 10% significance level; therefore, the Granger causality test was conducted without transforming the raw data.

Granger Causality Test

The results of the Granger causality test determined the lag interval between the variables, with a maximum of five lag intervals, as shown below (Table 4). The air temperature, air pressure, and flow rate variables were found to be causally significant at the 5% level, although the remaining variables were not significant. Regarding the final selected lag, the lag with the smallest *p*-value resulting from each tested variable was selected, irrespective of significance. Based on the smallest *p*-value in the Granger causality results, the maximum "t-3" time lag was confirmed between the explanatory variables (for the relative humidity and flow rate) and the soil moisture. Therefore, a predictive analysis was conducted using explanatory variables ranging from the current time (i.e., "t") to the "t-3" time lag.

Table 2. Summary Table of Variables

Variable	Air temperature (°C)	Relative humidity (%)	Wind speed (m/s)	Air pressure (hPa)	Precipitation (mm)	Flow rate (*l*/hour)	Soil moisture (%)
Minimum	11.3	22.5	0.0	996.4	0.0	0.0	38.9
Mean	25.1	83.8	2.0	1,006.5	0.2	38.0	46.5
Median	25.6	87.0	1.7	1,006.6	0.0	0.0	47.0
Maximum	34.2	99.9	8.3	1,014.6	24.0	2,730.0	52.0

Table 3. Augmented Dickey-Fuller Test Results

Variable	Air temperature	Relative humidity	Wind speed	Air pressure	Precipitation	Flow rate	Soil moisture
AIC[z]	−2.801	−6.123	−4.869	−6.252	−13.755	−12.578	−4.359
p-value	0.058	< 0.001	0.014	< 0.001	< 0.001	< 0.001	0.001

[z]AIC: Akaike information criterion.

Prediction Results

The performances of each of the models fitted with the meteorological variables of the various time lags ranging from "t" to "t-3" are shown in Table 5. When only the explanatory variables (meteorological data and flow rate) at the current time were used to predict soil moisture, RMSE had a range of 1.838 − 1.894 while R^2 had a range of 0.367 − 0.404 in the test set. On the other hand, when all explanatory variables were employed, ranging from the current time to the "t-3" time lag, RMSE was 1.542 − 1.695 and R^2 was 0.493 − 0.580 in the test set. The predictive performance of the models used in this study was found to be improved when all meteorological and flow rate data ranging from the previous to the current time were used rather than only data from the current time. The DNN model showed less bias and produced stable results overall compared to the RF and SVR methods. As a result, the DNN model showed the best performance using the time points of the input variable from "t" to "t-3", with the RMSE equal to 1.542 and a R^2 value of 0.580 in the test set.

Table 4. Granger Causality Test Results[z]

Time Lag	Air temperature	Relative humidity	Wind speed	Air Pressure	Precipitation	Flow rate
1	6.866 (0.009)	2.195 (0.139)	0.001 (0.980)	0.465 (0.495)	2.381 **(0.123)**	49.594 (< 0.001)
2	4.967 **(0.007)**	1.056 (0.348)	0.563 **(0.570)**	4.564 **(0.011)**	0.408 (0.665)	78.464 (< 0.001)
3	3.515 (0.015)	1.989 **(0.114)**	0.584 (0.625)	3.424 (0.017)	0.235 (0.872)	54.235 **(< 0.001)**
4	2.569 (0.036)	1.569 (0.180)	0.582 (0.676)	2.902 (0.021)	0.663 (0.618)	42.499 (< 0.001)
5	2.293 (0.043)	1.379 (0.229)	0.481 (0.791)	2.490 (0.029)	0.822 (0.534)	33.631 (< 0.001)
Final	2	3	2	2	1	3

[z]F statistics (*p*-value).

Table 5. Comparison of Models for Soil Moisture Predictions using Evaluation Metrics

Input[z]	Metrics	RF		SVR		DNN	
		Training set	Test set	Training set	Test set	Training set	Test set
(t)	RMSE	1.070	1.840	1.821	1.894	1.453	1.838
	R^2	0.798	0.403	0.414	0.367	0.627	0.404
(t, t-1)	RMSE	0.976	1.795	1.668	1.827	1.175	1.740
	R^2	0.832	0.431	0.509	0.415	0.756	0.465
(t, t-2)	RMSE	0.931	1.743	1.322	1.715	1.218	1.611
	R^2	0.847	0.463	0.692	0.481	0.738	0.542
(t, t-3)	RMSE	0.899	1.695	1.144	1.636	1.085	1.542
	R^2	0.857	0.493	0.769	0.527	0.792	0.580

[z](t, t-p) indicates the time points for the meteorological data and flow rate used as model inputs from "t" to "t-p" (where "t" is in hours).

Discussion

In this study, soil moisture prediction was performed using meteorological and flow rate data collected by KMA and from local orchards, respectively, from May 26 to August 25, 2022. For accurate predictions of soil moisture levels, the RF, SVR, and DNN were selected as the analysis models owing to their frequent use in previous studies (Gill et al., 2006; Adeyemi et al., 2018; Cai et al., 2019; Acharya et al., 2021). Before using meteorological and flow rate data in the models, a Granger causality test was conducted to determine the time lags over which these variables affect the soil moisture. Unlike previous studies, which subjectively determined the time points of the explanatory variables, this study provided a basis for determining the time point being used in the analysis through the Granger causality test.

It was confirmed that there were differences in the degree and time lags of the input variables of the soil moisture, while the air temperature, air pressure, and flow rate had a significant impact on the prediction of the soil moisture (Table 4). Precipitation and irrigation (flow rate in the analysis) were thought to have a significant impact on soil moisture in relation of the provision of water directly to the soil; however, no obvious causal relationship was found between precipitation and soil moisture. Because the change in soil moisture caused by irrigation is an increase in soil moisture through the water supply when the soil is "dry," there is a clear causal relationship between these two factors, and the flow rate can be a limiting factor with regard to soil moisture changes (Marshall et al., 2004). Precipitation, on the other hand, is a sporadic event that does not distinguish between "dry" and "wet" soil conditions. Because the soil moisture in arid or semi-arid regions is frequently low, it can be sensitive to even irregular precipitation (Wang et al., 2019). In contrast, the study site here has high annual rainfall levels (approximately 1,300 mm) such that the impact of rainfall on changes in soil moisture levels is likely to be relatively low. Due to a preceding rainfall event, the water absorbed by the soil can be imprinted in the soil moisture (McCabe et al., 2008; Dirmeyer et al., 2009), which may affect the soil moisture change during a subsequent rainfall event. If rainfall input exceeds the maximum permeability of the soil, soil water infiltration in the shallow layers may be limited, resulting in the shortest duration and the slowest permeating velocity at a shallow depth (Yan et al., 2021). These intricate factors may have prevented the moisture sensors buried in shallow soil layers (20 cm) from properly capturing changes in the soil moisture during rainfall events. As a result, the limitations of this study arising from the short data period as well as the complex heterogeneity of the rainfall-soil moisture relationship make it difficult to establish a causal relationship between the two factors.

The predictive performance of the models used in this study was found to improve when the meteorological data and flow rate from previous time points were also utilized instead of the data from the current time (Table 5). Considering the results of the test set, the DNN model showed the best performance using the time points of the input variable from "t" to "t-3", with a RMSE of 1.542 and R^2 value of 0.580.

In Gill et al. (2006), where the SVM model was employed to forecast soil moisture levels, the performance of the model dropped when only meteorological variables were used as the input compared to when both meteorological and soil moisture data were used. It is well known that earlier soil moisture data are important for predicting future soil moisture levels. Suppl. Table 4 shows the evaluation metrics of the model employing soil moisture and meteorological data, unlike the results in Table 5, which only used meteorological data as the model inputs. When the previous soil moisture was included in the model, the RMSE showed high prediction results of about 0.3 and a R^2 value close to 0.98. Past soil moisture data play a crucial role in future soil moisture predictions; however, obtaining soil moisture data can be

challenging for many farms engaged in outdoor cultivation. Therefore, this study, which predicts soil moisture only with meteorological and flow rate data without previous soil moisture information, could be an attractive option for farmers who lack access to soil data.

In addition, Table 1 shows that many previous studies used the daily average soil moisture as the unit of soil moisture prediction. However, soil moisture can vary greatly during the day depending on specific environmental factors. Here, we attempted to implement more real-time predictions by making hourly predictions of the soil moisture.

This study has limitations that complicate the generalizability of the findings because the prediction model presented here was not derived from extensive data in various regions. Therefore, we are planning a follow-up study to install soil moisture sensors and collect data that consider various environmental factors, including the soil properties, topography, and vegetation patterns, among other types, as presented in a number of earlier works (Grayson et al., 1997; Western et al., 1999; Qiu et al., 2001; Hupet and Vanclooster, 2002; Williams et al., 2003; Xu et al., 2008; Takagi and Lin, 2011; Zhu and Lin, 2011; Garcia et al., 2014). The soil moisture prediction performance of the model developed in this study is expected to improve if the asynchronous relationship between the soil and the atmosphere is identified and physiological factors are applied to the model input in future studies. Additionally, the developed model can serve as a framework for future soil moisture research to predict soil moisture levels with only meteorological and flow rate data in the absence of historical soil moisture data.

Literature Cited

Acharya U, Daigh AL, Oduor PG (2021) Machine learning for predicting field soil moisture using soil, crop, and nearby weather station data in the Red River Valley of the North. Soil Syst 5:57. doi:10.1016/S0022-1694(02)00016-1

Adeyemi O, Grove I, Peets S, Domun Y, Norton T (2018) Dynamic neural network modelling of soil moisture content for predictive irrigation scheduling. Sens 18:3408. doi:10.3390/s18103408

Aerts R (1997) Climate, leaf litter chemistry and leaf litter decomposition in terrestrial ecosystems: a triangular relationship. Oikos 79:439-449. doi:10.2307/3546886

Breiman L (1996) Bagging predictors. Mach Learn 24:123-140. doi:10.1007/BF00058655

Breiman L (2001) Random forests. Mach Learn 45:5-32. doi:10.1023/A:1010933404324

Bressler SL, Seth AK (2011) Wiener-Granger causality: a well established methodology. Neuroimage 58:323-329. doi:10.1016/j.neuroimage.2010.02.059

Cai Y, Zheng W, Zhang X, Zhangzhong L, Xue X (2019) Research on soil moisture prediction model based on deep learning. PLoS ONE 14:e0214508. doi:10.1371/journal.pone.0214508

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273-297. doi:10.1007/BF00994018

Cutler A, Cutler DR, Stevens JR (2012) Random Forests. In C Zhang, Y Ma, eds, Ensemble machine learning. Springer, New York, NY, pp 157-175. doi:10.1007/978-1-4419-9326-7_5

Deutsch ES, Bork EW, Willms WD (2010) Soil moisture and plant growth responses to litter and defoliation impacts in Park-land grasslands. Agric Ecosyst Environ 135:1-9. doi:10.1016/j.agee.2009.08.002

Dickey DA, Fuller WA (1979) Distribution of the estimators for autoregressive time series with a unit root. J Am Stat Assoc 74:427-431. doi:10.1080/01621459.1979.10482531

Dirmeyer PA, Schlosser CA, Brubaker KL (2009) Precipitation, recycling, and land memory: An integrated analysis. J Hydrometeorol 10:278-288. doi:10.1175/2008JHM1016.1

Filipović N, Brdar S, Mimić G, Marko O, Crnojević V (2022) Regional soil moisture prediction system based on Long Short-Term Memory network. Biosyst Eng 213:30-38. doi:10.1016/j.biosystemseng.2021.11.019

Gao P, Qiu H, Lan Y, Wang W, Chen W, Han X, Lu J (2022) Modeling for the prediction of soil moisture in litchi orchard with deep long short-term memory. Agriculture 12:25. doi:10.3390/agriculture12010025

Garcia GM, Pachepsky YA, Vereecken H (2014) Effect of soil hydraulic properties on the relationship between the spatial mean and variability of soil moisture. J Hydrol 516:154-160. doi:10.1016/j.jhydrol.2014.01.069

Gill MK, Asefa T, Kemblowski MW, McKee M (2006) Soil moisture prediction using support vector machines. J Am Water Resour Assoc

42:1033-1046. doi:10.1111/j.1752-1688.2006.tb04512.x

Gomathi R, Gururaja Rao PN, Chandran K, Selvi A (2015) Adaptive responses of sugarcane to waterlogging stress: An over view. Sugar Tech 17:325-338. doi:10.1007/s12355-014-0319-0

Granger CW (1969) Investigating causal relations by econometric models and cross-spectral methods. J Econom 424-438. doi:10.2307/1912791

Grayson RB, Western AW, Blöschl G (1997) Preferred states in spatial soil moisture patterns: local and non local controls. Water Resour Res 33:2897-2908. doi:10.1029/97WR02174

Hou XL, Feng YH, Wu GH, He YX, Chang DM, Yang H (2016) Application research on artificial neural network dynamic prediction model of soil moisture. Water Saving Irrig 7:70-72

Hupet F, Vanclooster M (2002) Intraseasonal dynamics of soil moisture variability within a small agricultural maize cropped field. J Hydrol 261:86-101. doi:10.1016/S0022-1694(02)00016-1

Ji R, Zhang S, Zheng L, Liu Q (2017) Prediction of soil moisture based on multilayer neural network with multi-valued neurons. Trans Chin Soc Agric Eng 33:126-131. doi:10.11975/j.issn.1002-6819.2017.z1.019

Kwon SH, Kim DH, Kim JS, Jung KY, Lee SH, Kwon JK (2020) Soil water flow patterns due to distance of two emitters of surface drip irrigation for horticultural crops. Hortic Sci Technol 38:631-644. doi:10.7235/HORT.20200058

Lamorski K, Pastuszka T, Krzyszczak J, Sławiński C, Witkowska-Walczak B (2013) Soil water dynamic modeling using the physical and support vector machine methods. Vadose Zone J 12:1-12. doi:10.2136/vzj2013.05.0085

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to hand-written zip code recognition. Neural Comput 1:541-551. doi:10.1162/neco.1989.1.4.541

Lewis-Beck C, Lewis-Beck M (2015) Multiple Regression: The Basics. In Applied Regression: An Introduction, Ed 2, Vol 22. Sage publications, California, USA, pp 55-74. doi:10.4135/9781483396774.n3

Lu CJ, Lee TS, Chiu CC (2009) Financial time series forecasting using independent component analysis and support vector regression. Decis Support Syst 47:115-125. doi:10.1016/j.dss.2009.02.001

Lu T, Sun J, Wu K, Yang Z (2018) High-speed channel modeling with machine learning methods for signal integrity analysis. IEEE Trans Electromagn Compat 60:1957-1964. doi:10.1109/TEMC.2017.2784833

Marshall CH, Pielke RA, Steyaert LT, Willard DA (2004) The impact of anthropogenic land-cover change on the Florida peninsula sea breezes and warm season sensible weather. Mon Weather Rev 132:28-52. doi:10.1175/1520-0493(2004)132<0028:TIOALC>2.0.CO;2

Matei O, Rusu T, Petrovan A, Mihuţ GA (2017) data mining system for real time soil moisture prediction. Procedia Eng 181:837-844. doi:10.1016/j.proeng.2017.02.475

McCabe MF, Wood EF, Wójcik R, Pan M, Sheffield J, Gao H, Su H (2008) Hydrological consistency using multi-sensor remote sensing data for water and energy cycle studies. Remote Sens Environ 112:430-444. doi:10.1016/j.rse.2007.03.027

Montavon G, Samek W, Müller KR (2018) Methods for interpreting and understanding deep neural networks. Digit Signal Process 73:1-15. doi:10.1016/j.dsp.2017.10.011

Mushtaq R (2011) Augmented dickey fuller test. SSRN 1-19. doi:10.2139/ssrn.1911068

Nam S, Hong C, An SK, Kim J (2023) Low substrate water content is efficient for the performance of *Ficus pumila* 'Variegata' indoors. Hortic Environ Biotechnol 64:583-591. doi:10.1007/s13580-023-00514-1

Neitsch SL, Arnold JG, Kiniry JR, Williams JR (2011) Equations: Atmospheric Water. In Soil and Water Assessment Tool Theoretical Documentation Version 2009. Texas Water Resources Institute, Texas, USA, pp 51-64

Ntukamazina N, Onwonga RN, Sommer R, Mukankusi CM, Mburu J, Rubyogo JC (2017) Effect of excessive and minimal soil moisture stress on agronomic performance of bush and climbing bean (*Phaseolus vulgaris* L.). Cogent Food Agric 3:1373414. doi:10.1080/23311932.2017.1373414

Prasad R, Deo RC, Li Y, Maraseni T (2018) Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition. Geoderma 330:136-161. doi:10.1016/j.geoderma.2018.05.035

Qi J, Du J, Siniscalchi SM, Ma X, Lee CH (2020) Analyzing upper bounds on mean absolute errors for deep neural net-work-based vector-to-vector regression. IEEE Trans Signal Process 68:3411-3422. doi:10.1109/TSP.2020.2993164

Qiu Y, Fu B, Wang J, Chen L (2001) Soil moisture variation in relation to topography and land use in a hillslope catchment of the Loess Plateau, China. J Hydrol 240:243-263. doi:10.1016/S0022-1694(00)00362-0

Rushton KR, Eilers VHM, Carter RC (2006) Improved soil moisture balance methodology for recharge estimation. J Hydrol 318:379-399. doi:10.1016/j.jhydrol.2005.06.022

Shi CY, Liu L, Li QL, Wei ZF, Gao DT (2022) Comparison of drought resistance of rootstocks 'M9-T337' and 'M26' grafted with 'Huashuo' apple. Hortic Environ Biotechnol 63:299-310. doi:10.1007/s13580-021-00398-z

Šimůnek J, Van Genuchten MT (2008) Modeling nonequilibrium flow and transport processes using HYDRUS. Vadose Zone J 7:782-797. doi:10.2136/vzj2007.0074

Takagi K, Lin HS (2011) Temporal dynamics of soil moisture spatial variability in the shale hills critical zone observatory. Vadose Zone J 10:832-842. doi:10.2136/vzj2010.0134

Van Dam JC, Huygen J, Wesseling JG, Feddes RA, Kabat P, Van Walsum PEV, Groenendijk P, Van Diepen CA (1997) Soil Water Flow. In Theory of SWAP version 2.0; Simulation of water flow, solute transport and plant growth in the soil-water-atmosphere-plant environment. DLO Winand Staring Centre, Wageningen, Netherlands, pp 21-38

Wang Y, Yang J, Chen Y, Fang G, Duan W, Li Y, De Maeyer P (2019) Quantifying the effects of climate and vegetation on soil moisture in an arid area, China. Water 11:767. doi:10.3390/w11040767

Western AW, Grayson RB, Blöschl G, Willgoose GR, McMahon TA (1999) Observed spatial organization of soil moisture and its relation to terrain indices. Water Resour Res 35:797-810. doi:10.1029/1998WR900065

Williams AG, Ternan JL, Fitzjohn C, De Alba S, Perez-Gonzalez A (2003) Soil moisture variability and land use in a season-ally arid environment. Hydrol Process 17:225-235. doi:10.1002/hyp.1120

Wu CH, Ho JM, Lee DT (2004) Travel-time prediction with support vector regression. IEEE Trans Intell Transp Syst 5:276-281. doi:10.1109/TITS.2004.837813

Xu XL, Ma KM, Fu BJ, Song CJ, Liu W (2008) Relationships between vegetation and soil and topography in a dry warm river valley, SW China. Catena 75:138-145. doi:10.1016/j.catena.2008.04.016

Yan W, Zhou Q, Peng D, Wei X, Tang X, Yuan E, Wang Y, Shi C (2021) Soil moisture responses under different vegetation types to winter rainfall events in a humid karst region. Environ Sci Pollut 28:56984-56995. doi:10.1007/s11356-021-14620-z

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. doi:10.1109/ACCESS.2020.3034984

Yu J, Zhang X, Xu L, Dong J, Zhangzhong L (2021) A hybrid CNN-GRU model for predicting soil moisture in maize root zone. Agric Water Manag 245:106649. doi:10.1016/j.agwat.2020.106649

Zhu Q, Lin H (2011) Influences of soil, terrain, and crop growth on soil moisture variation from transect to farm scales. Geoderma 163:45-54. doi:10.1016/j.geoderma.2011.03.015