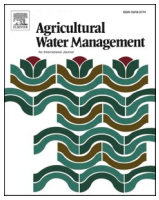
Agricultural Water Management 299 (2024) 108869

Contents lists available at ScienceDirect

Agricultural Water Management

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



Classification and Regression Tree (CART)-based estimation of soil water content based on meteorological inputs and explorations of hydrodynamics behind☆

Tsung-Hsi Wu [a,*], Pei-Yuan Chen [b], Chien-Chih Chen [a,c], Meng-Ju Chung [b], Zheng-Kai Ye [a], Ming-Hsu Li [b]

[a] Department of Earth Sciences, National Central University, Zongda Road, Taoyuan 32001, Taiwan
[b] Graduate Institute of Hydrological & Oceanic Sciences, National Central University, Zongda Road, Taoyuan 32001, Taiwan
[c] Earthquake-Disaster & Risk Evaluation and Management Center, Zongda Road, Taoyuan 32001, Taiwan

ARTICLE INFO

Keywords:
Estimation
Soil-water Content
Classification and Regression Tree (CART)
Meteorological features
Hydrodynamics
Ultisols

ABSTRACT

In this study, we investigate the feasibility of using the Classification and Regression Tree (CART) algorithm to estimate soil water content (SWC) using commonly available meteorological parameters. We trained and validated CART models using data collected in a grassland terrain in northern Taiwan throughout the year of 2018, with the goal of providing precise information for agricultural irrigation and flood risk assessment. Results indicate the effectiveness of CART in SWC estimation, with error levels acceptable for agricultural purposes. The mean absolute error is less than 4% (v/v) for 53 out of the total 60 models in the 12-fold Time-Series Cross-Validation for SWC at depths of 10, 30, 50, 70, and 100 cm. Furthermore, the effectiveness of meteorological parameters in different sets of time shifting and parameter types are assessed. Our findings reveal that the responsiveness of SWC to parameters derived from precipitation varies with soil depth and season, with SWC dynamics in response to precipitation being more pronounced in shallower layers ($\leq$50 cm) compared to deeper layers ($\geq$70 cm). The influence of precipitation-derived and non-precipitation parameters on SWC dynamics is manifested in their distinct feature-importance characteristics in a CART model. This study highlights the importance of understanding characteristics of rainfall and underlying hydrological dynamics, such as evapotranspiration and soil texture, in order to make accurate SWC predictions using CART. Since CART serves as the basis for a variety of top-performing machines like random forest and gradient-boosted trees, the discoveries from this study can also help estimate SWC with these advanced algorithms. Overall, the results of this study provide practical guidance for refining machine-learning based SWC estimations, contributing to more effective agricultural water management and irrigation strategies.

1. Introduction

Developing methods for accurately estimating soil water content (SWC) in the vadose/root zone and understanding its dynamics is crucial for various applications. Previous research addressing the topics of the influence of interaction between soil-water-plant on the environment (Maroufpoor et al., 2019), precision agriculture (Sobayo et al., 2018), and urban green space water management (Garg et al., 2020) has reviewed or highlighted the importance of understanding the dynamics of SWC and related hydraulic parameters for maintaining these systems. SWC estimates affect hydrological, biological, and meteorological processes as they control the partitioning of water and energy between the land surface and atmosphere. Moreover, precise irrigation based on accurately estimated SWC maintains crop growth while saving water resources (Peters et al., 2013; Angelaki et al., 2023). This similar concept applies to urban green infrastructure with different types of vegetation and soil, planting purposes, and irrigation management.

SWC can be directly measured via the gravimetric method, which

☆ This document is the results of the research project funded by MOST.
* Corresponding author.
E-mail addresses: tsung.hsi@g.ncu.edu.tw, okatsn@gmail.com (T.-H. Wu), pychen@ncu.edu.tw (P.-Y. Chen), chienchih.chen@g.ncu.edu.tw (C.-C. Chen), cc105626007@g.ncu.edu.tw (M.-J. Chung), supon3060@gmail.com (Z.-K. Ye), mli@ncu.edu.tw (M.-H. Li).

https://doi.org/10.1016/j.agwat.2024.108869
Received 2 August 2023; Received in revised form 21 October 2023; Accepted 6 May 2024
Available online 25 May 2024

involves drying soil samples and measuring the weight loss (Reynolds, 1970). Another method is the nuclear resonance method, which measures the interaction of water molecules with a magnetic field (Hendricks et al., 1999; Dawson et al., 2017; Shen et al., 2019). While these approaches can provide highly accurate estimations of SWC, their high costs make them impractical for in-situ deployment in most scenarios mentioned previously. In the context of field instrumentation, a variety of indirect approaches have been developed to make continuous SWC monitoring in fields viable. These include resistivity methods (Brunet et al., 2010), tension methods (Bittelli, 2010; Shekhar et al., 2017), dielectric methods (Toková et al., 2019; He et al., 2021; Robinson et al., 2003), and fiber optic sensing (Cosenza, 2016; Aldaba et al., 2018; Sun et al., 2022). Due to economic and deployment reasons, in-situ SWC monitoring is generally spatially sparse. Therefore, physically based numerical models and data-driven models are required on one hand to fill gaps in SWC measurement instrumentation. On the other hand, these models can be utilized to predict SWC and provide support for early responses in water management activities.

Conventionally, physically based numerical models allow the simulation of infiltration; for example, the current state-of-the-art models HYDRUS and SWAP are capable of simulating the movement of solutes in the soil-plant-atmospheric system (Simunek et al., 2016; Kroes et al., 2017; Pinheiro et al., 2019) by including partial differential equations that describe the flow of water in unsaturated porous media. Regional pedotransfer functions (PTFs) are also widely used to identify the parameters required in hydrological models; they are empirical statistical models used for estimating soil water-related properties using more readily available measurements (Wösten et al., 1999,2001; Pachepsky and van Genuchten, 2011; Patil and Singh, 2016). However, numerical models require accurate spatial-temporal initial conditions of parameters such as hydraulic conductivity, soil texture, inflow and outflow conditions; obtaining high-quality data for these parameters is not easy in most agricultural scenarios. Empirical statistical models such as PTFs, on the other hand, generally require a substantial amount of soil data to develop the function. In recent years, machine learning (ML) techniques have gained popularity for estimating and understanding the dynamics of soil water content, as well as correlated soil properties. Compared to numerical simulation, ML methods are attractive for their flexibility in prediction, as they generally require much less detailed and comprehensive hydraulic parameters. In addition, ML models can be trained on more diverse data sources without the need for explicit boundary conditions and initial conditions. Considering model performances, ML models can produce comparable and robust results to SWAP- or HYDRUS-like models, but require significantly fewer computational resources for estimating SWC in precision agriculture, environmental monitoring, or hydrological modeling (Li et al., 2020; Leonarduzzi et al., 2022; Yang and Mei, 2022; Lei et al., 2023).

ML models are data-driven and intrinsically lack physics-based and mechanistic understanding, as compared to traditional numerical models. In the context of applying ML models for soil water estimation or prediction, the purpose of training a model is to learn implicit patterns and relationships from observed data. In practice, the trained models can use various relational parameters such as meteorological parameters (e.g., Cai et al., 2019; Yu et al., 2020, 2021; An and Zhao, 2021; Huang et al., 2021; Tramblay and Quintana Seguí, 2022), remote sensing indices or imagery data (e.g., Liu et al., 2018; Adab et al., 2020; Araya et al., 2020; Greifeneder et al., 2021; Nguyen et al., 2022), hydraulic parameters and soil properties other than SWC (e.g., Leij et al., 2022; Lei et al., 2023), and solely SWC observations (e.g., Datta and Faroughi, 2023) to derive SWC values for SWC interpolation, nowcast, or forecast.

In our survey of studies using ML methods for SWC estimation within the past 5 years at the time of writing, we categorized the most popular ML methods into the following three categories: Decision Tree-based models (DT) (e.g., Liu et al., 2018; Greifeneder et al., 2021; Nguyen et al., 2022; Oliveira et al., 2021; Tramblay and Quintana Seguí, 2022), Support Vector Machines (SVM) (e.g., Huang et al., 2021), and Artificial Neural Networks (ANN) or "deep learning" methods (e.g., Cai et al., 2019; Yu et al., 2020, 2021; An and Zhao, 2021; Leij et al., 2022; Datta and Faroughi, 2023). Some studies aim to compare the performance of different methods from these three categories in specific scenarios of SWC estimation (e.g., Adab et al., 2020; Araya et al., 2020; Greifeneder et al., 2021; Oliveira et al., 2021). The top-performing models are often sophisticated DT or ANN models, demonstrating the effectiveness of DT and ANN methods in estimating SWC for purposes such as accurate predictions or fine interpolation of SWC in various climatic and geographic conditions.

For example, Araya et al. (2020) shows that in estimating SWC of depth 0–4 cm using high-resolution multispectral imagery, terrain attributes, and climate covariates, the boosted regression tree algorithm yielded the most accurate results compared to ANN and SVM methods. In estimating SWC of depths 10 and 20 cm in farmland of three sites in Beijing, China, using meteorological parameters as input features, a deep learning regression network model with two hidden layers outperforms both other ANN models and a SVM, demonstrating good data fitting and generalization capabilities (Cai et al., 2019). The works of Greifeneder et al. (2021) and Nguyen et al. (2022) both adopted DT methods for incorporating remote sensing data to provide a high-resolution SWC map; both of the results show that boosted trees can provide accurate SWC estimation better than Random Forest (RF) and SVM. In SWC prediction for corn production purposes, Yu et al. (2021) proposes a hybrid CNN-GRU model that integrates the strong feature expression ability of Conventional Neural Network (CNN) with the time series memory ability of Gated Recurrent Unit (GRU). It outperformed individual CNN and GRU models in terms of prediction accuracy and convergence rate, with improved accuracy at greater soil depths that could be a significant reference for agricultural irrigation applications. In the study of Tramblay and Quintana Seguí (2022), the RF model is capable of providing SWC estimations of comparable quality but greater robustness compared to regional-specific pedotransfer functions in the context of drought monitoring. Adab et al. (2020) compares the performance of RF, SVM, ANN, and a regression model in the task of estimating SWC at depths from 10 to 100 cm. The results show that RF outperforms others in a variety of performance metrics, such as Mean Absolute Error (MAE) and Nash-Sutcliffe efficiency. In Oliveira et al. (2021), RF, SVM, average neural network, and weighted k-nearest neighbor were used to model the spatiotemporal dynamics of SWC. The results show that RF was found not only to be the most accurate algorithm but also the most suitable one for predicting the spatiotemporal variability of SWC in an Atlantic forest remnant. Datta and Faroughi (2023) investigates long short-term memory (LSTM) models with SWC itself as input feature to forecast SWC at 10 and 30 cm depths up to one month in advance. The results show that a multihead LSTM (which is composed of several LSTM models) significantly outperforms individual LSTM models. Yu et al. (2020) proposed a ResBiLSTM model that combines residual network and bidirectional LSTM, taking gridded meteorological and SWC data as input to predict SWC at four depths from 20 to 50 cm in three growing stages of corn. The results show ResBiLSTM can achieve a good fit in different growing stages and significantly outperform SVM, RF, and deep learning regression network. These studies show great capabilities of DT and ANN approaches in a variety of scenarios of SWC estimation but also indicate that the increase in performance mostly came at the cost of higher model complexity.

In the scope of general machine learning in recent years, Classification and Regression Tree (CART) is one of the most popular DT algorithms for its flexibility (Jena and Dehuri, 2020). It is particularly excellent in handling heterogeneous datasets that contain both categorical and numerical data, which is still generally challenging for ANN-based models (Kadra et al., 2021). Because of its simplicity, CART commonly serves as the basis of the top-performing DT methods. It is often chosen as the weak learner for ensemble learning, such as random

forests, and gradient boosting machines, such as XGBoost (Chen and Guestrin, 2016), which have dominated Kaggle competitions in recent years (Bentéjac et al., 2021). In a comprehensive review of ML techniques used in soil science (Padarian et al., 2020), ANN models are considered highly performing in terms of predictive accuracy but are usually labeled as "black-box" models due to their complex nature. On the other hand, CART allows assessing the importance of variables and is especially preferred when the available dataset is small.

Precipitation is usually considered to have the most direct impact on SWC in the vadose zone, especially in regions with frequent rainfall like Taiwan. Persistent rainfall fills the near-surface soil and saturates it. Few studies using ML methods to estimate SWC have addressed how the saturation of soil and changes in evapotranspiration mechanism affect the model performance and stability. In this study, we use CART to estimate SWC for agricultural purposes. CART is chosen for its simplicity, intrinsic feature selection ability, and extensibility to more sophisticated tree algorithms like RF and boosted trees. The main objective of this study is to investigate the feasibility of using CART for estimating SWC and to assess the time-lag effects of precipitation-derived and non-precipitation meteorological parameters. CART has been shown to be a powerful algorithm for feature selection, automatically identifying the most important features during the tree induction process, as reported in previous studies (Questier et al., 2005; Grabczewski and Jankowski, 2005; Zhou et al., 2021). By leveraging this intrinsic ability for feature optimization, we were able to incorporate all meteorological parameters of interest, including air temperature, relative humidity, solar radiation, air pressure, and precipitation, with data observed up to the past 192 h for a proof-of–concept task of SWC estimation. Our dataset consisted of SWC measurements at five different depths and meteorological variables collected throughout the year 2018 in a grassland terrain site in northern Taiwan. We trained and validated our CART models using this dataset to estimate SWC one hour ahead, with the ultimate goal of providing precise information for agricultural irrigation and flood risk assessment.

The findings of this study indicate the effectiveness of using CART for estimating SWC in a subtropical grassland terrain, with error levels acceptable for agricultural purposes. The CART model was found to adapt distinct strategies for estimating SWC at depths of $\leq$50 cm and $\geq$70 cm, revealing a seasonal change in the infiltration mechanism. By validating our models throughout the different seasons of the year, we were able to explore the underlying hydrological dynamics of our study area and determine the key factors for a successful CART model.

Our findings also highlight the importance of having a general understanding of the characteristics of rainfall, evapotranspiration, and soil texture before applying ML methods to estimate SWC. For instance, if it is estimated that the low SWC may limit vegetation growth after 3 days, it is necessary to confirm in advance the irrigation water source and whether it is sufficient. This information is especially helpful for vegetations without readily supplied irrigation channels. However, new SWC observation program is just getting started and long-duration local SWC data is rarely found, which explains why this study uses a grassland with data available for the whole year as an example. SWC estimation gradually gains attentions recently in Taiwan due to frequently occurred droughts and floods events. The Central Weather Administration of Taiwan is also making efforts to improve the accuracy of SWC estimation in few-kilometer scale. It manifests the novelty and value of this study as a small-scale counterpart. Considering Taiwan's special geographical and hydrological characteristics, estimating SWC of a grassland in Taiwan also help verify the applicability of the machine learning method proposed in previous studies.

2. Data

The study area is located at the Atmospheric and Hydrological Observatory (National Central University, NCU, 2024) site, which has a 34-meter square grassland terrain with a vegetation height of 5–20 cm. The study area is relatively flat with no tall buildings or obstacles within 100 m of the site.

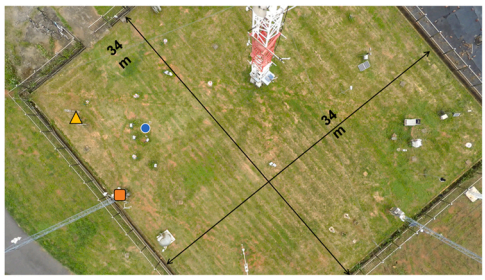


Fig. 1. Aerial view of the NCU Atmospheric and Hydrological Observatory site (Surface Hydrology Lab. of NCU, 2024a).

The meteorological data used in this study were obtained from various instruments, including precipitation (Ota keiki seisakusho/OW-34-BP), air temperature (Vaisala HUMICAP/HMP 155), relative humidity (Vaisala HUMICAP/HMP 155), wind speed (RM Young/05103), atmospheric pressure (Setra/278), and solar radiation (shortwave radiation) (Eppley/Precision Spectral Pyranometer). The SWC was measured using Sentek/EnviroSCAN, which was placed at 10 cm, 30 cm, 50 cm, 70 cm, and 100 cm below the surface.

In Fig. 1, the blue circle represents the location of the SWC sensor, while the yellow triangle represents a rain gauge and the orange square represents a 10-meter tower where additional instruments were placed, including a pressure gauge at 1 m, a temperature-hygrometer meter at 2 m, a pyranometer at 3 m, and an anemorumbograph at 10 m. The distance between the SWC and rainfall measurements is about 10 m, and the distance between the SWC and the meteorological variables is also about 10 m. The location and spacing devices are chosen mainly to avoid interfering with each other and the other existing devices unrelated to this study. Also, the soil texture of each layer is homogeneous, so the location of the soil moisture measurement is irrelevant to the estimated 1-dimensional SWC in this study. Moreover, the data during 2018 was selected because of the amount and quality of data.

The main wet season in the Taoyuan area occurs during the summer months (June to September), with an average annual rainfall of 2300 mm, an annual average temperature of around 22 degrees, an average summer temperature of 27 degrees, and an average winter temperature of 13 degrees (Li and Tseng, 2024).

The data used in this study cover the entire year of 2018, including any data deficiencies or missing data. The completeness of the data is shown in Fig. 2 (a). The original data, collected at a temporal resolution of 10 min, is down-sampled to hourly data by taking the last data point of each hour for modeling purposes. It is important to note that the raw data for precipitation represents the accumulated precipitation in 10-minute intervals, and accumulated precipitation for intervals from 1 h to 3 days is calculated by summing the raw data within the interval, with the time stamp of the derived data matching that of the last point of the raw data in the same interval. Any missing data is filled by using linear interpolation. Fig. 3 and 4 show the time series of meteorological data excluding precipitation, and the corresponding SWC at different depths along with precipitation data, respectively for the year 2018. The data used in this study were collected from the following three sources: the Surface Hydrology Laboratory (Surface Hydrology Lab. of NCU, 2024b) provides data on SWC, air temperature, and precipitation; the Cloud and Aerosol Laboratory (Aerosol and Cloud Laboratory of NCU, 2024) provides supplementary data on relative humidity and atmospheric pressure; and the Atmospheric Boundary Layer and Air Pollution Laboratory (Planetary Boundary Layer and Air Pollution Lab. of NCU, 2024) provides solar radiation data.

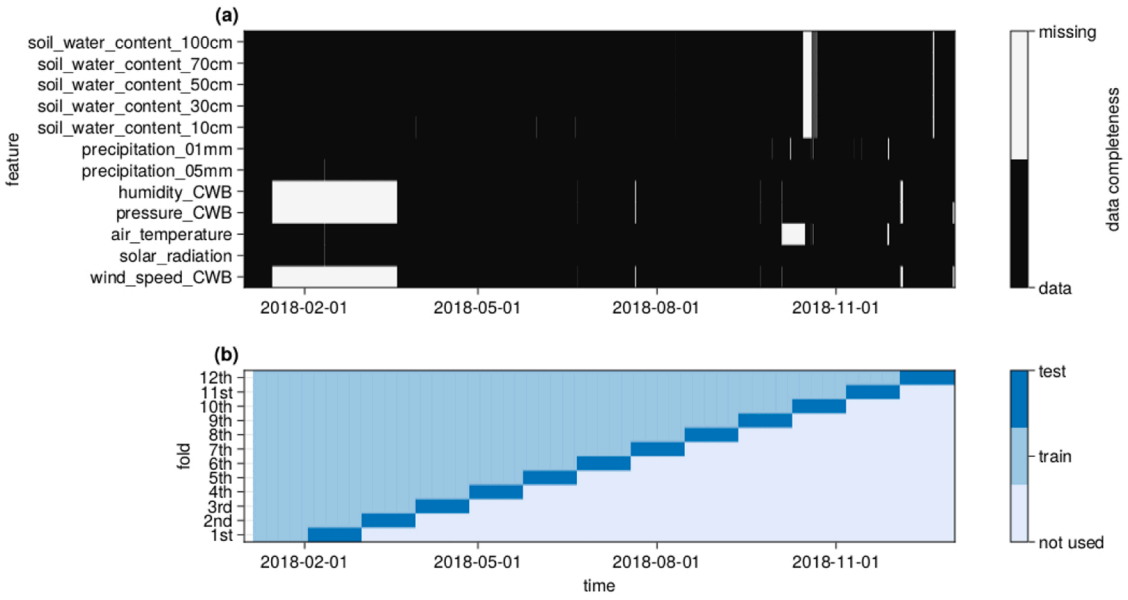


Fig. 2. (a) Missing data in the dataset used in this study. (b) Subsets of training-testing data for cross validation in this study.

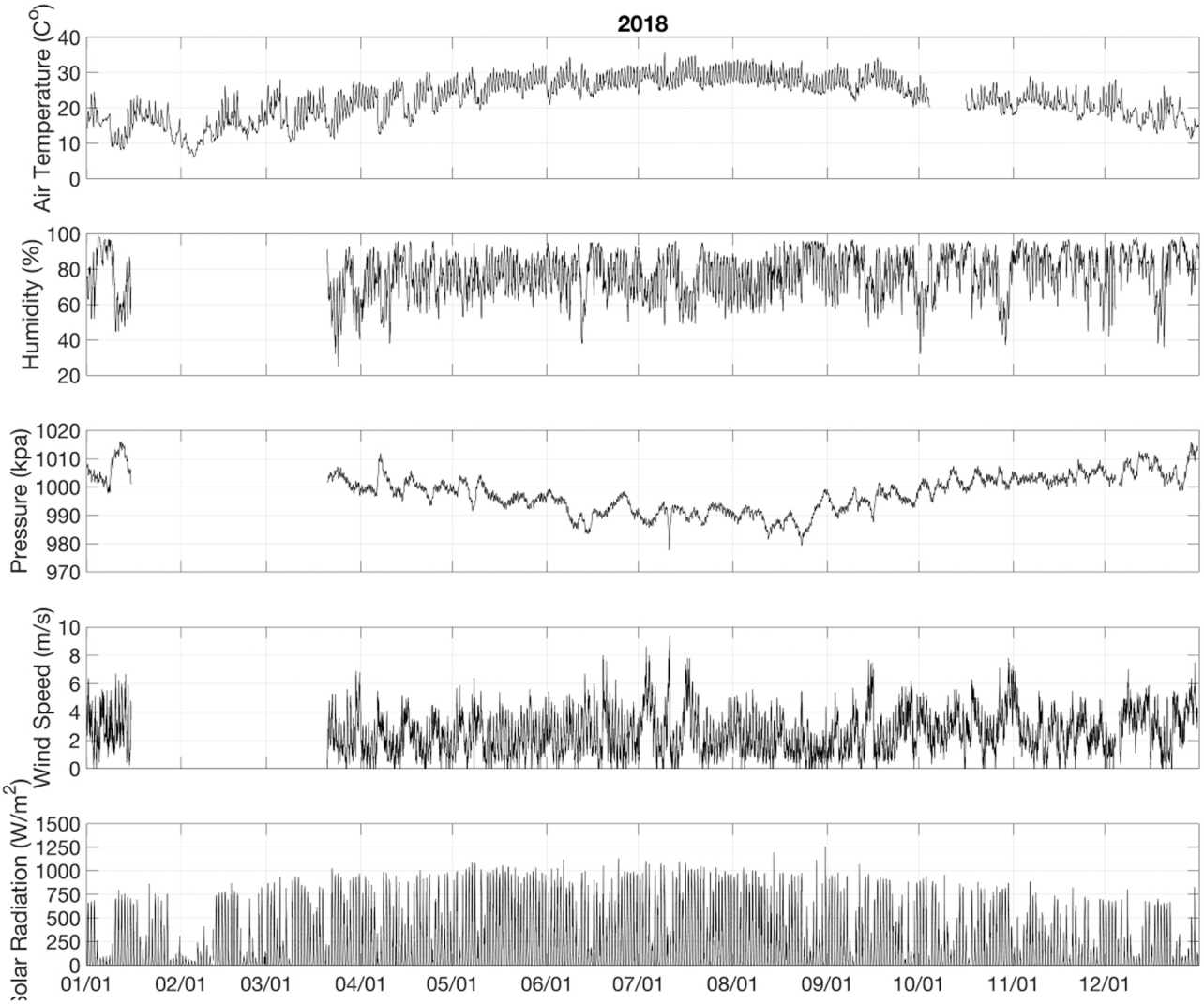


Fig. 3. The time-series data of 2018 for air temperature, humidity, atmospheric pressure, solar radiation, and wind speed.

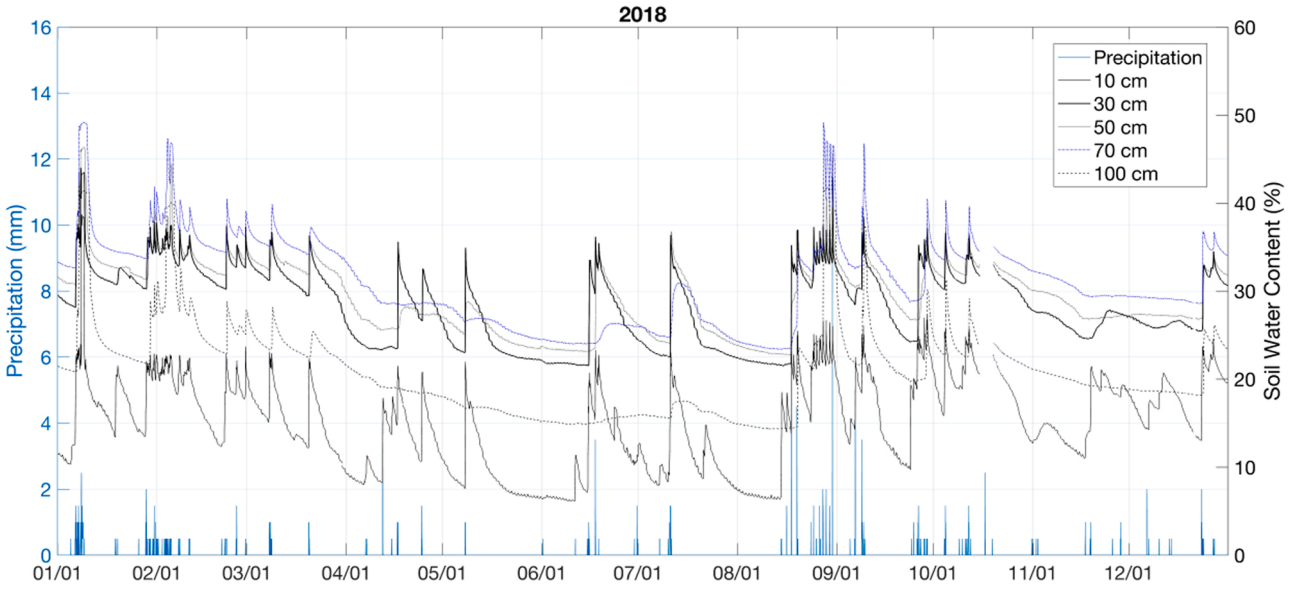


Fig. 4. The soil water content from 10 cm to 100 cm below the soil surface (black line of different styles and blue line) and the rainfall (blue bars).

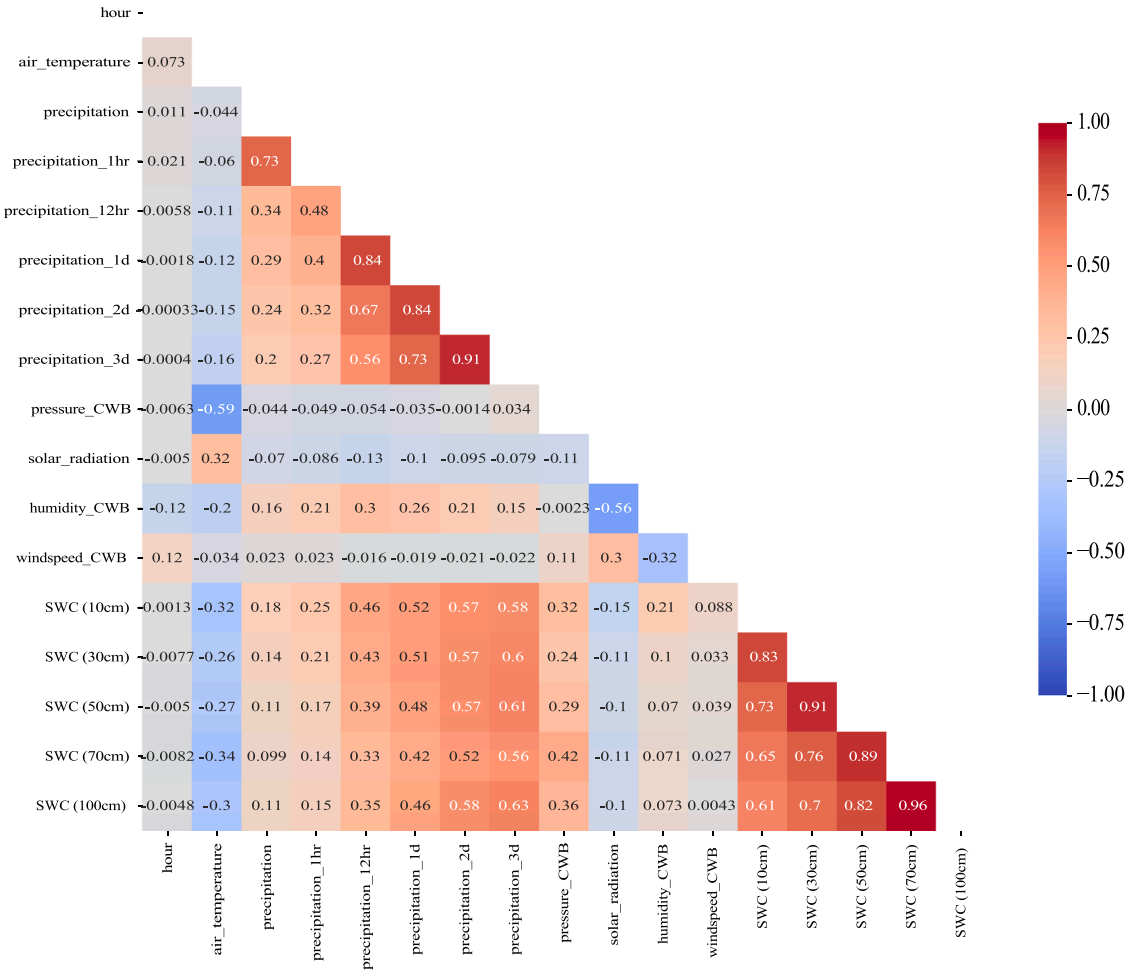


Fig. 5. Correlation matrix of the dataset features (corresponding to Fig. 2 (a)).

The correlation between meteorological data and SWC at the five depths is displayed as a correlation matrix in Fig. 5, with the value in each cell being the standard (Pearson) correlation coefficient. It shows that air temperature negatively correlates with SWC at all depth (~ -0.3); air pressure positively correlates with SWC at all depths ($\sim +0.3$); and the accumulated precipitation strongly positively correlates with SWC at all depths for the time interval of accumulation > 0.5 day(s). For other features, the correlation coefficient is mostly less than $\pm$ 0.2.

Table 1
Basic descriptions for all features.

variable	mean	min	median	max	unit
hour	11.5	0	11.5	23	hour
windspeed_CWB	2.46664	0.0	2.3	9.4	m/s
solar_radiation	163.532	0.0	3.467	1259.0	W/m^2
humidity_CWB	77.8678	25.0	79.0	98.0946	% RH[a]
pressure_CWB	998.503	977.6	1000.43	1016.0	hPa
air_temperature	22.5717	5.9393	23.0683	35.648	$^{\circ}C$
precipitation	0.0330226	0.0	0.0	9.0	mm
precipitation_1hr	0.16955	0.0	0.0	25.0	mm
precipitation_12hr	1.9041	0.0	0.0	68.7	mm
precipitation_1d	3.78865	0.0	0.0	113.5	mm
precipitation_2d	7.54727	0.0	0.2	178.0	mm
precipitation_3d	11.3072	0.0	1.4	226.0	mm
SWC_10cm	15.0258	6.145	15.08	38.71	% V/V[b]
SWC_30cm	28.3645	21.5	28.57	44.0	% V/V
SWC_50cm	29.4681	22.78	29.47	46.35	% V/V
SWC_70cm	31.1118	23.4	30.57	49.21	% V/V
SWC_100cm	20.8358	14.35	20.02	41.79	% V/V

[a] Relative Humidity
[b] Volumetric percentage

3. Methods

3.1. CART algorithm

Tree-based methods are a type of ML algorithm that involves constructing a decision tree to make predictions or decisions. In a decision tree, the root node represents the entire dataset, and the child nodes represent subgroups of the data based on specific conditions. In this study, we under the framework of MLJ (Blaom et al., 2020), apply CART algorithm provided by "DecisionTree.jl" (Sadeghi et al., 2022) for estimating SWC. CART is a popular decision tree algorithm that was first proposed by Breiman in 1984 (Breiman, 2017); it is a supervised ML method that can be used for both classification and regression tasks. In a regression task, such as SWC estimation, the goal is to predict a continuous numerical value; the algorithm works by repeatedly splitting the data based on the feature and threshold, where each split creates binary branches according to the criterion of minimizing the sum of squared error (SSE) (Hastie et al., 2009).

Considering a group of data with N observations, where each observation consists of p inputs and one response, the minimization of SSE in a split is written in the following equation:

$$\min_{j,s}[\min_{c_L}\sum_{x_i\in R_L(j,s)}(y_i-c_L)^2+\min_{c_R}\sum_{x_i\in R_R(j,s)}(y_i-c_R)^2]. \quad (1)$$

In this equation, $x_i=(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(p)})$ denotes the variable for the input features, and y_i denotes the variable of the target feature as the response, with subscript $i=1, 2, \ldots, N$ indicating the i_{th} observation. R denotes a subgroup of dataset at a certain split, with suffixed $_R$ or $_L$ denoting the "right" or "left" branch respectively. c_L or c_R denotes the average response in the subgroup of the left or right branch respectively. In the minimization of the average responses and the sum of squares in the equation, the feature j of threshold s that results in the split of the most distinctive subgroups of observations are obtained.

As pointed out in Hastie et al. (2009) and Carrizosa et al. (2021), the global minimum of SSE is computationally infeasible, and CART is a practical approach that builds the solution incrementally by making the locally optimal choice. Although CART often brings high computational efficiency and provides reasonably good solutions, it is important to note that CART in general does not guarantee the globally optimal partition in terms of SSE due to the lack of considering the entire solution space or future consequences of choices.

3.2. Time series-to-supervised dataset construction

In our task of SWC estimation, the full list of the time series data applied as the input or target features for CART is demonstrated in Table 1. In this table, meteorological parameters (air temperature, relative humidity, solar radiation, wind speed, air pressure, precipitation at the site) and the hour of the day were used for deriving input features; each SWC at a certain depth is the target feature as the response to the inputs. For the purpose of conciseness, the notation NP is used to refer to meteorological parameters that are not precipitation, P represents raw data of precipitation (accumulated in 10 min), and AP stands for accumulated precipitation derived from P throughout this document.

Table 2
Input features and their time-shifted variants of different sets. NP denotes parameters that are not precipitation or parameters derived from precipitation; P denotes the raw data of observed precipitation; and AP denotes the derived data from P. The suffix CWB denotes the data observed by Central Weather Administration of Taiwan.

Group tag/Feature name		Time shift (lag) of input features		
		baseline	longer	denser
NP	hour windspeed_CWB solar_radiation air_temperature pressure_CWB humidity_CWB	0	0	0
P	precipitation precipitation_1hr precipitation_12hr	0,−12, −24, −36, −48	0,−12, −24, −36, −48, −60, −72, −84, −96	0, −6,−12,−18,−24,−30,−36,−42,−48
AP	precipitation_1d precipitation_2d precipitation_3d			

Table 3
List of target features.

Target features		
Abbrev.	Feature name	Time shift
SWC	soil_water_content_10cm soil_water_content_30cm soil_water_content_50cm soil_water_content_70cm soil_water_content_100cm	+1

For the AP parameters, the period over which the amount of precipitation accumulates is indicated by a suffix in the variable name. The suffix is an integer value followed by "hr" (for hours) or "d" (for days). For example, the variable "precipitation_1hr" represents the accumulated precipitation over a period of 1 h. The hour of the day can provide rough information about the position of the sun, as it has been applied as an input feature to estimate SWC in previous studies, such as Pekel (2020). Therefore, we also investigate how this variable affects the estimation of SWC in our study.

To study the relationship between current SWC and earlier information from meteorological parameters, all input features except for "hour" were time-shifted variables, and the actual number of input features is multiplied by the number of time shifts of each parameter accordingly. Table 2 displays the three time-shift sets used in the study: the "baseline" set, which involves information up to −48 h with a step of −12 h; the "longer" set, which involves information up to −96 h with the same time-shift step; and the "denser" set, which has the same earliest available information as the "baseline" set, but twice the density of past information with a step of −6 h.

The target features are listed in Table 3. It should be noted that the five target features do not simultaneously respond to the input dataset; instead, a model is always trained and validated with only one target feature.

3.3. Model validation

To assess the validity of a set of trained models, we use mean absolute error (MAE) between the values given by the trained model (referred to as predicted values) and the real SWC observations (referred to as actual values). The data is split into 12 subsets for cross-validation, namely 12-fold Time-Series Cross-Validation (TSCV). Each subset is then partitioned into a training set used to train the tree model, and a testing set used to validate the performance of the model. For each fold i in TSCV, the model is trained on the folds from 1 to $i-1$ using ground truth data, and evaluated on the testing fold i by comparing the predicted SWC to the actual SWC. The overall performance of a model is determined by averaging its performance across the 12 folds. The 12 partitions have the same length of testing data but have varying lengths of training data, as demonstrated in Fig. 2 (b).

The main difference between TSCV and ordinary cross-validation (CV) is that TSCV preserves the temporal order of the data. This means that in TSCV, the training and testing set always contain consecutive data points, and the testing fold always occurs after the training fold. The choice of TSCV in our study is important because it ensures that the model is evaluated on data observed after the end of training, which is more representative of how the model will be used in a realistic deployment scenario.

3.4. Assessment of feature importance

CART works by greedily splitting the data into subsets based on the features that most reduce the impurity of the nodes, as introduced previously in Sec. 3.1. Due to this nature, CART provides two inherent metrics for assessing the importance of input features: Total Number of Node Splits (TNNS) and Total Decrease in Node Impurity (TDNI). TNNS refers to the total number of splits made by a feature in a tree model; it measures how often a feature was used to split the data during the tree construction process. TDNI represents the sum of impurity decrease over

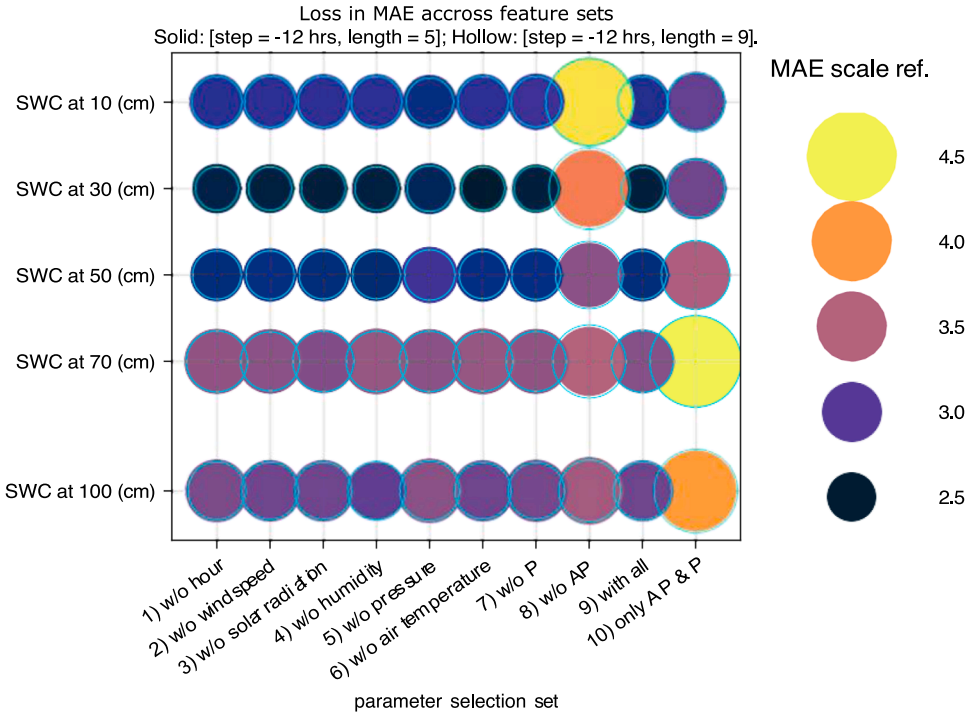


Fig. 6. MAE in the 12-fold TSCV of ten sets of parameter selection for estimating SWC at five depths. The magnitude of MAE is represented by the colors or sizes of circles, with solid circles representing results from "baseline" time shifting, and hollow circles on top of the solid circles representing results from "longer" time shifting. The parameter sets are labeled along the *x*, where "with all" denotes the use of the complete table of input features as listed in Table 2, "w/o" denotes *without* a certain parameter based on the "with all" set, and "only AP & P" denotes the use of only features with group tags AP and P.

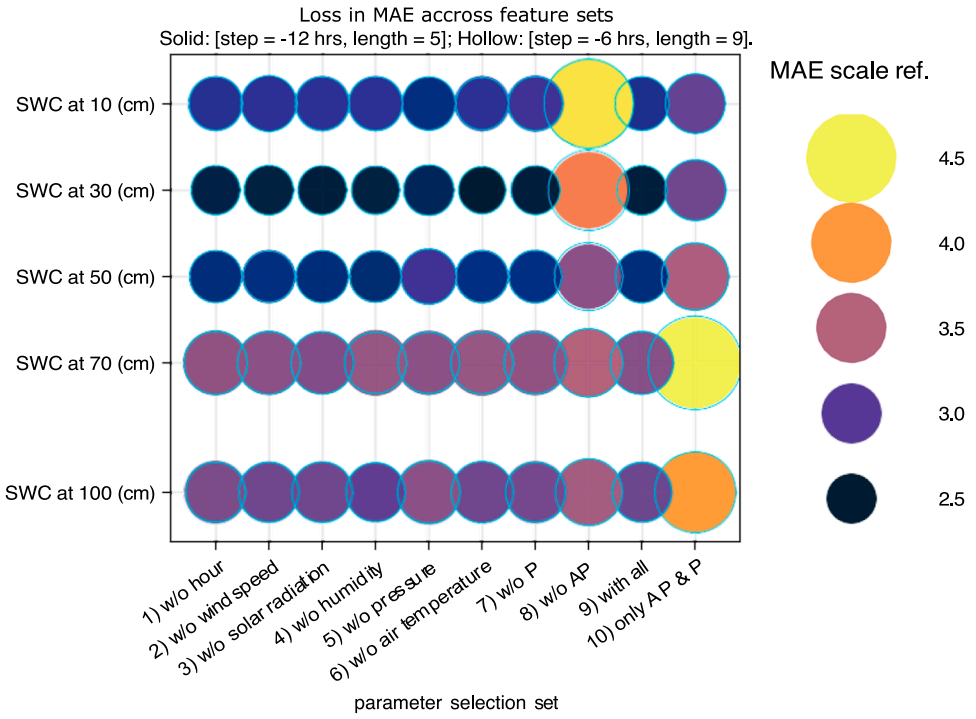


Fig. 7. The same as Fig. 6 but with hollow circles on top of the solid circles representing results from "denser" time shifting.

all nodes in the tree resulting from a feature splitting the training data; it measures how much the error a feature could reduce. In a CART model for regression in "DecisionTree.jl", the impurity decrease at the split of node k is defined as:

$$\Delta\text{impurity} = SSE_k - SSE_R - SSE_L. \tag{2}$$

In this equation, SSE_k, SSE_L, and SSE_R represent the SSE of target values at node k, the left child node of node k, and the right child node respectively.

It should be highlighted that SSE is proportional to the number of data points when Mean Squared Error (MSE) is the same. Thus, the decrease in impurity at a node of lower level (near the root) is generally larger than that at a node of a higher level (near the leaves).

CART assumes that the behavior of the target variable can be modeled through a series of decisions based on a set of input features using a hierarchy of binary branches. If a feature is irrelevant to the target, it will not be able to reduce the impurity of the nodes. Based on this assumption, features with low TDNI are considered to have minimal effects in making predictions. In comparison, a feature with a high TNNS is not necessarily effective, since TNNS does not provide information about the level in the hierarchy structure of a tree where a split occurs, which is critical in making predictions. In this study, TDNI is used to assess the impact of missing data and the importance of NP and AP & P parameters in the 12-fold TSCV to understand their role in dry and wet seasons. On the other hand, TNNS values for the same set of parameters are also calculated but serve as supplementary information.

4. Result and discussion

4.1. Analysis of individual parameters and time shifting effects

To study how meteorological information from earlier pasts and individual parameters within them take effect in SWC estimation, we employ three types of time shifting: "baseline", "longer", and "denser", as introduced in Sec. 3. For each set of time shifting, we examine the effect of individual precipitation (P) and non-precipitation (NP) parameters, as well as the synthetic effect of accumulated precipitation (AP) parameters, by considering ten different sets of parameter selections. Referring to the complete table of input features listed in Table 2, the ten sets of parameter selection are as follows: (1) without *hour*; (2) without *wind speed*; (3) without *solar radiation*; (4) without *humidity*; (5) without *pressure*; (6) without *air temperature*; (7) without *precipitation*; (8) without *AP parameters*; (9) with *all parameters*; and (10) with only *AP and P parameters*. The loss in MAE in the 12-fold TSCV of ten sets of parameter selection for three types of time shifting is demonstrated in Fig. 6 and 7. Both figures display the same results for the "baseline", where the MAE of cross-validation for each case is represented by the size of colored solid circles placed on the grid. On top of the "baseline", the results with "longer" and "denser" time shifting are separately plotted in Fig. 6 and 7, with the MAE represented by the size of hollow circles at the corresponding grid positions for comparison.

In Fig. 6, MAE of the cases with "longer" time shifting (represented by hollow circles) is smaller compared to the "baseline" cases (represented by solid colored circles) for almost every depth and parameter selection set. The comprehensive improvement in model performance for almost all "longer" cases indicates that meteorological information earlier than $t = -48$ h could be helpful in estimating SWC at these depths. On the other hand, MAE of "denser" cases are almost the same as those of "baseline", as demonstrated in Fig. 7, suggesting that temporarily more detailed information than the 12-hour resolution may not be essential for SWC estimation.

Looking at either Fig. 6 or Fig. 7 for the "baseline" results, the loss in MAE of parameter selection sets numbered from (1) to (7) is very close to that of set (9) for every depth, indicating that discarding a single NP or P parameter has minimal effect on the model's loss. This implies that the effect of the absence of a single parameter can be easily compensated for by the contributions of other parameters. On the other hand, the MAE of cases in set (8) is significantly larger than those of set (9), especially for SWC estimation at depths $\leq$30 cm. This indicates that AP parameters play a very critical role in influencing SWC at shallow depths, which is an unsurprising result that reflects the high correlations between AP parameters and SWC of all depths as displayed in Fig. 5. In set (10), using only AP and P parameters, the MAE for every depth is larger than that of set (9), indicating that the summary contribution of NP parameters is

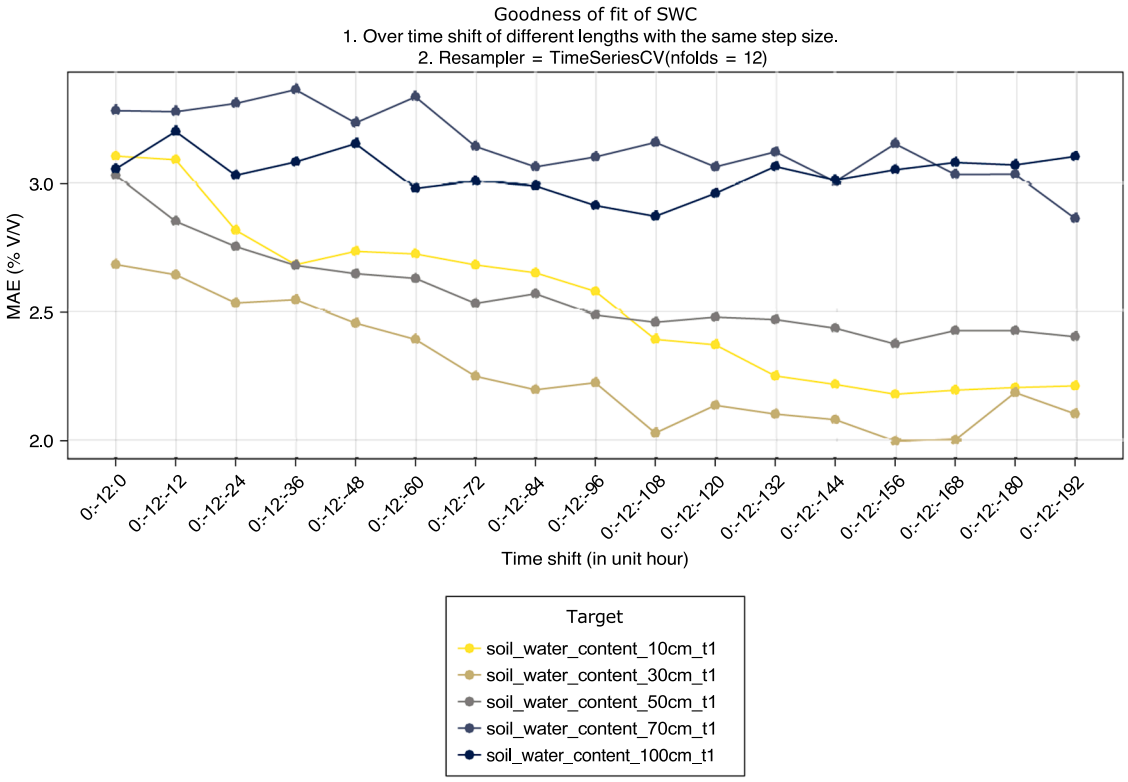


Fig. 8. Learning curve for elongated time lags, where models were trained and validated using all input features in Table 2.

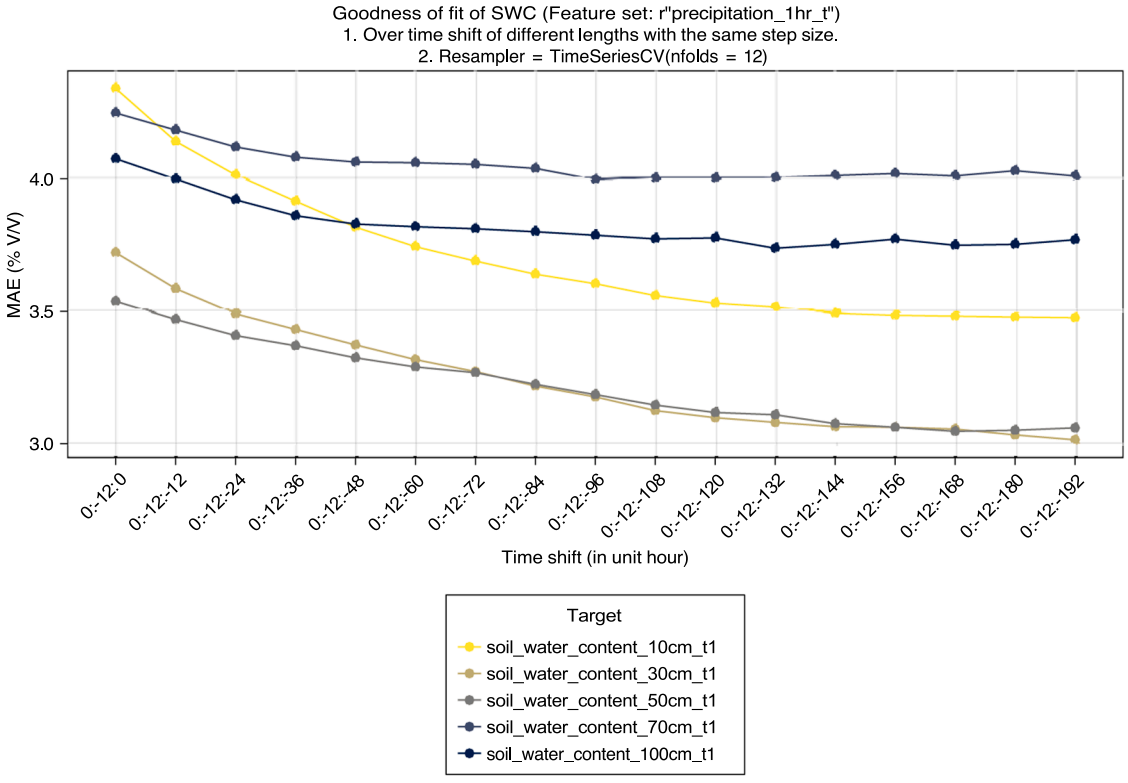


Fig. 9. Learning curve for elongated time lags, where models were trained and validated using only one-hour accumulative precipitation.

crucial for good SWC estimates. Furthermore, according to the results of set (8) and set (10), AP parameters are especially critical for estimating SWC at shallow depths (≤30 cm), and NP parameters are especially important for estimating SWC at deeper depths (≥70 cm).

4.2. Assessing the usefulness of earlier past data and observations on the influence of soil texture and evapotranspiration

In order to further examine to what extent the meteorological

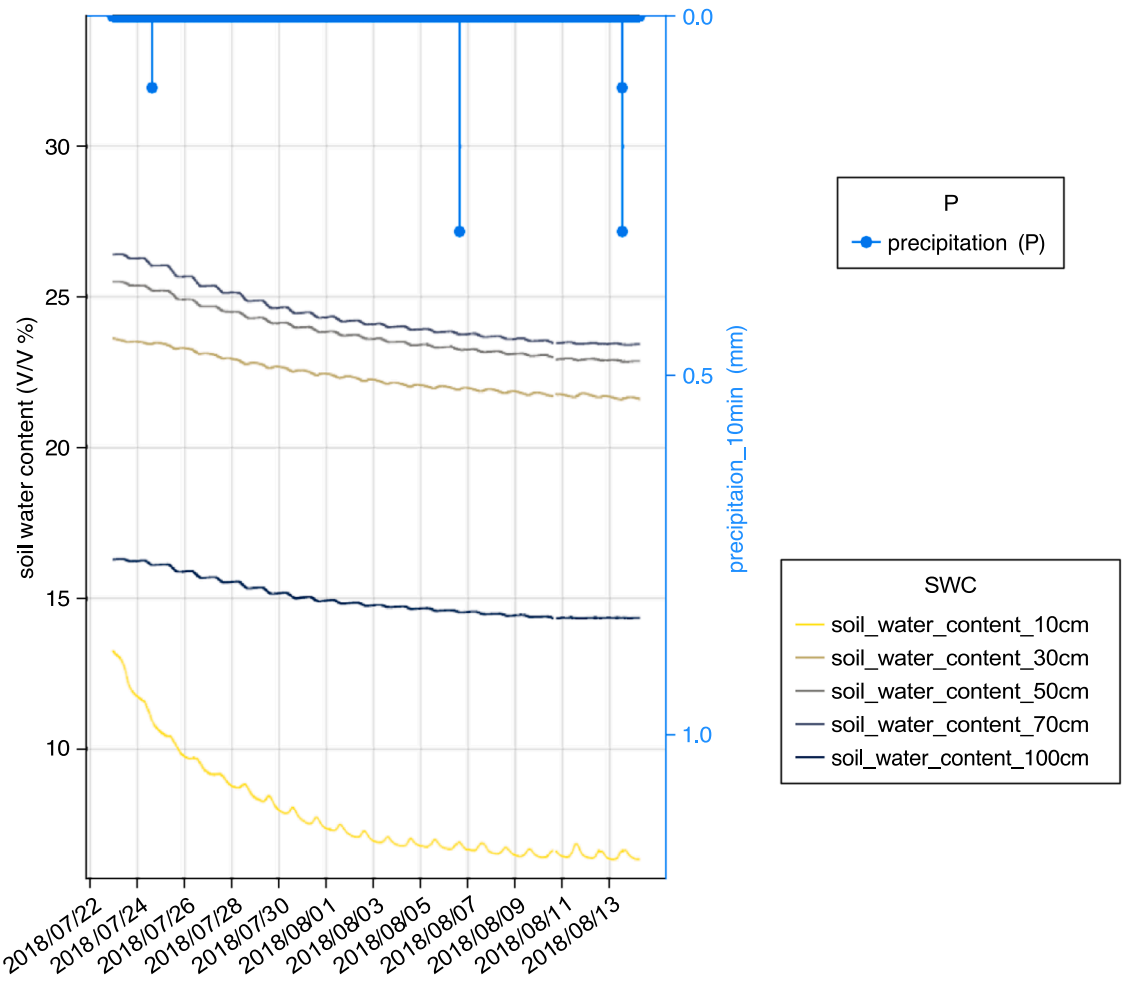


Fig. 10. SWC at five depths v.s. precipitation: a case event of an extreme long-drying summer.

information from earlier past could be useful, we incrementally added more time-shifted data as input features. We assessed model performance using the same 12-fold TSCV, and the results are demonstrated in Fig. 8. Fig. 8 illustrates the relationship between the time shift and the fitting loss in MAE for estimating SWC at five different depths using all the parameters listed in Table 2 with time shifting up to 192 h. Fig. 8 shows that the inclusion of meteorological information with time shifting up to 96 h (i.e., the time shift set "longer" in Table 2) is quite sufficient for SWC estimation at all depths. It is close to the longest rainfall duration of 89 h observed in 2018, if defining the separation of rainfall events as dry periods longer than 12 h.

We observed that cumulatively including information of earlier pasts leads to notable improvements in the estimation of SWC at depths $\leq$50 cm, but provides little improvement in the estimation of SWC at depths $\geq$70 cm. Supported by evidence related to soil texture and observations made during both a prolonged drying process in summer and a separate wetting event in January, we attribute this variation in estimation accuracy to the influence of the evapotranspiration (ET) mechanism. Considering the duration of dry days and the amounts of ET, the influence of the meteorological variables and thus ET at the study site can reach only depths $\leq$50 cm in most of the time in the year. Take an extreme long-drying summer case which is most likely to affect SWC at depth larger than 70 cm as an example, little change in SWC at the depth of 100 cm but large change at the depth of 10 cm is observed after drying for 96 h in late July with solar radiation up to 1100 W/m^2 (Fig. 10). By using only "precipitation_1hr" to estimate SWC, Fig. 9 further suggests that the information of precipitation earlier than $t = -48$ h is not helpful in estimating SWC at depths $\geq$70 cm, but on the other hand brings notable improvements in estimating SWC at depths $\leq$50 cm. The results presented in Fig. 8 and Fig. 9 appear to contradict the intuitive expectation that the influence of meteorological phenomena on SWC at greater depths would take effect at a more delayed time. However, as detailed in the following paragraph, the discrepancy can be attributed to the specific soil texture of the site.

The rock core samples obtained from the site are demonstrated in Fig. 12. Because the top (10 cm) soil layer contains relatively fine grains and the mid-layer (30 – 70 cm) contains coarse soil particles in which water flows easily based on rock core samples, water infiltrates through the preferential-flow path to the deep layer (100 cm) faster than expected. The wetting and beginning of the drying process are demonstrated in Fig. 11 with 82-hour 229.5 mm rain water infiltrating into soil in a relatively dry condition. Although SWC at 10 cm responds first to the precipitation, soil at 100 cm and 70 cm become saturated earlier than the shallow layer. Then, the water accumulated so that 50 cm and 30 cm become saturated as well because the soil at 100 cm seems to be an aquitard in which water is difficult to pass through or be held. Moreover, the SWC at 100 cm also declines faster than SWC at 10 cm after the rainfall intensity decreases to 1 mm per 10 min near the end of the event. The early response and slow recession of the top layer provide evidence that precipitation at earlier time steps, up to 144 h in Fig. 9, influences SWC more at 10 cm than at 100 cm. On the other hand, SWC at 30 – 70 cm shows incremental changes in between corresponding to the core samples. The durations of the saturated periods for SWC at different depths between the rising and recession curves depend on both the rainfall duration and the soil texture and vary from event to event. This 82-hour 229.5 mm event has the largest rainfall amount and is

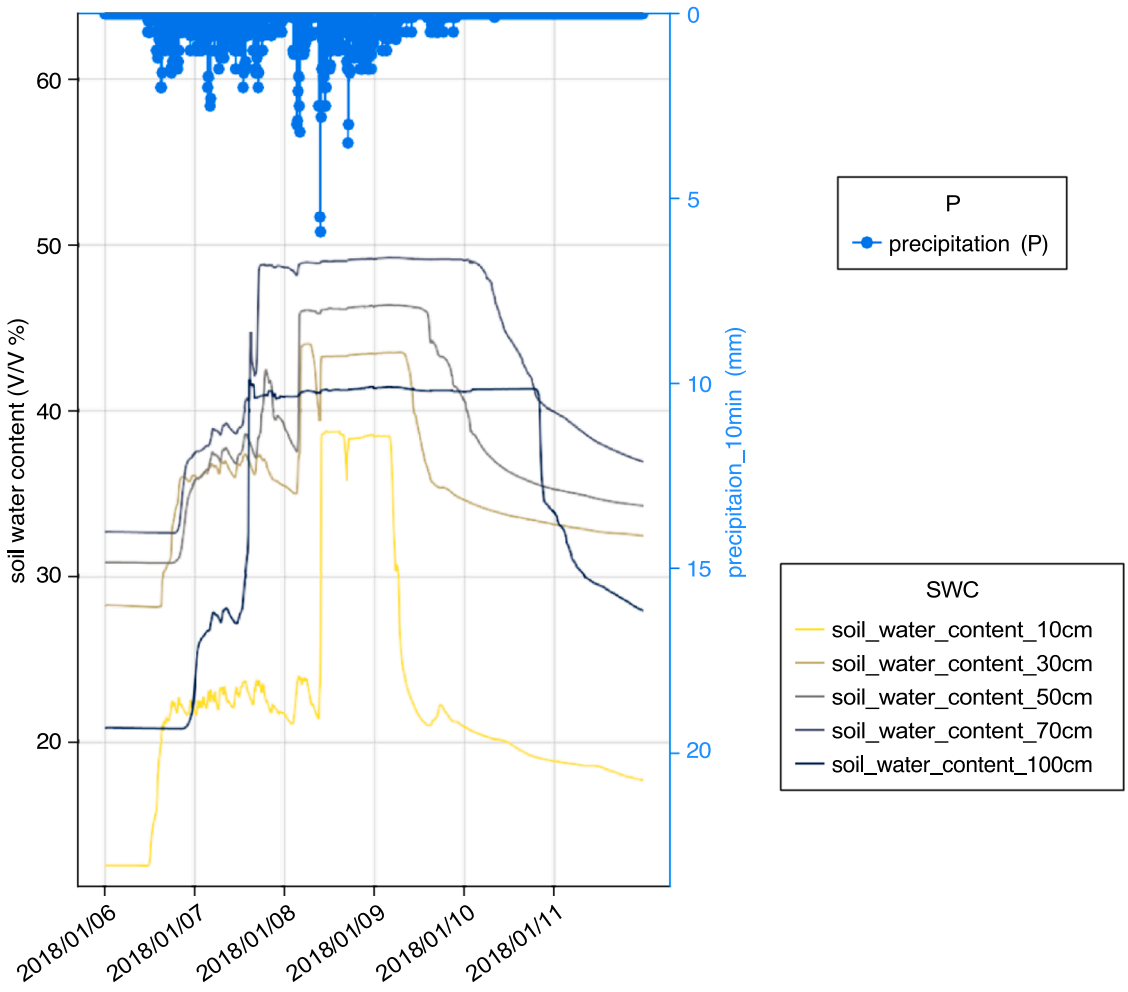


Fig. 11. SWC at five depths v.s. precipitation: an 82-hour 229.5 mm rainfall event.

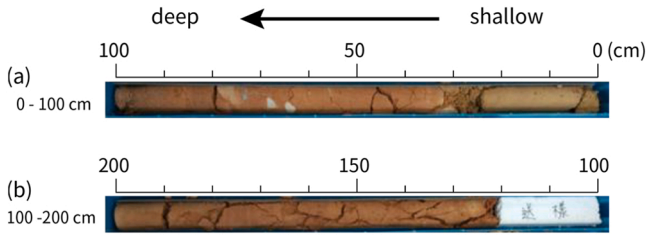


Fig. 12. Rock core samples obtained from the study area. Layer (a): fine-grained, yellow-brown with red-brown streaks clay soil; layer (b): fine-grained, yellow-brown clay soil. This figure is modified from (Tseng, 2019).

much longer than all except one rainfall event observed in 2018, which shows the rare situation that the significant and persistent changes in SWC at 100 cm is most likely close to that at 10 cm. For smaller rainfall or shorter rainfall duration, the 100 cm SWC has only little change in a limited time period, which explains that in almost all rainfall events only rainfall within 48 h significantly improves the estimation of SWC at 100 cm.

4.3. Insights into the role of AP and NP parameters in 12-fold TSCV

The fitting results of the 12-fold TSCV for all depths of SWC measurements with the time shift set "longer" are displayed in Fig. 13, and the corresponding performance estimated by MAE for each fold is displayed in Fig. 14. The "*n*th-fold" marked in the figures or mentioned in the article, directly refers to the time period of testing data of the *n*th fold validation. In Fig. 14, there is a clear increase in the loss of the 3rd fold validation for all models, and a drastic increase in loss of the 8th fold validation for models targeting SWC of depths ≥70 cm. In all other cases, the MAE is less than 4% (v/v). Referring to Fig. 13, discrepancies between the predicted and actual values during the 3rd and 8th fold are clearly observable. We attribute the misfit in the 3rd fold to the inability of CART to extrapolate outside the range of the training data. In CART, the leaf nodes are always summarized from the training data, and this nature makes CART intrinsically unable to provide predictions outside the range of the training dataset. From the second half of the 2nd fold to the end of the 3rd fold, the SWC continuously decreased and reached historically low levels that were not present in the training data for the 3rd fold. In addition, when comparing the MAE of the 2nd fold validation, we observed exceptionally worse performances for models targeting SWC depths ≥30 cm. This observation aligns with the fact that the lowest SWC in the 3rd fold validation for depths ≥30 cm is much lower than the minimum SWC ever reached before. For the misfit in the 8th fold validation, we speculate that a much lower responsiveness of SWC to non-intensive precipitation at deeper depths is the major cause. It can be clearly observed in the interval from the 3rd fold to the 7th fold that, as demonstrated in Fig. 13, the SWC at shallow depths (≤30 cm) responds promptly to precipitation, while the SWC at greater depths (≥70 cm) shows little to no response to precipitation.

To understand the role of precipitation in the 12-fold TSCV, feature importance of NP, P, and AP parameters are assessed for each fold using TDNI and TNNS as metrics, as demonstrated in Fig. 15. Both TDNI and

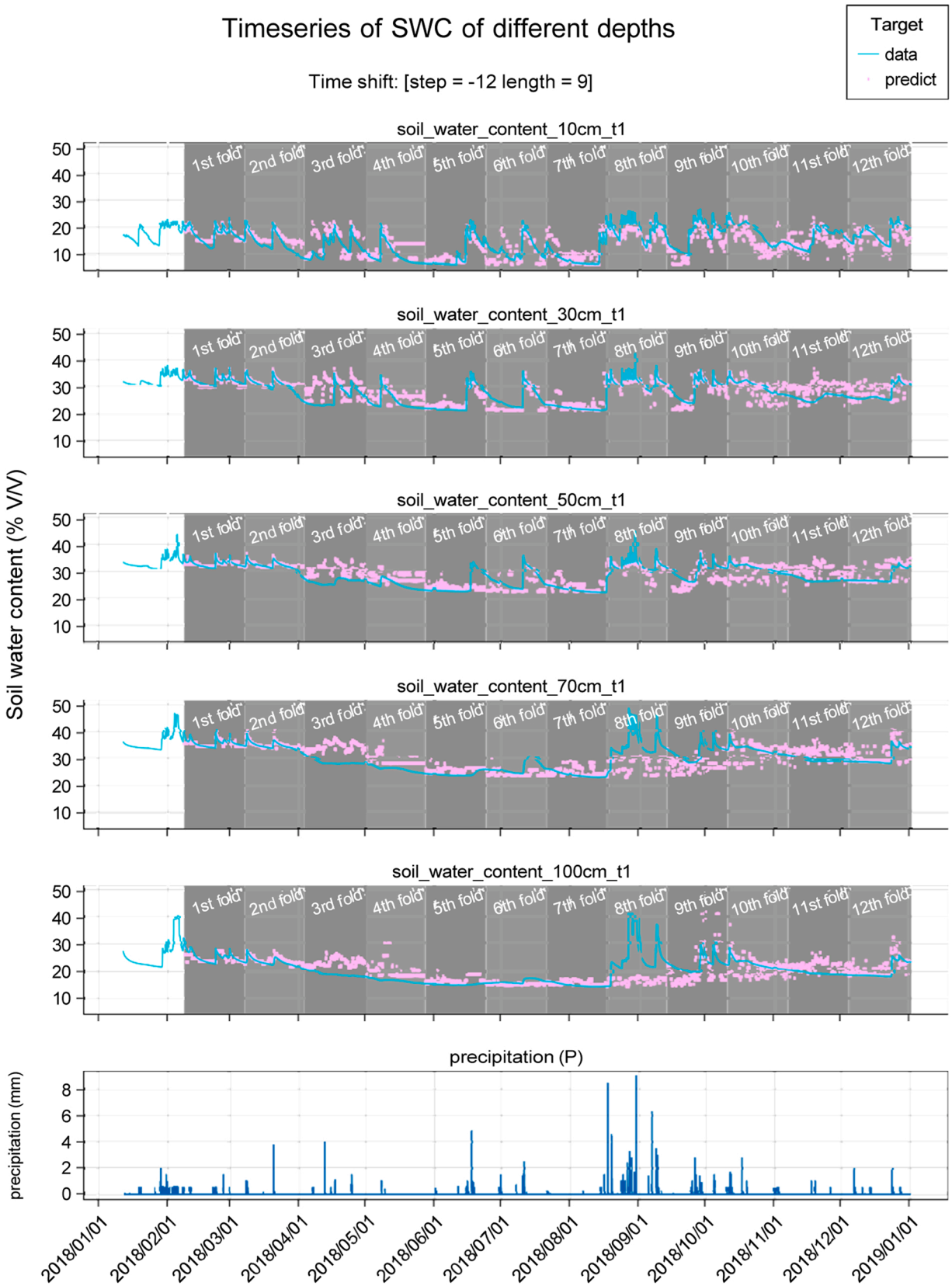


Fig. 13. Time-series of soil water content predictions in the testing phases of the 12 folds (scattered points), referencing actual observations (soild line); with results of soil water content of different depths separately demonstrated in each subplot. The models are trained using all features in Table 2 of time-shift set "longer".

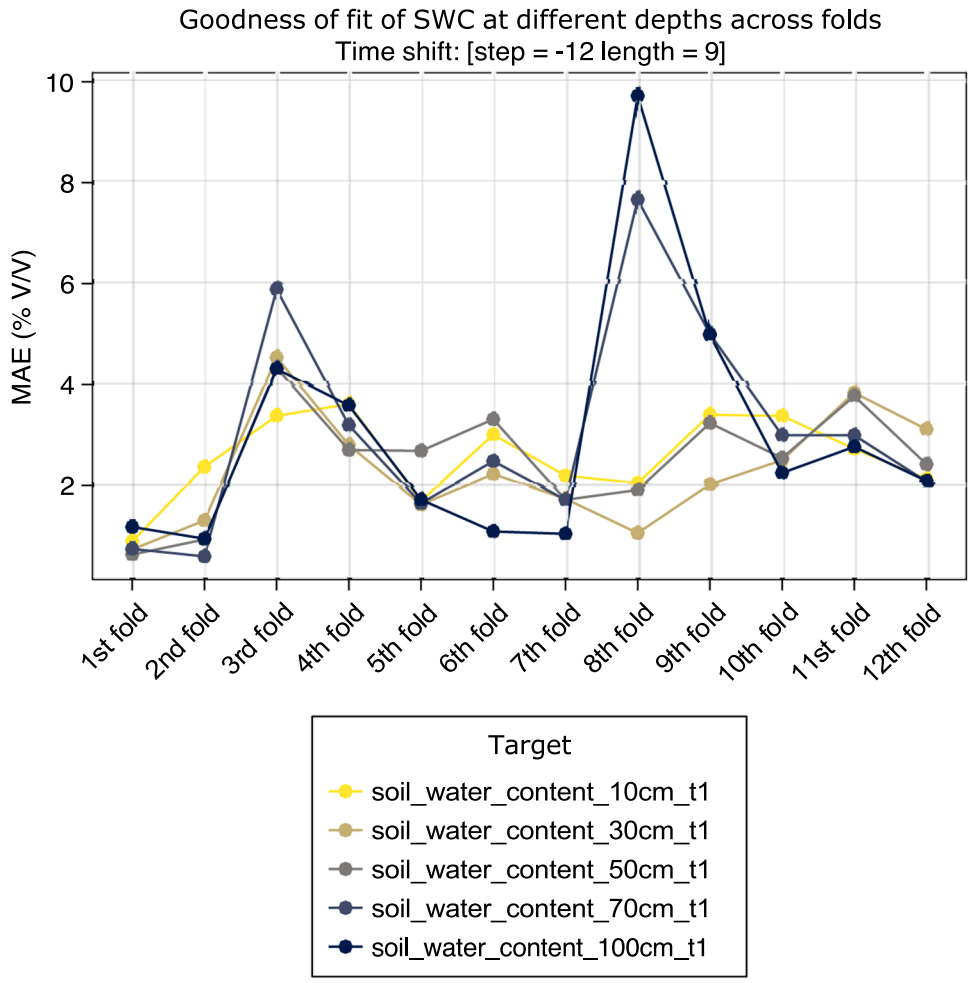


Fig. 14. Loss in MAE across 12 folds for all SWC depths, referring the fitting goodness of the results demonstrated in Fig. 13.

TNNS reflect the relevance between an input feature and the target variable but have different connotations, as explained in Sec. 3.4. In this study, the fundamental assumption and rule that the CART operates on is that the behavior of underground SWC can be solved through a series of binary decisions based on meteorological parameters. Based on this assumption, we use TDNI and TNNS to assess the responsiveness of SWC to the collection of NP parameters and to the collection of AP & P parameters. To make the results comparable, TDNI and TNNS are normalized for each CART model. It should be noted that the TDNI and TNNS for the ith fold are derived from the model trained using data from the beginning of 2018 to the end of the $(i-1)$th fold. Therefore, the TDNI and TNNS for the ith fold reflect the relevance of the input and target features observed up to but not including the time of the ith fold. In addition, the feature importance measured by TDNI and TNNS is relative, and the sum of both TDNI and TNNS for all input features is always 1. As a consequence, the summary effect of NP parameters in terms of TDNI or TNNS is always complementary to the summary effect of AP & P parameters in each fold. In Fig. 15, the TDNI of AP & P parameters for the first three folds is high for SWC at depths $\leq$50 cm and starts lowering after the 3rd fold until the 8th fold. For SWC at depths $\geq$70 cm, the lowering starts at the 2nd fold and reaches a significantly lower level around the 7th fold compared to the shallowest SWC. After the 8th fold, a rebound of the TDNI of AP & P parameters can be observed for SWC at every depth, and the TDNI remains stable from the 9th to the 12th fold. The TDNI pattern of AP & P parameters in the 12-fold TSCV indicates the responsiveness of SWC to precipitation between the 3rd and 7th fold, becoming lower for all depths and significantly lower for SWC at deeper depths ($\geq$70 cm). The rebound of TDNI for AP & P parameters at 9th fold is especially strong for SWC at the depth of 100 cm, indicating that the responsiveness of SWC at the deepest depth becomes prominent again during the 8th fold. The corresponding time of the data included in the training dataset that instigate these behaviors of TDNI of AP & P parameters coincides with the time in the year of 2018 from April to July when the rainfall events become less intensive, and the time of August when heavy rainfall events consecutively occurred. On the other hand, the variance of TNNS across different folds is trivial, and the TNNS of NP parameters is significantly higher than that of AP & P parameters for SWC at every depth and the model of every fold. For NP parameters, high TNNS with low TDNI indicates that the portion of decision nodes made by NP parameters in a tree is virtually the same, but these nodes are located at higher levels (near the leaf). This suggests that even though AP and P are the most crucial factors in affecting underground SWC during seasons with intensive rainfall, such as the time from mid-August to September, and the time from January to February, NP parameters can contribute to a finer prediction. Besides, the good performance of models in validating the 4th to 7th folds indicates that NP parameters are effective in providing accurate estimations of SWC during seasons with only non-intensive rainfall. Considering the seasonal variation of TDNI of SWC at deeper depths and the seasonal variation in ET mechanisms inferred by the two specific drying and wetting cases, as well as the rock core samples obtained at the site, we attribute the exceptionally large MAE of the 8th fold TSCV for SWC at deeper depths ($\geq$70 cm) to the sudden change in the responsiveness characteristics of SWC to precipitation caused by the intensive rainfall events occurring during the time of the 8th fold.

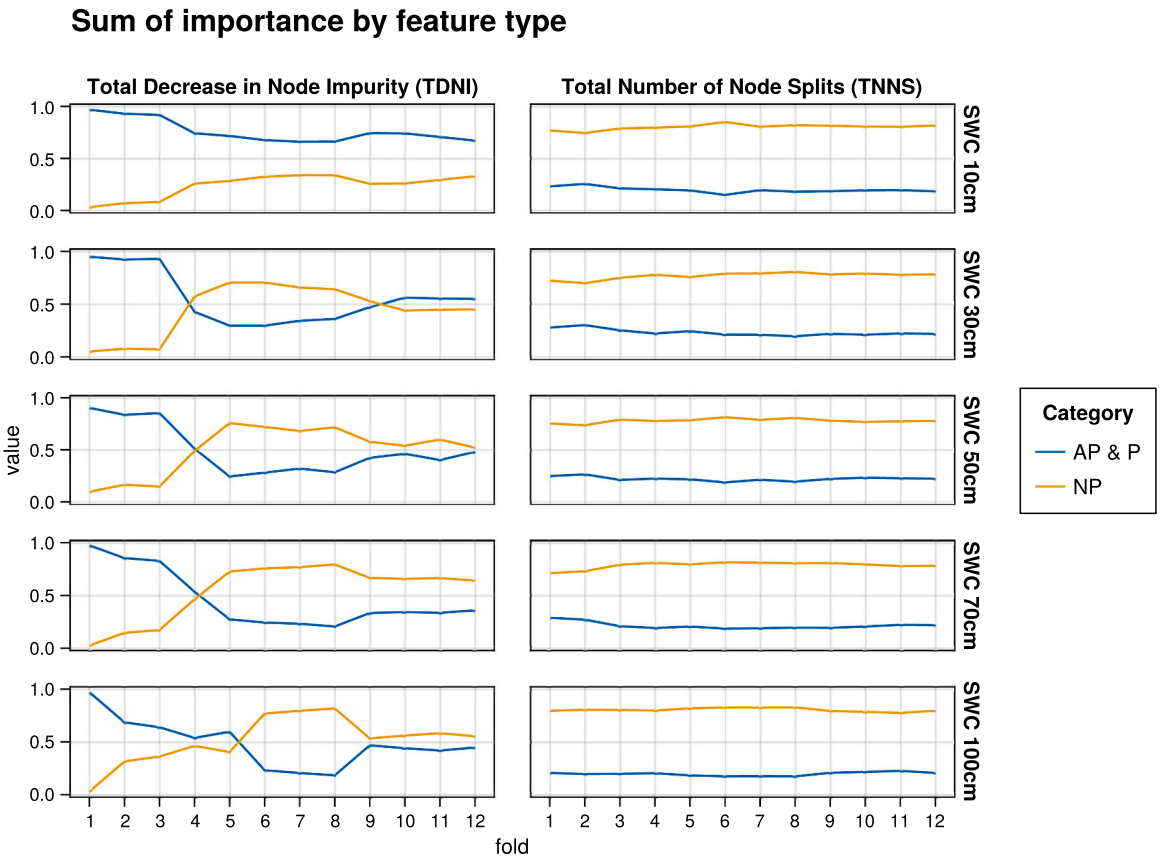


Fig. 15. The sum of feature importance by categories as denoted in the legend.

4.4. Assessing the impact of missing data

From January to March, there is a continuous time span of more than two months where humidity, pressure, and wind speed observations are missing, as shown in Fig. 2 (a). To understand how these missing parameters may affect predictions and validation results, we calculate TDNI and TNNS for the models during the 12-fold TSCV, as shown in Fig. 16. The results show that in the first three folds, the TDNI of humidity, pressure, and wind speed for almost all models is very low. This indicates that even when the input data for these parameters are inferred values using linear interpolation, they cannot result in substantial deviation in predicted SWC in the 1st and 2nd fold validation. The very low TDNI of the model in the 3rd fold validation further guarantees that the misfit of this fold is very unlikely due to the handling of missing data.

4.5. Seasonal variations in SWC mechanisms and practical implications for agriculture in Taiwan

In the 12-fold TSCV for the "baseline" case with *all parameters*, despite three models targeting SWC of depths 50, 70, 100 cm in the 3rd fold and four models targeting SWC of depths 70, 100 cm in the 8th and 9th fold, the remaining 53 models perform well in the validation, with MAE of less than 4 % (v/v), as demonstrated in Fig. 14. Considering that root depth is usually similar to the height of the vegetation, soil within 20 cm depth represented by SWC at 10 cm is the most of use for agricultural irrigation. With the three-day rainfall forecast more reliable than the weekly forecast in Taiwan, the CART model helps to produce hourly estimations of SWC within three days in practice for collecting water beforehand in case of possible drought events. Notice that the effects of error in rainfall forecasts require further studies. Moreover, water shortage within the tolerable range $15-20$ % of the total water demand barely influences crop production. Overestimating SWC by 4 % for SWC at 10 cm results in an 8 mm or around 7.2 tons of water deficit per 0.1 hectare farmland. 7.2 tons is 7.2 ‰ of water demand for rice per 0.1 hectare, while the water demands of soybeans and wheat are $\frac{1}{2}$ and $\frac{1}{10}$ of that for rice, respectively. It indicates that the continuous estimation error of more than 20 days may start to influence rice production, which manifests the value of the estimated SWC information in this study. However, crops may require specific water during certain growth periods. Therefore, the crop type and growth period will affect the tolerable accumulated days of SWC estimation error, which needs to be further clarified by future studies.

The misfit of the 8th-fold validation and the responsiveness of SWC inferred by TDNI can also be explained from hydrological aspects. The observed SWC at shallow layers ($\leq$50 cm) decreases to lower values in April than in previous months because of smaller rainfall ($6-22.5$ mm), less frequent rainfall events, and larger average solar radiation, providing more spaces in the soil due to ET for rainwater detention. Barely any hydrological response is found for 70 cm until mid-June and for 100 cm until mid-July because of small rainfall amounts and small hydraulic conductivity caused by low SWC, unlike high SWC in January. SWC at each layer is thus close to the wilting points before the long-lasting event in mid-August with large rainfall amounts. In contrast, in the other interval containing the mid-August event from the 8th to the 9th fold, the SWC at all depths shows a clear response to precipitation although the antecedent water content before the mid-August event is low. It can be attributed to the rainfall characteristics, including the number, the rainfall amounts, the rainfall duration, and the rainfall intensity of the rainfall events. The consecutive intense rainfall events increase SWC from a very dry condition to saturation, and the phenomenon unseen in the other folds of the data results in the misfitting tree model for the 8th fold. The hydraulic conductivity increases rapidly as well during the wetting process, which increases the transmissivity of water to the deep soil layer and therefore positively feedbacks to

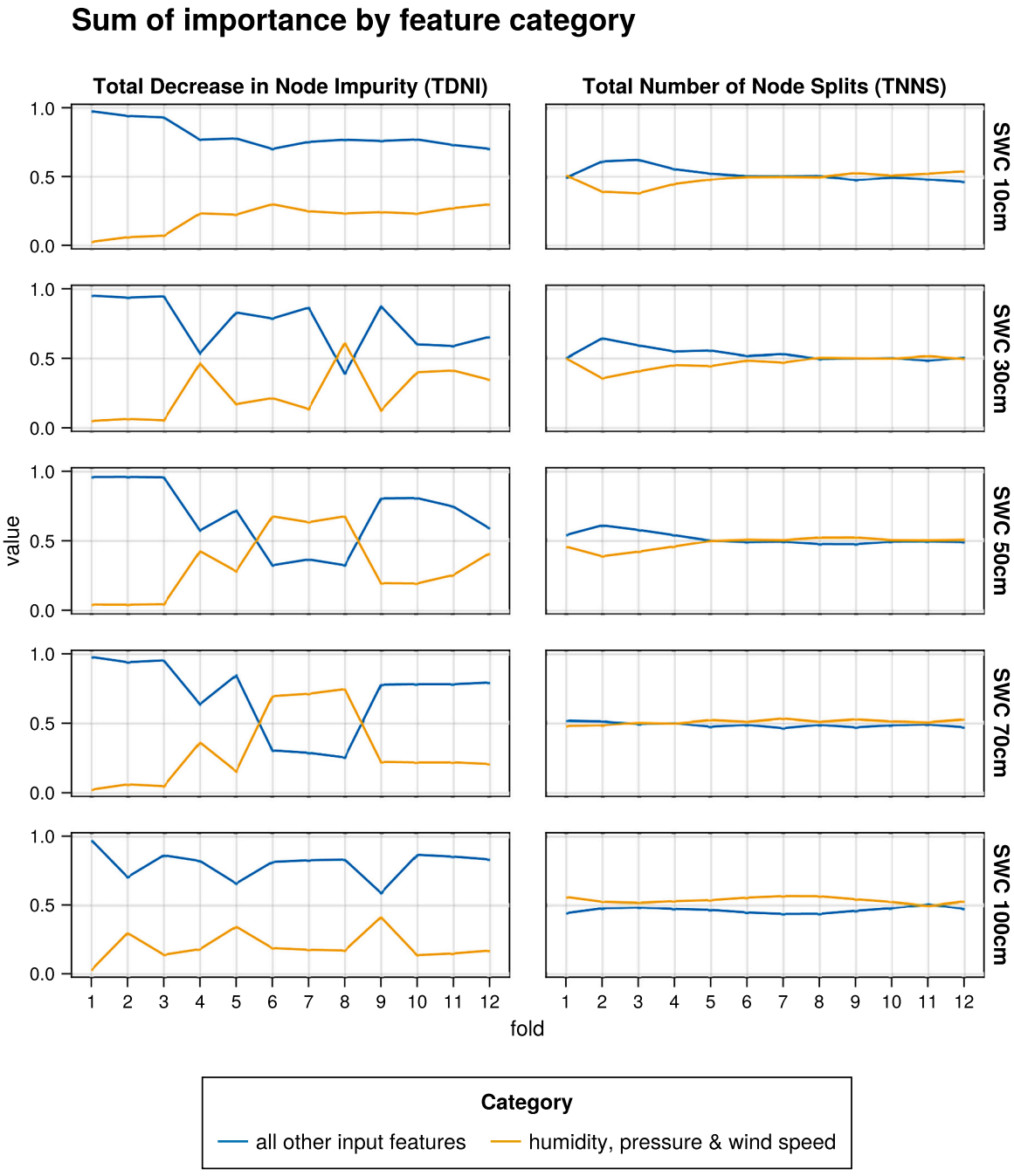


Fig. 16. The sum of feature importance by categories as denoted in the legend.

accelerate the increase of SWC to saturation.

Considering all the results above, it is implied that the mechanism of SWC response at this site changes seasonally due to changes in the rainfall and ET characteristics. In most drying and wetting processes, the influence of the rainfall and ET only reaches soil depths ≤50 cm, which may be partly due to the fine grains in the topsoil layer. It also explains why the model performs better for SWC ≤50 cm (Fig. 13), because more drying-wetting events are experienced in the shallow layers than in the deep layers and the model has more events to learn. From the results in Taoyuan with 2300 mm rainfall and 1030 mm evaporating-pan records annually shown in this study, it is suggested to have a general idea of the characteristics of rainfall, ET, and soil texture before applying CART to estimate SWC. In this case, the model's performance is explainable and precise information can be provided for agricultural irrigation or even flood early warning.

5. Conclusion

In this study, we assess the capability of CART in estimating SWC in a site of grassland terrain in northern Taiwan, and inspect the effectiveness of meteorological parameters in sets of different time shifting and parameter type to provide guidance for refining SWC estimations in agricultural settings. During the assessment of the time shifting effect, we found that including earlier meteorological information as input features can bring substantial reduction of error. However, using temporally more detailed information than the 12-hour resolution may not be necessary for hourly SWC estimation. When analyzing the learning curve of models that incrementally included data observed from earlier periods as input features, we observed two distinct patterns for SWC depths ≤50 cm and depths ≥70 cm. This difference aligns with the high water retention ability of the topsoil and the presence of

preferential flow in the mid-layer, which were inferred from soil characteristics of the rock core samples obtained at the site and SWC observations during a drying process in summer. In the performance analysis across 10 parameter selection sets, we found that for the depth of 10 cm, CART relied more on AP & P parameters, while for the depth of 100 cm, it relied more on NP parameters to deal with the inconsistency in the sensitivity of SWC to precipitation across dry and wet seasons. During the feature importance analysis, the variation of TDNI values of NP parameters and AP & P parameters revealed changes in the responsiveness of SWC to precipitation throughout the year. This explains the distinct strategies that the CART model developed for estimating SWC at the shallowest depth (10 cm) and the deepest depth (100 cm) at the study site. Furthermore, the consistently higher TNNS values associated with NP parameters suggest the efficacy of NP parameters in finer adjustments in making predictions. Additionally, the CART algorithm exhibits resilience in the face of superfluous input meteorological data up to 192 h before the current time, as demonstrated in the time shifting effect analysis. It also exhibited robustness against missing data, as shown in Sec. 4.4.

The mechanism of SWC response at this site changes seasonally, and thus influences the performance of the CART model. More specifically, the response of SWC in shallow layers is frequent during both rainy and dry seasons. On the other hand, saturation of soil at a depth larger than 70 cm, which occurs earlier than that of soil at a depth smaller than 50 cm because of the fine grains in the topsoil layer, only occurs for long-duration large-intensity precipitation in the rainy season. Overall, these results not only provide hints about the near-surface soil structure of the study area but also have implications for better estimation of SWC, which is not exclusively for the CART model.

CART is the commonly chosen weak learner for state-of-the-art DT methods. The methodology and insights on feature selection for SWC estimation presented in this study can be easily applied when using these advanced DT techniques. Furthermore, since CART is naturally compatible with heterogeneous datasets, this study can serve as the basis for future works that apply CART-based DT algorithms using meteorological data in combination with other variables such as land use category. However, it should be noted, as encountered and discussed in this study, that CART has limitations in its ability to extrapolate outside the training range. Additionally, the fundamental assumption of CART in our application of SWC estimation, as detailed in Sec. 3, is reasonable but has not been thoroughly examined. Hence, caution with these limitations must be exercised in conducting future works based on this study.

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.

Data Availability

The authors do not have permission to share data.

Acknowledgement

The authors of the manuscript gratefully acknowledge the laboratories providing valuable data, namely the Surface Hydrology Lab., Aerosol and Cloud Lab., and Planetary Boundary Layer and Air Pollution Lab. of National Central University. The authors also would like to express their gratitude to the National Science and Technology Council (R.O.C) for providing funding for this research through grant number MOST 110–2634-F-008–008 and MOST 111–2634-F-008–001. Our work is assisted by ChatGPT of OpenAI and Bard of Google for English editing and grammar verification purposes. The use of these tools is limited to improving readability and language, and the authors revised the content as needed and take full responsibility for the content of the publication.

References

Adab, H., Morbidelli, R., Saltalippi, C., Moradian, M., Ghalhari, G.A.F., 2020. Machine learning to estimate surface soil moisture from remote sensing data. Water 12, 3223. https://doi.org/10.3390/w12113223.

Aerosol and Cloud Laboratory of NCU, 2024.Aerosol and Cloud Laboratory.⟨http://aerosol.atm.ncu.edu.tw⟩.

Aldaba, A.L., Lopez-Torres, D., Campo-Bescós, M.A., López, J.J., Yerro, D., Elosua, C., Arregui, F.J., Auguste, J.L., Jamier, R., Roy, P., López-Amo, M., 2018. Microstructured optical fiber sensor for soil moisture measurements, In: 26th International Conference on Optical Fiber Sensors (2018), Paper WF41, Optica Publishing Group. WF41.10.1364/OFS.2018.WF41.

An, X., Zhao, F., 2021. Prediction of soil moisture based on BP neural network optimized search algorithm. IOP Conf. Ser.: Earth Environ. Sci. 714, 022046 https://doi.org/10.1088/1755-1315/714/2/022046.

Angelaki, A., Bota, V., Chalkidis, I., 2023. Estimation of hydraulic parameters from the soil water characteristic curve. Sustainability 15, 6714. https://doi.org/10.3390/su15086714.

Araya, S.N., Fryjoff-Hung, A., Anderson, A., Viers, J.H., Ghezzehei, T.A., 2020. Machine Learning Based Soil Moisture Retrieval from Unmanned Aircraft System Multispectral Remote Sensing, In: IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, 4598-4601. 10.1109/IGARSS39084.2020.9324117.

Bentéjac, C., Csörgő, A., Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. Artif. Intell. Rev. 54, 1937–1967. https://doi.org/10.1007/s10462-020-09896-5.

Bittelli, M., 2010. Measuring soil water potential for water management in agriculture: a review. Sustainability 2, 1226–1251. https://doi.org/10.3390/su2051226.

Blaom, A.D., Kiraly, F., Lienart, T., Simillides, Y., Arenas, D., Vollmer, S.J., 2020. MLJ: a julia package for composable machine learning. J. Open Source Softw. 5, 2704. https://doi.org/10.21105/joss.02704.

Breiman, L., 2017. Classification and Regression Trees, First ed. Routledge. https://doi.org/10.1201/9781315139470.

Brunet, P., Clément, R., Bouvier, C., 2010. Monitoring soil water content and deficit using electrical resistivity tomography (ert) - a case study in the cevennes area, france. J. Hydrol. 380, 146–153. ⟨https://api.semanticscholar.org/CorpusID:128434141⟩.

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. https://doi.org/10.1371/journal.pone.0214508.

Carrizosa, E., Molero-Río, C., Romero Morales, D., 2021. Mathematical optimization in classification and regression trees. TOP 29, 5–33. https://doi.org/10.1007/s11750-021-00594-1.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Cosenza, P., 2016.Indirect determination of soil water content.⟨https://api.semanticscholar.org/CorpusID:3932194⟩.

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

Dawson, C.B., Day-Lewis, F.D., Johnson, C.D., Lane, J.W., Robinson, J., Slater, L.D., 2017.Borehole nuclear magnetic resonance (nmr): a valuable tool for environmental site management.⟨https://api.semanticscholar.org/CorpusID:135163659⟩.

Garg, A., Gadi, V.K., Feng, Y.C., Lin, P., Qinhua, W., Ganesan, S., Mei, G., 2020. Dynamics of soil water content using field monitoring and AI: a case study of a vegetated soil in an urban environment in China. Sustain. Comput.: Inform. Syst. 28 https://doi.org/10.1016/j.suscom.2019.01.003.

Grabczewski, K., Jankowski, N., 2005.Feature selection with decision tree criterion, In: Fifth International Conference on Hybrid Intelligent Systems (HIS'05), 6 pp.–. 10.1109/ICHIS.2005.43.

Greifeneder, F., Notarnicola, C., Wagner, W., 2021. A machine learning-based approach for surface soil moisture estimations with google earth engine. Remote Sens. 13, 2099. https://doi.org/10.3390/rs13112099.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-84858-7.

He, H., Aogu, K., Li, M., Xu, J., Sheng, W., Jones, S.B., González-Teruel, J.D., Robinson, D.A., Horton, R., Bristow, K., Dyck, M., Filipović, V., Noborio, K., Wu, Q., Jin, H., Feng, H., Si, B., Lv, J., 2021.Chapter three - a review of time domain reflectometry (tdr) applications in porous media, Academic Press. volume 168 of Advances in Agronomy, 83-155.⟨https://www.sciencedirect.com/science/article/pii/S0065211321000316⟩, 10.1016/bs.agron.2021.02.003.

J. Hendricks, T. Yao, A. Kearns, 1999.Nuclear Magnetic Resonance Imaging of Water Content in the Subsurface.Technical Report DOE/ER/14732, 750970. 10.2172/750970.

Huang, Y., Jiang, H., Wang, W.f., Wang, W., Sun, D., 2021. Soil moisture content prediction model for tea plantations based on SVM optimised by the bald eagle search algorithm. Cogn. Comput. Syst. 3, 351–360. https://doi.org/10.1049/ccs2.12034.

Jena, M., Dehuri, S., 2020. Decisiontree for classification and regression: a state-of-the art review. Informatica 44. https://doi.org/10.31449/inf.v44i4.3023.
Kadra, A., Lindauer, M., Hutter, F., Grabocka, J., 2021. Well-tuned simple nets excel on tabular datasets. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (Eds.), Advances in Neural Information Processing Systems. ⟨https://openreview.net/forum?id=d3k38LTDCyO⟩.
Kroes, J., AU, Bartholomeus, R., Groenendijk, P., Heinen, M., Supit, I., Walsum, P., 2017. Swap Version 4: Theory Description and User Manual.
Lei, G., Zeng, W., Yu, J., Huang, J., 2023. A comparison of physical-based and machine learning modeling for soil salt dynamics in crop fields. Agric. Water Manag. 277, 108115 https://doi.org/10.1016/j.agwat.2022.108115.
Leij, F.J., Dane, J.H., Sciortino, A., 2022. Hierarchical prediction of soil water content time series. CATENA 209, 105841. https://doi.org/10.1016/j.catena.2021.105841.
Leonarduzzi, E., Tran, H., Bansal, V., Hull, R.B., De la Fuente, L., Bearup, L.A., Melchior, P., Condon, L.E., Maxwell, R.M., 2022. Training machine learning with physics-based simulations to predict 2D soil moisture fields in a changing climate. Front. Water 4.
Li, M., Tseng, I., 2024.Linking evapotranspiration and infiltration dynamics to variations in soil moistures and changes in groundwater levels, In: Proceedings of the European Geosciences Union (EGU), EGU2019-3720, 207-210.
Li, P., Zha, Y., Shi, L., Tso, C.H.M., Zhang, Y., Zeng, W., 2020. Comparison of the use of a physical-based model with data assimilation and machine learning methods for simulating soil water dynamics. J. Hydrol. 584, 124692 https://doi.org/10.1016/j.jhydrol.2020.124692.
Liu, Y., Yang, Y., Jing, W., Yue, X., 2018. Comparison of different machine learning approaches for monthly satellite-based soil moisture downscaling over Northeast China. Remote Sens. 10, 31. https://doi.org/10.3390/rs10010031.
Maroufpoor, S., Maroufpoor, E., Bozorg-Haddad, O., Shiri, J., MundherYaseen, Z., 2019. Soil moisture simulation using hybrid artificial intelligent model: Hybridization of adaptive neuro fuzzy inference system with grey wolf optimizer algorithm. J. Hydrol. 575, 544–556. https://doi.org/10.1016/j.jhydrol.2019.05.045.
National Central University, NCU, 2024.Atmospheric and Hydrological Observatory.⟨http://naho.atm.ncu.edu.tw/Home.html⟩.
Nguyen, T.T., Ngo, H.H., Guo, W., Chang, S.W., Nguyen, D.D., Nguyen, C.T., Zhang, J., Liang, S., Bui, X.T., Hoang, N.B., 2022. A low-cost approach for soil moisture prediction using multi-sensor data and machine learning algorithm. Sci. Total Environ. 833, 155066 https://doi.org/10.1016/j.scitotenv.2022.155066. ⟨https://www.sciencedirect.com/science/article/pii/S0048969722021593⟩.
Oliveira, V.A., Rodrigues, A., Morais, M.A.V., Terra, M., Guo, L., Mello, C.R., 2021. Spatiotemporal modelling of soil moisture in an atlantic forest through machine learning algorithms. Eur. J. Soil Sci. 72, 1969–1987. https://doi.org/10.1111/ejss.13123.
Pachepsky, Y.A., van Genuchten, M.T., 2011.Pedotransfer Functions.⟨https://api.semanticscholar.org/CorpusID:222010432⟩.
Padarian, J., Minasny, B., McBratney, A.B., 2020. Machine learning and soil sciences: a review aided by machine learning tools. SOIL 6, 35–52. https://doi.org/10.5194/soil-6-35-2020.
Patil, N.G., Singh, S.K., 2016. Pedotransfer functions for estimating soil hydraulic properties: A review. Pedosphere 26, 417–430. https://doi.org/10.1016/S1002-0160(15)60054-6. ⟨https://www.sciencedirect.com/science/article/pii/S1002016015600546⟩.
Pekel, E., 2020. Estimation of soil moisture using decision tree regression. Theor. Appl. Climatol. 139, 1111–1119. https://doi.org/10.1007/s00704-019-03048-8.
Peters, R.T., Desta, K.G., Nelson, L., 2013. Practical use of soil moisture sensors and their data for irrigation scheduling. Fact Sheet FS083E. Washington State University Extension. ⟨https://api.semanticscholar.org/CorpusID:131978473⟩.
Pinheiro, E.A.R., de Jong van Lier, Q., Šimúnek, J., 2019. The role of soil hydraulic properties in crop water use efficiency: a process-based analysis for some brazilian scenarios. Agric. Syst. 173, 364–377. https://doi.org/10.1016/j.agsy.2019.03.019. ⟨https://www.sciencedirect.com/science/article/pii/S0308521×18314227⟩.
Planetary Boundary Layer and Air Pollution Lab. of NCU, 2024.Planetary Boundary Layer and Air Pollution Lab.⟨http://pblap.atm.ncu.edu.tw/⟩.
Questier, F., Put, R., Coomans, D., Walczak, B., Heyden, Y.V., 2005. The use of CART and multivariate regression trees for supervised and unsupervised feature selection. Chemom. Intell. Lab. Syst. 76, 45–54. https://doi.org/10.1016/j.chemolab.2004.09.003.
Reynolds, S.G., 1970. The gravimetric method of soil moisture determination Part I A study of equipment, and methodological problems. J. Hydrol. 11, 258–273. https://doi.org/10.1016/0022-1694(70)90066-1.
Robinson, D.A., Jones, S.B., Wraith, J.M., Or, D., Friedman, S.P., 2003. A review of advances in dielectric and electrical conductivity measurement in soils using time domain reflectometry. Vadose Zone J. 2, 444–475. https://doi.org/10.2113/2.4.444. ⟨https://pubs.geoscienceworld.org/vzj/article-pdf/2/4/444/2990507/444.pdf⟩.
Sadeghi, B., Chiarawongse, P., Squire, K., Jones, D.C., Noack, A., St-Jean, C., Huijzer, R., Schätzle, R., Butterworth, I., Peng, Y.F., Blaom, A., 2022.DecisionTree.jl - A Julia implementation of the CART Decision Tree and Random Forest algorithms.10.5281/zenodo.7359268.
Shekhar, S., Kumar, M., Kumari, A., Jain, S.K., 2017. Soil moisture profile analysis using tensiometer under different discharge rates of drip emitter. Int. J. Curr. Microbiol. Appl. Sci. 6, 908–917. https://doi.org/10.20546/ijcmas.2017.611.106.
Shen, S., Guo, P., Wu, J., Ding, Y., Chen, F., Meng, F., Xu, Z., 2019. Optimized inside-out magnetic resonance probe for soil moisture measuring in situ. J. Magn. Reson. 307, 106565 https://doi.org/10.1016/j.jmr.2019.07.052. ⟨https://www.sciencedirect.com/science/article/pii/S109078071930182X⟩.
Simunek, Jirka, J., Van Genuchten, M., Šejna, M., 2016. Recent developments and applications of the HYDRUS computer software packages. Vadose Zone J. 6. https://doi.org/10.2136/vzj2016.04.0033.
Sobayo, R., Wu, H.H., Ray, R., Qian, L., 2018.Integration of convolutional neural network and thermal images into soil moisture estimation, In: 2018 1st International Conference on Data Intelligence and Security (ICDIS), IEEE.207-210.
Sun, M.Y., Shi, B., Guo, J.Y., Zhu, H.H., Jiang, H.T., Liu, J., Wei, G.Q., Zheng, X., 2022. Development and application of fiber-optic sensing technology for monitoring soil moisture field. Front. Sens. 2.
Surface Hydrology Lab. of NCU, 2024a.Aerial view of NCU Atmospheric and Hydrological Observatory.⟨https://hydro.ihs.ncu.edu.tw/⟩.
Surface Hydrology Lab. of NCU, 2024b.Surface hydrology lab.⟨https://hydro.ihs.ncu.edu.tw/⟩.
Toková, L., Igaz, D., Aydín, E., 2019. Measurement of volumetric water content by gravimetric and time domain reflectometry methods at field experiment with biochar and n fertilizer. Acta Hortic. Et. Regiotect. 22, 61–64. ⟨https://api.semanticscholar.org/CorpusID:207961140⟩.
Tramblay, Y., QuintanaSeguí, 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, 1325–1334. https://doi.org/10.5194/nhess-22-1325-2022.
Tseng, I., 2019.Effects of evapotranspiration and infiltration on variations in soil moisture and changes in groundwater levels.National Digital Library of Theses and Dissertations in Taiwan ⟨https://hdl.handle.net/11296/q37cpq⟩.
Wösten, J., Lilly, A., Nemes, A., Le Bas, C., 1999. Development and use of a database of hydraulic properties of european soils. Geoderma 90, 169–185. https://doi.org/10.1016/S0016-7061(98)00132-3. ⟨https://www.sciencedirect.com/science/article/pii/S0016706198001323⟩.
Wösten, J., Pachepsky, Y., Rawls, W., 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. J. Hydrol. 251, 123–150. https://doi.org/10.1016/S0022-1694(01)00464-4. ⟨https://www.sciencedirect.com/science/article/pii/S0022169401004644⟩.
Yang, Y., Mei, G., 2022. A deep learning-based approach for a numerical investigation of soil– water vertical infiltration with physics-informed neural networks. Mathematics 10, 2945. https://doi.org/10.3390/math10162945.
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. https://doi.org/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 https://doi.org/10.1016/j.agwat.2020.106649.
Zhou, H., Zhang, J., Zhou, Y., Guo, X., Ma, Y., 2021. A feature selection algorithm of decision tree based on feature weight. Expert Syst. Appl. 164, 113842 https://doi.org/10.1016/j.eswa.2020.113842.