

## Simulating and predicting soil water dynamics using three models for the Taihu Lake region of China

Can Chen , Qing Lv  and Qian Tang 

College of Applied Meteorology, Nanjing University of Information Science and Technology, Nanjing 210044, China

\*Corresponding author. E-mail: chencan2010203@sohu.com

 CC, 0000-0003-3517-5567; QL, 0000-0003-3751-1456; QT, 0000-0002-9445-6599

### ABSTRACT

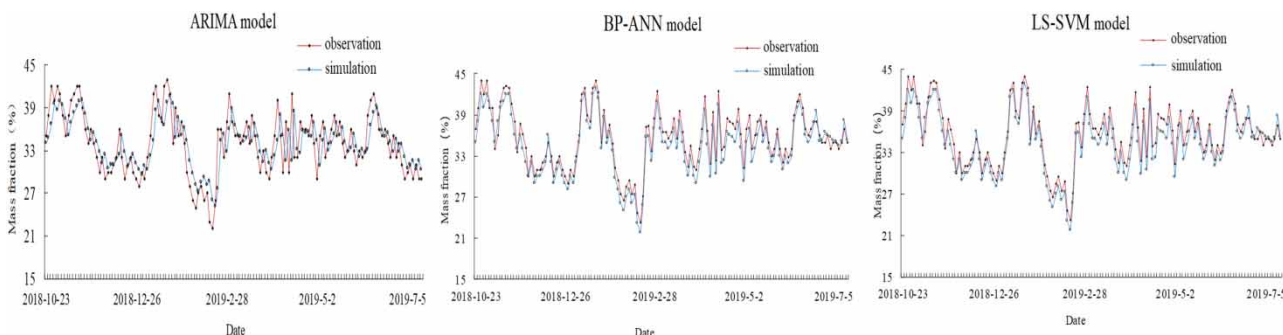
Drought stress under a changing climate can significantly affect agricultural production. Simulation of soil water dynamics in field conditions becomes necessary to understand changes of soil water conditions to develop irrigation guidelines. In this study, three models including Auto-Regressive Integrated Moving Average (ARIMA), Back-Propagation Artificial Neural Network (BP-ANN), and Least Squares Support Vector Machine (LS-SVM) were used to simulate the soil water content in the 0–14 cm and 14–33 cm soil layers across the Taihu Lake region of China. Rainfall, evaporation, temperature, humidity and wind speed that affect soil water content changes were considered in the BP-ANN and LS-SVM, but not in ARIMA. The results showed that the variability of soil water content in the 0–14 cm soil layer was greater than that in 14–33 cm. Correlation coefficients ( $r$ ) of soil water content between simulations and observations were highest (0.9827) using LS-SVM in the 14–33 cm soil layer, while they were the lowest (0.7019) using ARIMA in the 0–14 cm soil layer; but no significant difference in  $r$  values was observed between the two soil layers with the BP-ANN model. Compared with the other two models, the LS-SVM model seems to be more accurate for forecasting the dynamics of soil moisture. The results suggested that agro-climatic data can be used to predict the severity of soil drought stress and provide guidance for irrigation to increase crop production in the Taihu Lake region of China.

**Key words:** ARIMA model, BP-ANN model, LS-SVM model, simulating and predicting, soil water dynamics, Taihu Lake region

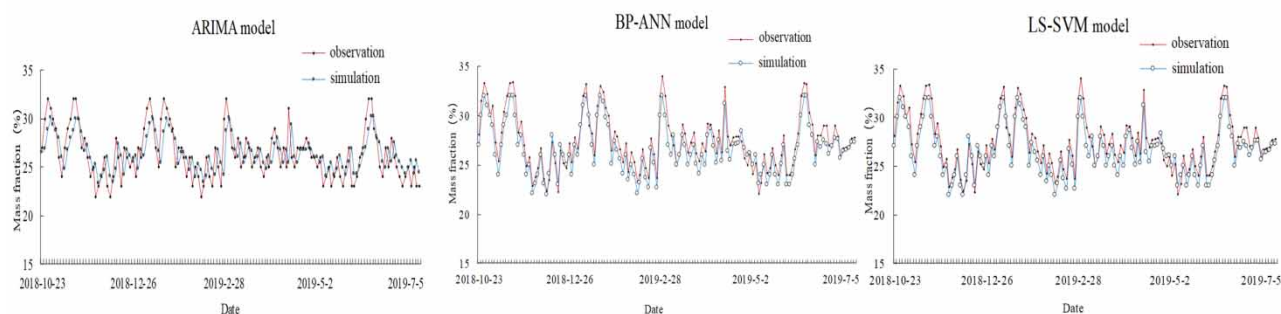
### HIGHLIGHTS

- To understand the dynamics of soil water in the Taihu Lake region of China.
- The simulation accuracy of LS-SVM was the highest.
- To predict the trend of soil water content in the study area.

## GRAPHICAL ABSTRACT



Daily soil water changes prediction by the three model vs. Measurements in 0-14 cm



Daily soil water changes prediction by the three model vs. Measurements in 14-33 cm

## 1. INTRODUCTION

Soil water content is one of the key factors that influence agricultural production. Soil water dynamics depend heavily on local environmental conditions, especially meteorological factors. In recent years, scientists have focused attention on soil water dynamics (Bialkowski & Buttle 2015; Vereecken *et al.* 2015; Bohara *et al.* 2019; Bai *et al.* 2020) and their research methods can generally be grouped into two categories. One is the statistical forecasting method (or empirical modeling) based on correlations between soil water content and weather parameters (Zucco *et al.* 2014). The other is the theoretical modeling approach based on the soil water balance equation and soil water dynamics (Kashani *et al.* 2020). These methods obtain satisfactory results in most cases when all necessary boundary conditions are properly processed. However, there are problems in real application. Values of the various parameters used in these models must be determined through experimental measurement or statistical analysis. Generally, theoretical soil water prediction models are too complicated to apply because of the difficulty of acquiring all parameter values.

The parameter values that are used for predictive models can dynamically change with natural conditions and need to be numerically determined through experimental measurements or statistical inferences. Natural conditions' variability over temporal and spatial horizons makes parameterization extremely difficult, thus impeding the effective use of models. In addition, comprehensive and complex models' outcomes are extremely sensitive to parameter definition. If, on the other hand, the models are composed of easily defined parameters, the simplified versions lack the flexibility to be comprehensive and universally applicable.

Recently various stochastic models have been developed and applied in water resources and hydrology, including the soil water simulations by Back-Propagation Artificial Neural Network (BP-ANN), Auto-Regressive Integrated Moving Average (ARIMA), and Least Squares Support Vector Machine (LS-SVM) (Aitkenhead & Coull 2016; Mojid *et al.* 2019; Asquith 2020). The BP-ANN, ARIMA, and LS-SVM techniques are widely used in water quality prediction and estimation of water demand growth for various purposes (Parmar & Bhardwaj 2015; Zounemat-Kermani *et al.* 2016; Tiyasha *et al.*

2020). These studies indicated that BP-ANN, ARIMA, and LS-SVM could successfully model the complex relationship between meteorological factors and soil water in agriculture.

The BP-ANN is a mathematics-based model, whose architecture has been inspired by biological neural networks (Almomani 2020; Jin *et al.* 2021). The BP-ANN is very appropriate for modeling nonlinear processes, and is particularly suitable for soil water and salt dynamics (Morshed & Kaluarachchi 1998). The ARIMA model is popular because of its simplicity and robust statistical properties (Hanh *et al.* 2010). The ARIMA is a linear prediction model which assumes that, the current data has a direct relationship with the past data and its errors (Narayanan *et al.* 2013). In hydrological studies, the ARIMA model has been used for forecasting monthly temperature, humidity and precipitation (Jahanbakhsh & Babapour 2003), stream flow data and prediction of soil water dynamics (Panda & Kumar 2011; Patle *et al.* 2015). LS-SVM is a good tool for system identification especially from I/O data or shortly learning from data. LS-SVM is an emerging modeling technique which combines the advantages of neural networks (handling large amounts of highly nonlinear data) and nonlinear regression (high generalization) (Suykens *et al.* 2001; Fan *et al.* 2008). Yu *et al.* (2017) showed that the LS-SVM is an appropriate and high-performance data-driven model for prediction of daily runoff, compared with neural network models.

The Taihu Lake region is one of China's most intensive agricultural regions, with a long history of cultivation. Under a changing climate, the frequency of extreme weather events has increased and drought has become the foremost natural disaster for agricultural production (Nguyen *et al.* 2018). During drought times, understanding of soil water dynamics is critical to developing irrigation schedules and guaranteeing agricultural production.

The objective of this study was to understand the dynamics of soil water, use ARIMA, BP-ANN and LS-SVM models to predict the trend of soil water content, and to determine the most suitable simulation model in the Taihu Lake region of China.

## 2. MATERIALS AND METHODS

### 2.1. Experimental site, soil and meteorological data collection

The experimental area is located at Changshu City, southeast of Jiangsu Province of China (31°30'N, 120°33'E). The climate is classified as subtropical monsoon with an annual precipitation of 1,100–1,200 mm, annual average temperature of 16 °C, annual sunshine greater than 2,000 hours, and frost-free period of more than 230 days. The dominant cropping rotation is rice–wheat. The major soil type is loam (local name Wushan soil).

The experimental platform consisted of nine Free Air Carbon Dioxide Enrichment (FACE) circulation systems and one control. The spatial interval between circulation systems was 20 m. The experiments were designed with two CO<sub>2</sub> levels: ambient CO<sub>2</sub> (AC); elevated CO<sub>2</sub>: ambient + 200 µmol L<sup>-1</sup> (EC), and two temperature levels: ambient temperature (AT) and elevated temperature: ambient +2 °C (ET). There are four treatments in this experiment: ambient (AC and AT), C (EC and AT), CT (EC and ET), and T (AC and ET), and each treatment had three replicates. Experimental plots were arranged in a randomized complete block design.

Soil water content changes were monitored from October 23, 2018, to July 15, 2019, and soil samples were collected every two days at depths of 0–14 cm and 14–33 cm. The water content of the soil samples was determined by drying (105–110 °C for eight hours in an oven).

Soil bulk density and saturated conductivity were measured by the ring knife method. The pH was determined in a 1:5 soil: water suspension using a Thermo Orion pH meter with a combination electrode. Soil organic matter was determined by potassium dichromate external heating (Bao 2000). Soil clay, sand and silt content were measured by the pipette method. Physical and chemical properties of the soil were determined (Table 1).

**Table 1** | The physical and chemical properties of the tested soil

| Soil layer  | Depth (cm) | Saturated hydraulic conductivity (cm·s <sup>-1</sup> ) | Soil organic matter (%) | Bulk density (g·cm <sup>-3</sup> ) | pH  | Porosity (%) | Mechanical composition (g·kg <sup>-1</sup> ) |               |            |
|-------------|------------|--|-------------------------|------------------------------------|-----|--------------|--|---------------|------------|
|             |            |  |                         |                                    |     |              | > 0.02 mm                                    | 0.02–0.002 mm | < 0.002 mm |
| Cultivation | 0–14       | $7.04 \times 10^{-4}$                                  | 3.72                    | 1.21                               | 7.0 | 54.34        | 337.42                                       | 386.23        | 276.35     |
| Plowpan     | 14–33      | $1.26 \times 10^{-4}$                                  | 2.91                    | 1.47                               | 7.2 | 44.53        | 278.62                                       | 392.44        | 328.94     |

Meteorological data including daily rainfall, average daily temperature, daily maximum and minimum temperatures, daily cumulative evaporation, daily average relative humidity, daily average wind speed, and daily average land surface temperature were collected from Changshu weather station.

A total of 134 sets of meteorological parameters and soil water contents were divided into two groups. Group A contained 120 sets which were used for model training, while group B comprised 14 sets used for model simulation.

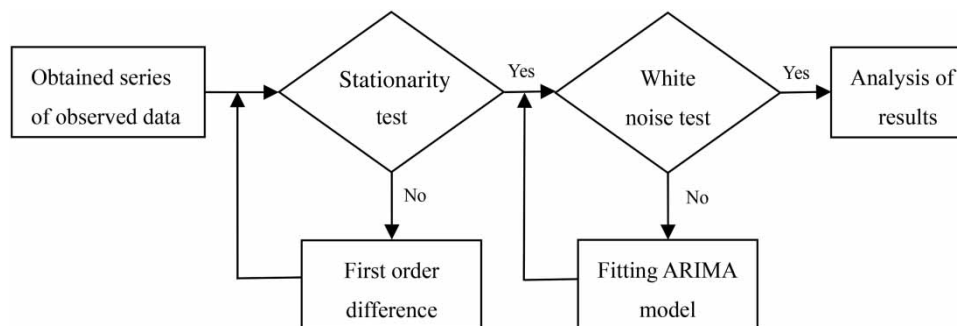
## 2.2. Model simulation

### 2.2.1. Auto-regressive integrated moving average model (ARIMA)

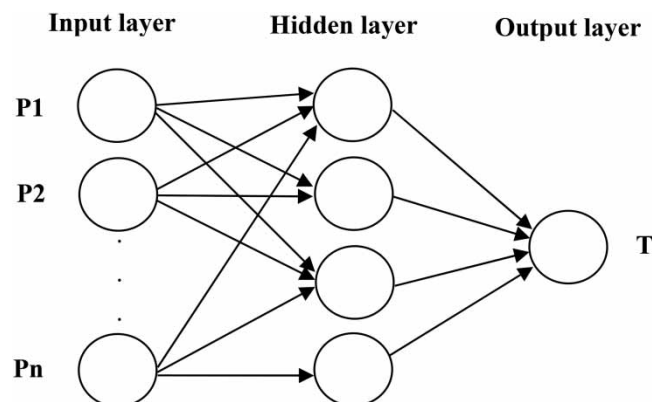
The ARIMA model, based on time series analysis, can solve a real problem using random data series. This approach assumes that the data is independent, but time series analysis focuses on dealing with the dependency of data series. The process of model analysis was as follows: (1) data arrangement, (2) time series model selection, (3) model performance, and (4) model prediction. A time series model flow chart is shown in Figure 1.

### 2.2.2. Back-propagation ANN (BP-ANN)

An Artificial Neural Network (BP-ANN) is a mathematical model imitating the behavior of an animal's neural network to perform distributed information processing, with a strong nonlinear mapping ability (Deng *et al.* 2011). BP-ANN is appropriate for modeling nonlinear processes, such as soil water and salt dynamics (Morshed & Kaluarachchi 1998). The BP-ANN is a multi-layer feed-forward network with error back-propagation. The network output error propagates back to modify the network weights and thresholds, so as to realize the nonlinear mapping of the network. BP-ANN is relatively mature and widely used at present. The basic structure of a BP-ANN is shown in Figure 2.



**Figure 1** | Schematic diagram depicting time series modeling (Xiao & Guo 2009).



**Figure 2** | Elements of a BP-ANN.

The specific algorithm formula for the nonlinear action function,  $f()$ , was described by Zhang (2006). For the  $i^{\text{th}}$  node of the neural network, the node output is:

$$y_j = f\left(\sum w_{ij} \times x_i - \theta_j\right) \quad (1)$$

where  $y_j$  is the output of the  $i^{\text{th}}$  node,  $f()$  depicts the nonlinear action function,  $x_i$  is the input of the  $i^{\text{th}}$  node,  $w_{ij}$  is the connection weight of the  $i^{\text{th}}$  and  $j^{\text{th}}$  nodes, and  $\theta_j$  is the threshold of the  $i^{\text{th}}$  neuron. The weights associated with node connections may be modified:

$$\Delta w_{ij(n+1)} = \alpha \times e_i \times y_j + \eta \times \Delta w_{ij(n)} \quad (2)$$

where  $\alpha$  is the dynamically adjusted learning factor according to output error,  $\eta$  is the momentum factor,  $e_i$  is the calculation error, and  $y_j$  is the output of the  $i^{\text{th}}$  node. The error term is:

$$e_p = 0.5 \times \sum (t_{pi} - o_{pi})^2 \quad (3)$$

where  $e_p$  is the estimation error of the  $i^{\text{th}}$  node,  $t_{pi}$  is the expected output value of node  $i$ , while  $o_{pi}$  is the calculated output value of node  $i$ . The hidden layer is a hyperbolic tangent transfer function (TANSIG):

$$f(x) = \frac{2}{1 + e^{-2x}} - 1 \quad (4)$$

### 2.2.3. Least squares support vector machine (LS-SVM)

Least squares support vector machine (LS-SVM) based on statistical learning theory is a method developed by Cortes & Vapnik (1995). Because of its powerful capabilities in classification and regression, LS-SVM has been widely applied in the artificial intelligence field. The LS-SVM developed by Suykens & Vandewalle (1999) is an improved algorithm, using equality-type constraints instead of inequalities.

Given a training set of  $N$  data points,  $\{x_i, y_i\}_{i=1}^N$ , with input data  $x_i \in R^n$ , output data  $y_i \in R$  and total number of data patterns  $N$ , the nonlinear function of the LS-SVM is defined as:

$$y(x) = w^T \phi(x) + b \quad (5)$$

where  $w^T$  is the weight vector,  $\phi(x)$  is the mapping function that maps  $x$  into the high-dimensional feature vector, and  $b$  is the bias.

An LS-SVM optimization problem is formulated as follows:

$$\min J(w, e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^N e_i^2 \quad (6)$$

Equation (6) is subject to the constraints:

$$y(x_i) = w^T \phi(x_i) + b + e_i \quad i = 1, 2, 3, \dots, N \quad (7)$$

where  $\gamma$  is the regularization constant parameter,  $e_i$  is the error vector for  $x_i$ , and  $b$  is the bias. One defines the Lagrangian:

$$L(w, b, e, a) = J(w, b, e) - \sum_{i=1}^n a_i \{w^T \phi(x_i) + b - y_i + e_i\} \quad (8)$$

where  $a_i$  is the Lagrangian multipliers. The conditions for optimality lead to a set of linear equations. Equation (9) is the solution for Equations (6) and (7):

$$\begin{bmatrix} 0 & -Y^T \\ Y & k \left( x, x_i + \frac{1}{r} I \right) \end{bmatrix} \begin{bmatrix} b \\ a_i \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (9)$$

where  $Y = (y_1; y_2; \dots; y_n)$  and  $k(x, x_i)$  is the common kernel function that is satisfied with Mercer's condition in the form of polynomial, sigmoid, Gaussian, and radial basis (RBF) kernel functions. In this study, the RBF was used and defined as:

$$k(x, x_i) = \exp\left(-\frac{(x - x_i)^2}{2\sigma^2}\right) \quad (10)$$

where  $\sigma^2$  is the width of the RBF,  $i = 1, 2, 3 \dots N$ .

The LS-SVM regression formulation is then defined:

$$f(x) = \sum_{i=1}^N a_i k(x, x_i) + b \quad (11)$$

## 2.3. Preliminary treatment

### 2.3.1. Auto-regressive integrated moving average model (ARIMA)

A time series analysis must undergo the stationarity test (Figure 1). If the series is not stationarity, i.e. the auto-correlation of the data set is significant and the data exhibits a time-dependent trend, the time series must be filtered by an appropriate mathematical model to remove the trend. Thus, the auto-correlation of the residual white noise series should not be significant (Wu *et al.* 1997). The time series of soil water content of both 0–14 cm and 14–33 cm layers showed regular and periodic fluctuations. The first-order difference was applied to obtain stationary time series.

The resulting residual series would then undergo further testing for randomness (i.e. the white noise test in Figure 1). With increasing lag time, the coefficients of the corresponding auto-correlation function (ACF) exhibited a tailing phenomenon (Wang *et al.* 2011) and gradually fell within the confidence intervals. As the average of ACF approached zero, the residual time series were stationary and the ARIMA was chosen as the optimal model. The Akaike information criterion (AIC) in the following format was used to test the ARIMA:

$$AIC(p) = N \ln \sigma_{\theta}^2 + 2p \quad (12)$$

where  $p$  is model order,  $N$  is length of time series  $\{x_t\}$  and  $\sigma_{\theta}^2$  is the maximum of the likelihood equation,

$$\sigma_{\theta}^2 = \frac{1}{N} \sum_{t=1}^N \partial_t^2 \quad (13)$$

where  $\partial_t^2$  is the residual series.

A better model results in smaller AIC values (Tian 2006). The least-squares method was employed to determine the optimal model order defined by Equation (13) according to the partial auto-correlation function (PACF) of the residual series. Accordingly, the minimum AIC was 294 corresponding to ARIMA (2, 1, 0). The ACF of the ARIMA (2, 1, 0) model showed that the correlations among the residuals were not significant and the coefficients of ACF residuals passed the white noise test. Subsequently, the soil water content in the 0–14 cm and 14–33 cm soil layers were forecast by combining the outcomes of the first-order difference stationary series model and the ARIMA (2, 1, 0) residual series model.

### 2.3.2. Back-propagation artificial neural network (BP-ANN)

The BP-ANN model was built by designating daily measurements of mean, maximum, and minimum temperatures, cumulative rainfall, cumulative evaporation, mean wind speed, mean humidity and mean soil surface temperature as the eight layers of neural inputs and the soil water content of the 0–14 cm and 14–33 cm soil layers as outputs.

The hidden layer contained the TANSIG and PURELIN that were the transfer functions to mathematically facilitate the conversion of input neurons to the outputs. The transfer function between input and hidden layers was a sigmoid function (TANSIG). The transfer function between hidden and output layers was a linear function (PURELIN). The network learning process employed the training data set to calibrate the function parameters through a trial- and-error iteration process. The error tolerance and max-iteration of the BP-ANN were 0.02 and 1,000, respectively. There were five hidden layer neurons,

and the network structure was 8:5:1, for the input, hidden, and output layers, respectively. Transfer functions between the layers were designed by the Elman network (Wen *et al.* 2009; Deng *et al.* 2021).

The general network output vector of the sub-values should be between  $-1$  and  $1$ . To ensure that the large input neuron falls within the large gradient area of the neuron activation function, the components of input vectors should also fall between  $-1$  and  $1$ . Before network training, the input and output variables were normalized such that:

$$X'_i = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \quad (14)$$

where  $X'_i$  and  $X_i$  are the  $i^{\text{th}}$  observation of a neuron input layer before and after processing, and  $X_{\min}$  and  $X_{\max}$  are the maximum and minimum of each neuron input layer, respectively.

### 2.3.3. Least squares support vector machine (LS-SVM)

The LS-SVM is an improved version of the support vector machine, using equality-type constraints instead of inequality constraints, which increases the speed of problem solving. We chose the radial basis kernel function (RBF) as the common kernel function of LS-SVM. The regularization constant parameter ( $r = 7,000$ ) and nuclear parameter ( $\sigma = 0.4$ ) were determined by robust cross-validation.

In this paper, some MATLAB2010b modules, including the fuzzy logic toolbox, neural network toolbox and wavelet toolbox, were used to develop the simulation system based on LS-SVM, with the meteorological factors and prediction system based on chaotic time series analysis.

## 3. MODEL PERFORMANCES

Model performance in this study was evaluated on the basis of the root-mean square error (RMSE), the mean absolute relative error (MARE), and correlation coefficient ( $r$ ) (Chen *et al.* 1998). The performance functions were defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n [\hat{y}(i) - y(i)]^2}{n-1}} \quad (15)$$

$$MARE = \frac{1}{n} \sum_{i=1}^n \frac{|\hat{y}(i) - y(i)|}{y(i)} \quad (16)$$

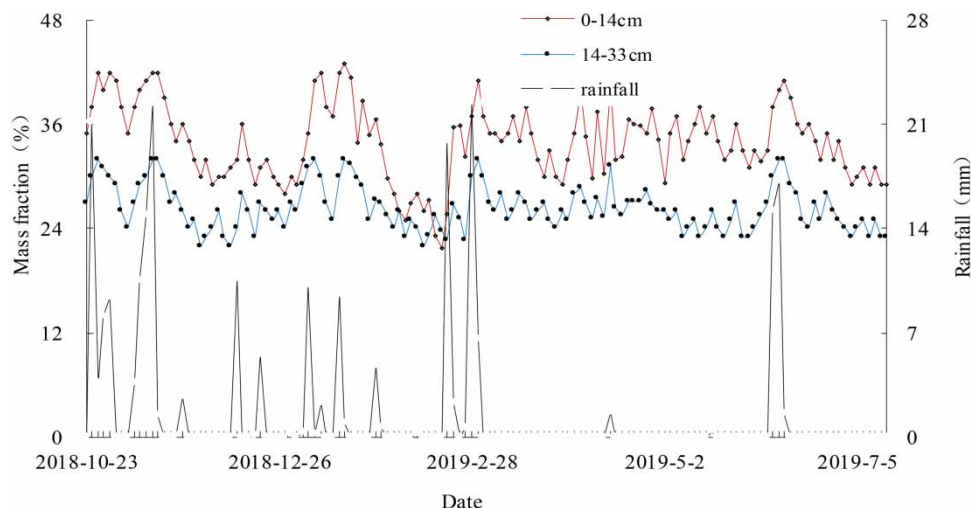
$$r = \frac{\sum_{i=1}^n \{[\hat{y}(i) - \hat{y}_m][y(i) - y_m]\}}{\sqrt{\sum_{i=1}^n [\hat{y}(i) - \hat{y}_m]^2 \sum_{i=1}^n [y(i) - y_m]^2}} \quad (17)$$

where  $n$  is the total number of data points,  $\hat{y}(i)$  is the simulated value,  $y(i)$  is the original measured value,  $\hat{y}_m$  is the mean of the simulated value, and  $y_m$  is the mean of the original measured value.

## 4. RESULTS

### 4.1. Dynamics of soil water content

The soil water content in the 0–14 cm soil layer was higher than in the 14–33 cm soil layer (except for 25 February 2019, which was due to more than ten days without precipitation). The magnitude of soil water content fluctuations with time was greater in the 0–14 cm soil layer than in the 14–33 cm soil layer (Figure 3). The coefficients of variation of soil water content in the 0–14 cm and 14–33 cm soil layers were 12.93% and 9.98%, respectively. The germination stage of wheat is from November to December when wheat needs a lot of water to germinate. Usually, there is enough water stored in soil for wheat germination in the study area. The heading–grouting stage of wheat is from the end of March to the middle of April. During this stage, wheat growth needs more irrigation. The mature stage of wheat is between the end of May and the start of July and needs less water than does the heading–grouting stage.



**Figure 3** | Measured daily soil water fluctuations of the soil.

#### 4.2. Time series modeling of soil water content

The ARIMA (2, 1, 0) developed from the testing data set provided reasonable estimates and the predicted magnitude and trend were in close agreement with the corresponding measurements (Figure 4). Time series soil moisture contents in both 0–14 cm and 14–33 cm soil layers tended to be of periodic fluctuations. Therefore, the first-order difference was applied to get stationary time series in this study. The mean absolute relative errors between the predictions and measurements in the 0–14 cm and 14–33 cm soil layers were 0.20 and 0.11, respectively (Table 2). The ARIMA (2, 1, 0) time series model was suitable for simulation and prediction of soil water in this region.

#### 4.3. BP-ANN modeling of soil water content

The BP-ANN model was built by daily rainfall, average daily temperature, daily maximum and minimum temperatures, daily cumulative evaporation, daily average relative humidity, daily average wind speed, and daily average land surface temperature as input variation, and soil water at 0–14 cm and 14–33 cm as output variation. Soil water contents in the 0–14 cm and 14–33 cm depth layers were predicted by the BP-ANN model (Figure 5).

The soil water estimates were comparable to the observed values in the 0–14 cm and 14–33 cm soil layers (Figure 5). The maximum relative errors for the two soil depths were 0.14 and 0.16, respectively. Correlation coefficients ( $r$ ) between observed and simulated values were 0.8397 and 0.8721 for the 0–14 cm and 14–33 cm layers respectively (Table 2).

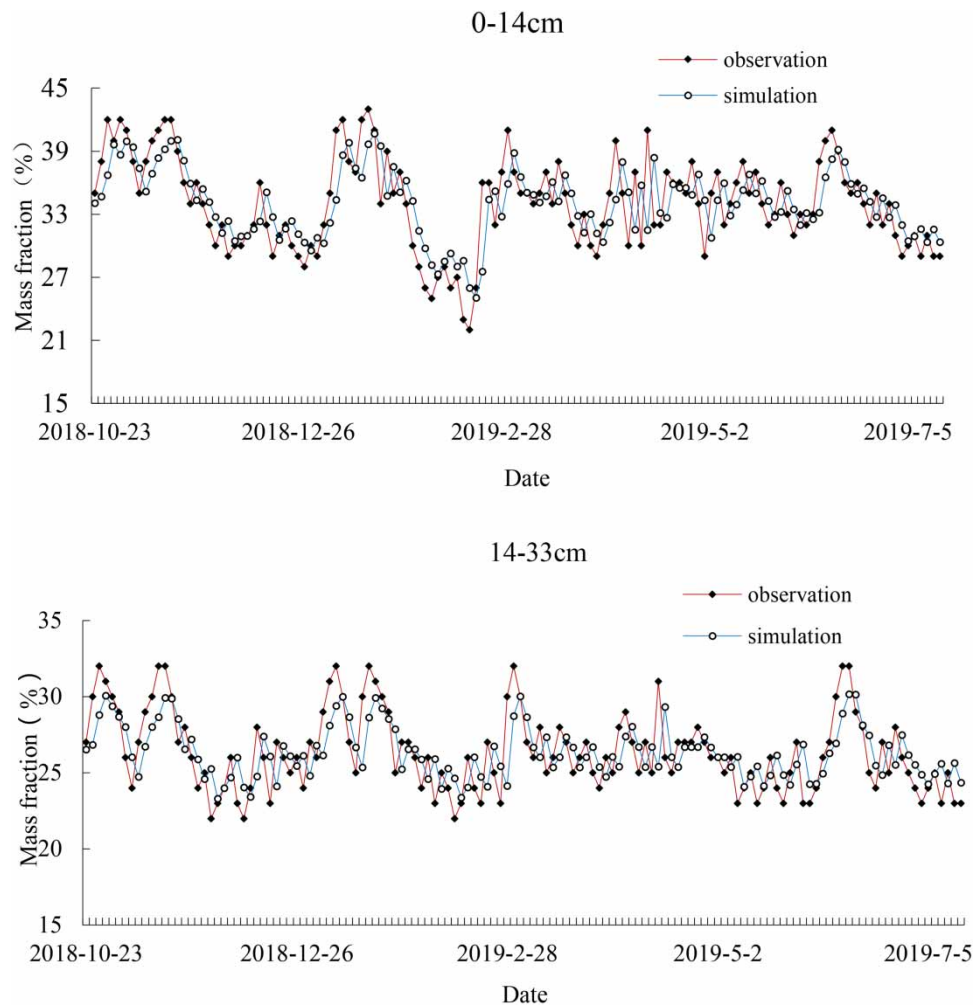
#### 4.4. LS-SVM modeling of soil water content

The LS-SVM is an improved method of the support vector machine, using equality-type constraints instead of inequality-type constraints, which will increase the speed of solving problems. The soil water content predicted by the LS-SVM model agreed well with observed values. The correlation coefficient ( $r$ ) of the estimates was 0.9664 and 0.9827 for the 0–14 cm soil layer and 14–33 cm soil layer, respectively. The simulated values have good correspondence with the observed values (Figure 6). The mean absolute relative error in the 0–14 cm soil layer and 14–33 cm soil layer was 0.05 and 0.04, respectively (Table 2).

### 5. DISCUSSION

All three models (ARIMA, BP-ANN and LS-SVM) provided good predictions of soil water changes (Table 2). In comparison with the ARIMA and BP-ANN, the LS-SVM training speed was quicker.

The MARE represents the cumulative error of the predictions versus measurements distributed over the entire data set; the MARE of perfect agreement will be zero. The MARE of the model predictions varied from 0.04 to 0.20, indicating all three models provided reasonably accurate predictions of daily water changes in both 0–14 cm and 14–33 cm soil layers. The models that relied on inputs of daily meteorological information (i.e. BP-ANN and LS-SVM) performed slightly better with MARE ranging from 0.04 to 0.16. The time-series-based model, ARIMA, relying strictly on the trends was slightly less



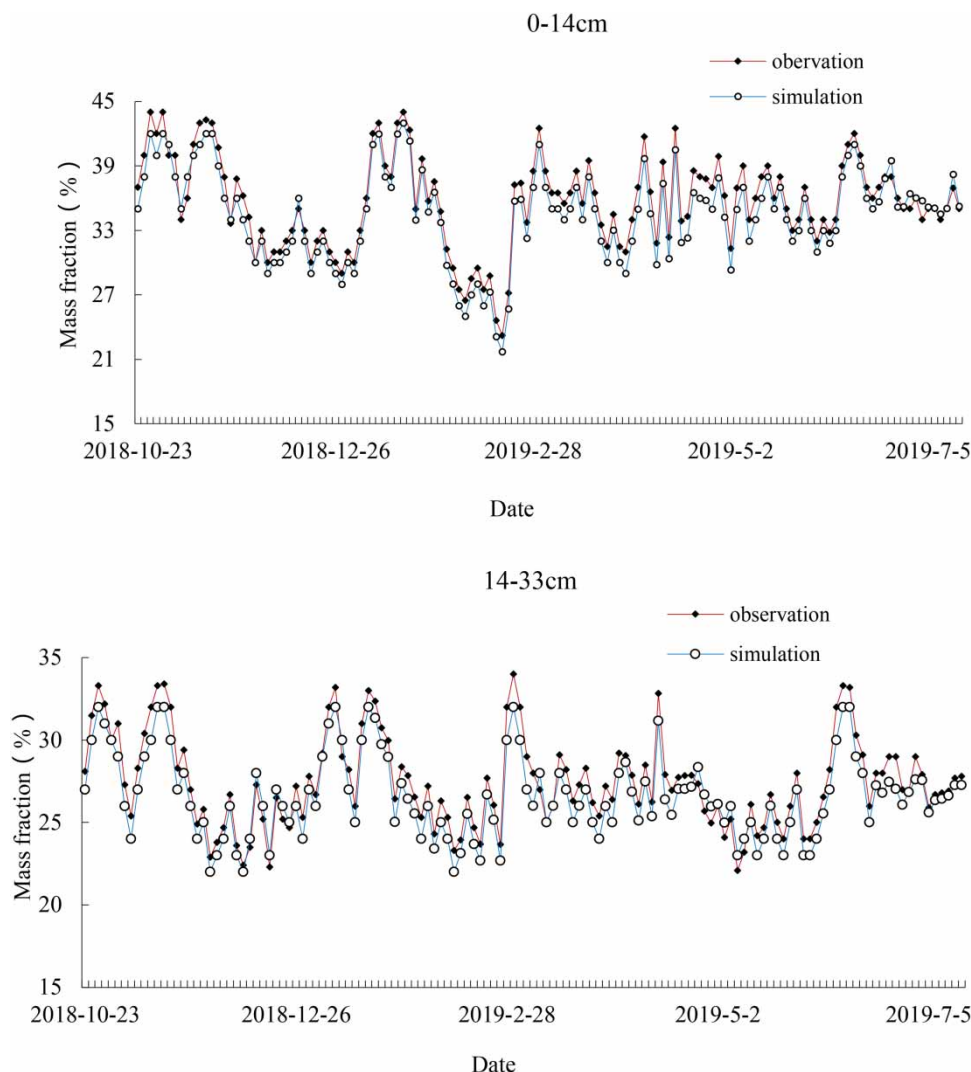
**Figure 4** | Daily soil water changes prediction by the ARIMA model vs measurements.

**Table 2** | The simulation performance statistics of the three models

| Model  | Soil depth (cm) | RMSE | MARE | <i>r</i> |
|--------|-----------------|------|------|----------|
| ARIMA  | 0–14            | 2.21 | 0.20 | 0.7019   |
|        | 14–33           | 1.10 | 0.11 | 0.7868   |
| BP-ANN | 0–14            | 1.45 | 0.14 | 0.8397   |
|        | 14–33           | 1.33 | 0.16 | 0.8721   |
| LS-SVM | 0–14            | 0.51 | 0.05 | 0.9664   |
|        | 14–33           | 0.46 | 0.04 | 0.9827   |

responsive, with MARE ranging from 0.11 to 0.20. The RMSE represents the cumulative error of the squared terms distributed over the entire data set.

Comparing the simulated values with the observed values, the root-mean-square error (RMSE) of soil moisture content is higher for the 0–14 cm soil layer than for the 14–33 cm soil layer, and the correlation coefficient (*r*) is lower for the 0–14 cm soil layer than for the 14–33 cm soil layer, because the variability in soil moisture content was greater for the 0–14 cm soil layer than for the 14–33 cm soil layer, and the soil moisture in the 0–14 cm soil layer was more influenced by meteorological factors. In this regard, it is biased toward larger deviations between predictions and measurements. In relative terms, a higher



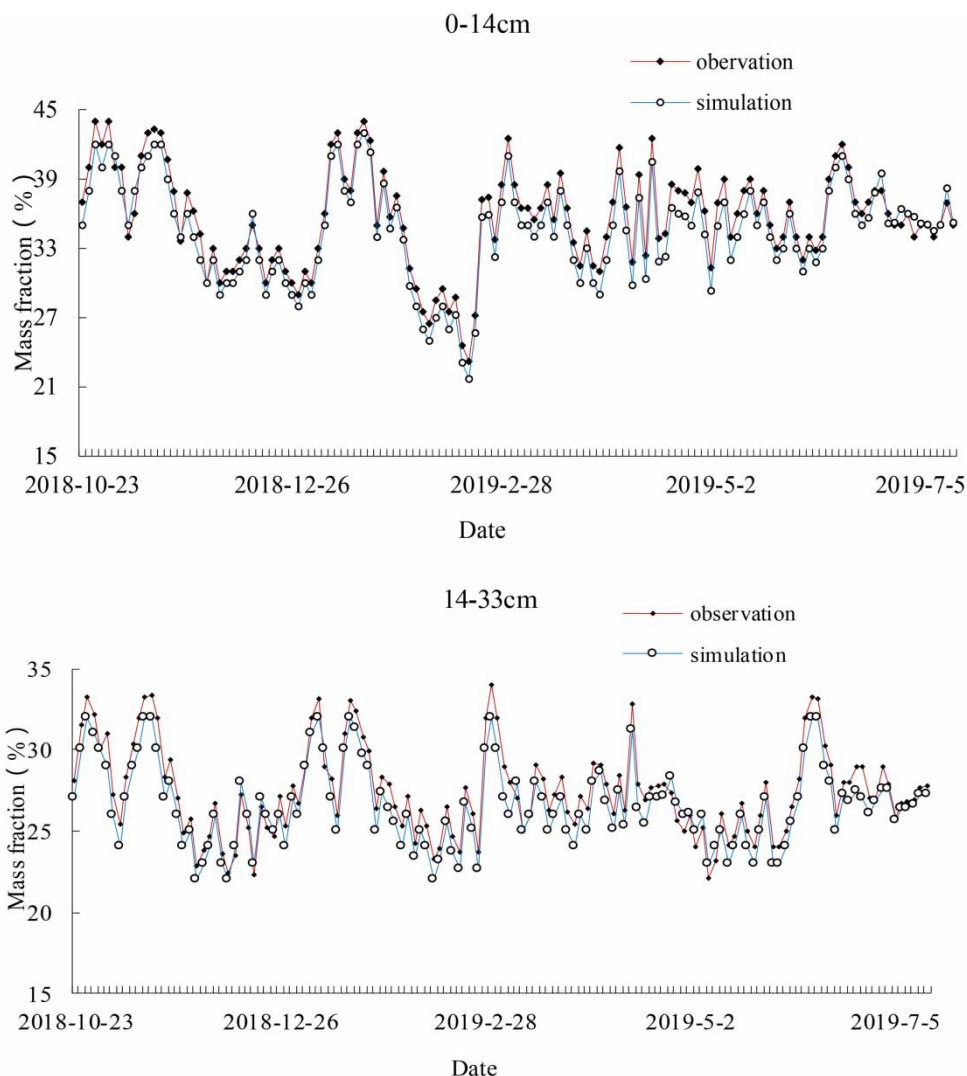
**Figure 5** | Daily soil moisture changes prediction by the BP-ANN model vs measurements.

RMSE denotes more and greater deviations between the predictions and measurements. The RMSE of the predictions varied from 0.46 to 2.21, an indication that the deviations from the measured data were tolerable. The models that relied on inputs of daily meteorological information, i.e. BP-ANN and LS-SVM, exhibited lesser degrees of deviation with RMSE ranging from 0.46 to 1.45. The time-series-based model, ARIMA, which relied strictly on the trends, exhibited a greater extent of deviation with RMSE ranging from 1.10 to 2.21.

The correlation coefficient indicates the trend exhibited by two parallel data sets. The correlations of ARIMA-, BP-ANN- and LS-SVM-based predictions were all significant with average  $r$  equal to 0.7019, 0.8397, and 0.9664 respectively in the 0–14 cm soil layer. Judging from the indices of their predictive performances, the LS-SVM-based model exhibited more consistent and accurate predictions over the ARIMA- and BP-ANN-based models. [Deng et al. \(2011\)](#) also found that the LS-SVM model performed better in simulating the dynamic trend of soil water in the red soil region of China, especially under a changing climate.

## 6. CONCLUSIONS

The ARIMA, BP-ANN and LS-SVM models were used to predict soil water dynamics in the 0–14 cm and 14–33 cm layers of cultivated soils in the Taihu Lake region of China. They were all successful in predicting the trend and magnitudes of daily soil water fluctuations in terms of precision and relative maximum errors. Rainfall, evaporation, temperature, humidity and wind



**Figure 6** | Daily soil water changes prediction by the LS-SVM model vs measurements.

speed that affect soil water content changes were considered in the BP-ANN and LS-SVM, but not in ARIMA. The results were most satisfactory when meteorological factors were taken as input variables, and the simulation accuracy of LS-SVM was the highest. Therefore, the LS-SVM model can be used to predict soil water dynamics in the study area.

## ACKNOWLEDGEMENTS

The authors thank Dr Christopher Ogden for his checking of the English language and comments on this paper. This research was funded by the Natural Science Foundation of China (42107478), Regional Innovation and Development Joint Fund of National Natural Science Foundation of China (U20A2098), the Natural Science Foundation of Jiangsu Province, China (No. BK 20150909), State Key Laboratory of Soil and Sustainable Agriculture (Institute of Soil Science, Chinese Academy of Sciences) (No. Y20160038), Foundation of Chinese postdoctoral (2016M591884).

## CONFLICT OF INTEREST

The authors declare no competing interests.

## DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

## REFERENCES

- Aitkenhead, M. J. & Coull, M. C. 2016 Mapping soil carbon stocks across Scotland using a neural network model. *Geoderma* **262**, 187–198.
- Almomani, F. 2020 Prediction the performance of multistage moving bed biological process using artificial neural network (ANN). *Sci. Total Environ.* **744**, 140854.
- Asquith, W. H. 2020 The use of support vectors from support vector machines for hydrometeorologic monitoring network analyses. *J. Hydrol.* **583**, 124522.
- Bai, X., Jia, X. X., Jia, Y. H., Shao, M. A. & Hu, W. 2020 Modeling long-term soil water dynamics in response to land-use change in a semi-arid area. *J. Hydrol.* **585**, 124824.
- Bao, S. D. 2000 *Soil Agricultural Chemical Analysis*, 3rd edn. Agriculture Press, Beijing, China (in Chinese).
- Bialkowski, R. & Buttle, J. M. 2015 Stemflow and throughfall contributions to soil water recharge under trees with differing branch architectures. *Hydrol. Processes* **29**, 4069–4082.
- Bohara, H., Dodla, S., Wang, J. J., Darapuneni, M., Acharya, B. S., Magdi, S. & Pavuluri, K. 2019 Influence of poultry litter and biochar on soil water dynamics and nutrient leaching from a very fine sandy loam soil. *Soil Tillage Res.* **189**, 44–51.
- Chen, J. L., Islam, S. & Biswas, P. 1998 Nonlinear dynamics of hourly ozone concentrations: nonparametric short term prediction. *Atmos. Environ.* **32**, 1839–1848.
- Cortes, C. & Vapnik, V. 1995 Support-vector networks. *Mach. Learn.* **20**, 273–297.
- Deng, J. Q., Chen, X. M., Du, Z. J. & Zhang, Y. 2011 Soil water simulation and prediction using stochastic models based on LS-SVM for red soil region of China. *Water Resour. Manage.* **25**, 2823–2836 (in Chinese).
- Deng, W., Shang, S., Cai, X., Zhao, H., Zhou, Y., Chen, H. & Deng, W. 2021 Quantum differential evolution with cooperative coevolution framework and hybrid mutation strategy for large scale optimization. *Knowledge-Based Syst.* **224**, 107080.
- Fan, M. Q., Wang, H. X. & Li, S. K. 2008 Restudy on SVD-based watermarking scheme. *Appl. Math. Comput.* **203** (2), 926–930.
- Hanh, P. T. M., Anh, N. V., Ba, D. T., Sthiannopkao, S. & Kim, K.-W. 2010 Analysis of variation and relation of climate, hydrology and water quality in the lower Mekong River. *Water Sci. Technol.* **62** (7), 1587–1594.
- Jahanbakhsh, S. & Babapour Basseri, E. A. 2003 Studying and forecasting of the mean monthly temperature of Tabriz, using ARIMA model. *J. Geog. Res.* **15** (3), 34–46.
- Jin, T., Xia, H., Deng, W., Li, Y. & Chen, H. 2021 Uncertain fractional-order multi-objective optimization based on reliability analysis and application to fractional-order circuit with Caputo type. *Circuits Syst. Signal Process.* **40**, 5955–5982.
- Kashani, M. H., Ghorbani, M. A., Shahabi, M., Naganna, S. R. & Diop, L. 2020 Multiple AI model integration strategy – application to saturated hydraulic conductivity prediction from easily available soil properties. *Soil Tillage Res.* **196**, 104449.
- Mojid, M. A., Hossain, A. B. M. Z. & Ashraf, M. A. 2019 Artificial neural network model to predict transport parameters of reactive solutes from basic soil properties. *Environ. Pollut.* **255**, 113355.
- Morshed, J. & Kaluarachchi, J. J. 1998 Application of artificial neural network and genetic algorithm in flow and transport simulations. *Adv. Water Resour.* **22**, 145–158.
- Narayanan, P., Basistha, A., Sarkar, S. & Sachdeva, K. 2013 Trend analysis and ARIMA modelling of pre-monsoon rainfall data for western India. *C.R. Geosci.* **345** (1), 22–27.
- Nguyen, L. T. T., Osanai, Y., Lai, K., Anderson, I. C., Bange, M. P., Tissue, D. T. & Singh, B. K. 2018 Responses of the soil microbial community to nitrogen fertilizer regimes and historical exposure to extreme weather events: flooding or prolonged-drought. *Soil Biol. Biochem.* **118**, 227–236.
- Panda, D. K. & Kumar, A. 2011 Evaluation of an over-used coastal aquifer (Orissa, India) using statistical approaches. *Hydrol. Sci. J.* **56** (3), 486–497.
- Parmar, K. S. & Bhardwaj, R. 2015 Statistical, time series, and fractal analysis of full stretch of river Yamuna (India) for water quality management. *Environ. Sci. Pollut. Res.* **22** (1), 397–414.
- Patle, G. T., Singh, D. K., Sarangi, A., Rai, A., Khanna, M. & Sahoo, R. N. 2015 Time series analysis of groundwater levels and projection of future trend. *J. Geol. Soc. India* **85**, 232–242.
- Suykens, J. A. K. & Vandewalle, J. 1999 Least squares support vector machine classifiers. *Neural Process. Lett.* **9**, 293–300.
- Suykens, J. A. K., Vandewalle, J. & De Moor, B. 2001 Optimal control by least squares support vector machines. *Neural Networks* **14** (1), 23–35.
- Tian, Z. 2006 *Time Series Theory and Methods*. Higher Education Press, Beijing, China, pp. 214–250 (in Chinese).
- Tiyasha, T., Tung, T. M. & Yaseen, Z. M. 2020 A survey on river water quality modelling using artificial intelligence models: 2000–2020. *J. Hydrol.* **585**, 124670.
- Vereecken, H., Huisman, J. A., Franssen, H. J. H., Brüggemann, N., Bogaen, H. R., Kollet, S., Javaux, M., van der Kruk, J. & Vanderborght, J. 2015 Soil hydrology: recent methodological advances, challenges, and perspectives. *Water Resour. Res.* **51**, 2616–2633.
- Wang, F., Chen, S. K. & Feng, G. S. 2011 *SAS Statistical Analysis and Application*, vol. 3. Publishing House of Electronics Industry, Beijing, China, pp. 9–31 (in Chinese).

- Wen, X., Zhou, L. & Wang, L. 2009 *Matlab Neural Network Design Application*. Beijing Science Press, Beijing, China, pp. 200–221 (in Chinese).
- Wu, L., Jury, W. A., Chang, A. C. & Allmaras, R. R. 1997 [Time series analysis of field-measured water content of a sandy soil](#). *Soil Sci. Soc. Am. J.* **61**, 736–742.
- Xiao, Z. F. & Guo, M. Y. 2009 *Time Series Analysis and SAS Application*. Wuhan University Press, Wuhan, China, pp. 145–155 (in Chinese).
- Yu, P. S., Yang, T. C., Chen, S. Y., Kuo, C. M. & Tseng, H. W. 2017 [Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting](#). *J. Hydrol.* **552**, 92–104.
- Zhang, D. Y. 2006 *Neural Network Theory and Methods*. Tsinghu University Press, Beijing, China, pp. 34–35 (in Chinese).
- Zounemat-Kermani, M., Kişi, Ö., Adamowski, J. & Ramezani-Charmahineh, A. 2016 [Evaluation of data driven models for river suspended sediment concentration modeling](#). *J. Hydrol.* **535**, 457–472.
- Zucco, G., Brocca, L., Moramarco, T. & Morbidelli, R. 2014 [Influence of land use on soil moisture spatial-temporal variability and monitoring](#). *J. Hydrol.* **516**, 193–199.

First received 26 October 2021; accepted in revised form 14 January 2022. Available online 29 January 2022