Ain Shams Engineering Journal xxx (xxxx) xxx

Contents lists available at ScienceDirect

Ain Shams Engineering Journal

journal homepage: www.sciencedirect.com

Civil Engineering

Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia

Ahmedbahaaaldin Ibrahem Ahmed Osman [a], Ali Najah Ahmed [b], Ming Fai Chow [c], Yuk Feng Huang [d,*], Ahmed El-Shafie [e,f]

[a] Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), Kajang 43000, Selangor Darul Ehsan, Malaysia
[b] Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), 43000 Selangor, Malaysia
[c] Institute of Sustainable Energy (ISE), Universiti Tenaga Nasional (UNITEN), 43000 Selangor, Malaysia
[d] Department of Civil Engineering, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Selangor, Malaysia
[e] Department of Civil Engineering, Faculty of Engineering, University of Malaya (UM), 50603 Kuala Lumpur, Malaysia
[f] National Water and Energy Center (NWC), United Arab Emirates University, P.O. Box. 15551, Al Ain, United Arab Emirates

ARTICLE INFO

Article history:
Received 14 August 2020
Revised 17 October 2020
Accepted 19 November 2020
Available online xxxx

Keywords:
Groundwater level prediction
Machine learning algorithm
Artificial neural network
Support vector regression
Cross-correlation

ABSTRACT

Groundwater levels have been declining recently in Malaysia. This is why, the current study was aimed to propose an accurate groundwater levels prediction model using machine learning algorithms in highly populated towns in Selangor, Malaysia. The models developed used 11 months of previously recorded data of rainfall, temperature and evaporation to predict groundwater levels. Three machine learning models have been tested and evaluated; Xgboost, Artificial Neural Network, and Support Vector Regression. The results showed that for the first scenario, which had combinations of 1,2 and 3 days delayed of rainfall data only considered as an input, the models' performance was the worst. while in the second scenario the proposed Xgboost model outperformed both the Artificial Neural Network and Support Vector Regression models for all different input combinations. A significant increase in performance was achieved in the third scenario, when using 1 day delayed of groundwater levels as an input as well where R^2 equal to 0.92 in the Xgboost model in scenario 3 and 0.16, 0.11 in scenarios 2 and 1 respectively. The results obtained in this study serves as a great benchmark for future groundwater levels prediction using Xgboost algorithm.

1. Introduction

Change in climate and weather conditions, increase in water demand, and water pollution are all factors contributed to recent water crisis around the world [9]. The United Nations (UN) estimates that around 1.2 billion people live in regions with total water scarcity, and around a quarter more of the world population are approaching the same situation [16]. Groundwater is the largest and one of the most important source of freshwater [29]. However, the increase in groundwater extraction and climate change played a significant role in increasing groundwater level decline [21]. Accurate prediction of groundwater levels is a major challenge in water resources management of aquifer systems, and its modeling is of significant importance in regions where there is insufficient amount of radially available surface water [7]. The change in groundwater levels is influenced by boundary constraints as well as change in hydrological and meteorological variables including precipitation, temperature and evaporation [30]. Furthermore, the way these variables interact with each other has been difficult to accurately characterize. These changes affect the design, planning and management of irrigation systems and water resources. The significance of groundwater and its influence on water balance is emphasized through its relation to climate change. The recent results of meteorological models show that climate change has caused a decrease in groundwater recharge which is projected to have serious impacts on managing water resources

* Corresponding author.
E-mail addresses: ahmedbahaaaldin@uniten.edu.my (A. Ibrahem Ahmed Osman), mahfoodh@uniten.edu.my (A. Najah Ahmed), Chowmf@uniten.edu.my (M.F. Chow), huangyf@utar.edu.my (Y. Feng Huang), elshafie@um.edu.my (A. El-Shafie).

Peer review under responsibility of Ain Shams University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.asej.2020.11.011
2090-4479/

and reducing accessibility to clean water [23]. Thus, the need for accurate prediction models and understanding the influence of the hydrological and meteorological variables on groundwater levels have become an important part of implementing water resources management plans and to ensure water use efficiency, especially as water demand increase in urban areas [32], and future climate change projections indicate an increase in our dependency on groundwater [23]. To predict groundwater levels, a range of parameters recorded at meteorological stations are used as an input for the models. The most important of these parameters are relative humidity, sunshine hours, wind speed, rainfall, evaporation, and temperature [11,14,22,28,30]. Use irrigation demand, streamflow, climate change indices, rainfall, and temperature as input parameters for predicting groundwater level change at agricultural regions across the US. Their results show that including irrigation demand, rainfall and temperature make the second and third most significant contribution to groundwater level predication. Using artificial neural network [17]. Investigate the relationship between groundwater level and meteorological parameters including rainfall, maximum temperature, minimum temperature, solar radiation, wind speed, relative humidity, elevation of area, polygon area and water depth using data from meteorological stations in Punjab, Pakistan. And concluded that rainfall has the most significant effect on groundwater level variations than any other parameter used in the study. Moghaddam et al. [26] use data of monthly evaporation, average temperature, discharge, and water tables of previous months for the period 2002 to 2014 as inputs parameters into conceptual model (MODFLOW), ANN, and Bayesian network (BN) to predict monthly groundwater level in Khorasan, Iran. The study reports RMSE values of 0.091, 0.11, and 0.023 for the models respectively. An accurate prediction of groundwater levels using different modeling methods and climate variables has been the objective of many researchers in recent years [26,39].

As cities expand across the world, governments and planners look for ways to accurately predict the groundwater level so they can make sure the water demand is met as the population growth keeps increasing and also as a very important factor when designing new cities.

Physical models have long been used to predict groundwater levels. These types of models require large amount of data and the take considerable amount of time to construct, especially for systems such as a groundwater system. The non-linearity nature of the groundwater system and its response to climatic variables makes it difficult to simulate using physical models because of the large amount of data needed to produce such accurate interpretation. Therefore, new modeling approaches based on machine learning algorithms become more reasonable to use for groundwater level prediction, as such modeling methods do not require information about the system's properties psychical characteristics. Machine learning algorithms utilize mathematical concepts in finding an ideal function from the data provided and use this function to classify, predict or detect certain outputs. The algorithms learn automatically by identifying patterns in the input data.

All around the world and for the past decade, different studies have been carried out using a variety of machine learning algorithms to model groundwater for the purpose of predicting and forecasting the fluctuations in its levels, the algorithms used including, Artificial Neural Network (ANN) [2,38], Support Vector Machine (SVM) [25,37], and Genetic Programming (GP) [19]. The results of using machine learning algorithms for the groundwater modeling showed the capability of these models as it achieved better performance compared to physical modelling methods [27,34]. However stand-alone machine learning models vary largely in their performance at different lead times [36]. So, to increase the performance, researchers both in academia and industry are actively developing new models by combining different algorithms. For example, a wavelet transformer (WA) is used to decompose a data series into its components that carry most of the information, then it is fed to a neural network model to make predictions [5]. This can significantly increase the prediction accuracy, as the WA allows the extraction of data series that carry the most important information and remove the noisy ones [1]. Genetic algorithms (GA) is also used as a method to pre-process the data before it is fed to predictive model, this can significantly increase the model accuracy [14]. However, even though hybrid models can achieve better performance than stand-alone models, there is normally a trade-off between model performance and training time required. So, a model that can achieve both great performance and fast training time is needed.

In this study three models (Artificial Neural Network, Support Vector Regression, and Xgboost)) with three different input parameters are investigated. Cross-correlation was preformed to select the best input parameters. The models then tested at a selected study location in Malaysia. Three scenarios based on input combination were investigated and the proposed models were tested in its ability to predict the groundwater levels. Then the models' performances were compared using different statistical performance indices.

2. Materials and methods

2.1. Study area and dataset

Selangor is a province in Malaysia. It is one of the warmest regions in the country with high temperature and humidity all year round. Selangor considered one of the states in Malaysia that faced severe droughts and water crises in the past [6]. The area in this study shown in Fig. 1, is located in the southern part of Selangor, With a total area of 1429.1 km^2 , and a total population of 1.259 million [10]. The two districts (Hulu Langat and Sepang) that the study area is comprised of, have a population density of 1,288 people/km^2 and 320 people/km^2 respectively.

This study is conducted to predict the groundwater levels in five towns, namely Jenderam, Bangi, Beranag, Kajang and Paya Indah Wetland. Table 1 shows data obtained from 20 October 2017 to 24 July 2018 of groundwater level, rainfall, temperature, and evaporation used as models' inputs. And Table 2 presents the descriptive statistics of the data. The input data is divided into 70% training and 30% testing respectively, previous researches indicated that the mentioned percentages should be sufficient for predictive models [15,27].

2.2. Input variable Selection for the models

In machine learning models, choosing the right input variables is crucial for the model's performance. Selection of input variables is done to ensure that the model can capture the relationship between inputs and the target variable during the training process [18]. In this current study, the sliding dot product also known as cross-correlation is used to find the similarity between the target variable and the delayed values of the inputs. For example, to find the cross-reaction coefficient between x_t and y_{t+k} the following equation is used:

$$r_k = \frac{\sum_{i=1}^{n-k}\left(x_i-\bar{x}\right)\left(Y_{i+k}-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}} \quad (1)$$

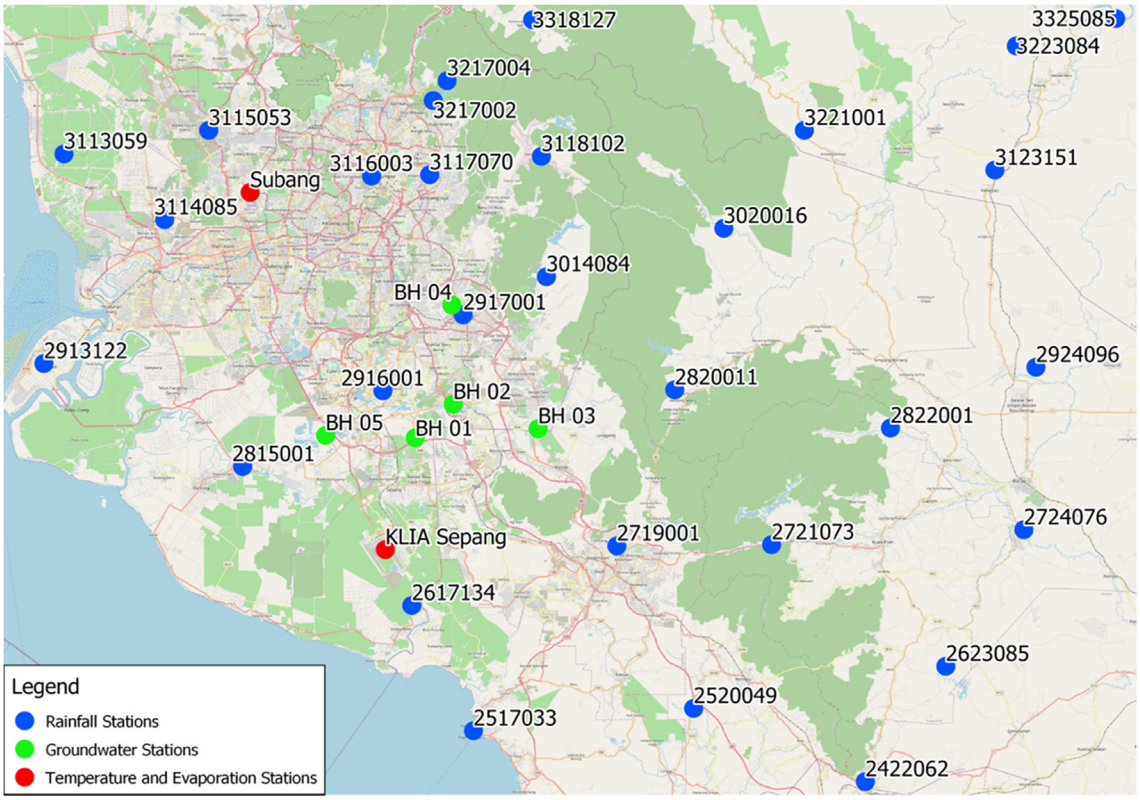

Fig. 1. Location of the study area, with the locations of the mentoring wells, rainfall stations, and temperature/evaporation stations.

Table 1
Rainfall, temperature, evaporation, and groundwater level data used to conduct the study.

Date	Rainfall (mm)	Temperature (℃)	Evaporation (mm)	GWL BH1 (m)	GWL BH2 (m)	GWL BH3 (m)	GWL BH4 (m)	GWL BH5 (m)
20-Oct-2017	0.15	29.95	4.90	16.75	15.15	16.37	13.95	16.22
21-Oct-2017	0.15	30.15	6.90	16.70	15.35	16.35	13.93	16.20
22-Oct-2017	0.05	30.30	7.35	16.64	15.12	16.33	13.89	16.18
23-Oct-2017	0.00	30.15	6.65	16.58	15.09	16.31	13.86	16.17
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
20-Jul-2018	0.00	29.00	5.30	17.00	14.92	16.19	13.67	16.25
21-Jul-2018	0.89	28.85	6.20	16.99	14.93	16.19	13.66	16.24
22-Jul-2018	2.67	27.90	3.05	17.00	14.94	16.19	13.67	16.24
23-Jul-2018	4.39	28.25	3.65	17.05	14.95	16.19	13.65	16.23
24-Jul-2018	0.33	28.00	3.65	16.82	14.93	16.19	13.65	16.21

Table 2
Descriptive statistics of climatic and hydrological variables.

Parameter	Rainfall (mm)	Temperature (℃)	Evaporation (mm)	GWL BH1 (m)	GWL BH2 (m)	GWL BH3 (m)	GWL BH4 (m)	GWL BH5 (m)
Mean	6.23	27.81	4.47	17.10	15.31	16.36	14.00	16.39
Standard Error	0.41	0.07	0.09	0.03	0.02	0.01	0.01	0.01
Median	3.82	27.85	4.40	17.07	15.23	16.35	13.92	16.34
Mode	0.00	27.85	4.10	17.00	15.17	16.37	13.85	16.26
Standard Deviation	6.90	1.14	1.55	0.45	0.30	0.14	0.23	0.17
Sample Variance	47.60	1.30	2.40	0.20	0.09	0.02	0.05	0.03
Kurtosis	2.66	1.56	0.46	−0.78	2.97	0.01	4.46	−0.91
Skewness	1.48	−0.51	0.13	0.33	1.60	0.48	1.83	0.49
Range	42.25	7.55	10.40	1.83	1.60	0.68	1.57	0.64
Minimum	0.00	22.85	−1.10	16.37	14.91	16.08	13.65	16.12
Maximum	42.25	30.40	9.30	18.20	16.51	16.76	15.22	16.76
Count	278.00	278.00	278.00	278.00	278.00	278.00	278.00	278.00

Where x variable specifies the first variable to be cross correlated, y variable is the second variable to be cross correlated, $\bar{x}$ is the mean of the first variable values, and $\bar{y}$ is the mean of the second variable values. While, k represents the time index, which can be positive or negative. And r_k is the k^{th} order cross-correlation coefficient.

2.3. Artificial neural network (ANN)

ANN is a computational system comprised of artificial neurons that resample the neurons in the human brain in terms of information transfer to gain knowledge, and it has been utilized heavily in different branches of engineering especially in hydrology where it has been used as a method of prediction [4,33]. ANN has different types of networks that consist of single or multiple layers , such as the feedforward back propagation neural network (FFBP) and it consists of an input layer which is the first layer that receives the data, then a hidden layer, and lastly an output layer that generates specific results drawn from the input data. The flow of information in this type of neural network is only in one direction.

In this study, the FFBP models devolved with an input layer that has one to four neurons, one hidden layer and an output neuron representing the predicted daily groundwater level. The input and hidden layers of the network were connected through a transformation weights (w_{ji}). And a sigmoid transfer was applied as the transformation function for the hidden layer. Fig. 2 shows the structure of the neural network model. Additionally, series of weights (w_{kj})acted as the connection from between the hidden layer and output layer [3], with a linear transfer activation function.

The predicted outcome of the model ($\hat{y}_k$) is given by [13]as follows:

$$\text{Å}\cdot_t = f\left(\sum_{j=1}^{n} w_{kj}h_j + b_k\right) \quad (2)$$

Where w_{kj} is the Weight connecting neuron k of the output layer and the neuron j of the hidden layer, h_j is the output from the hidden layer neuron, n represent the number of hidden neurons, f is the output layers' activation function, and b_j represent the bias of neuron k of the output layer.

2.4. Support vector regression

Support vector machine (SVM) has been used for all kinds of predictive models both in classification and regression, and when

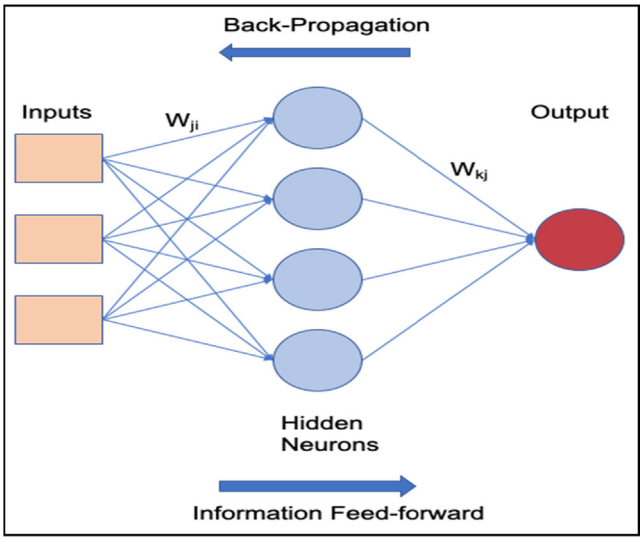

Fig. 2. Simple architecture of a neural network model.

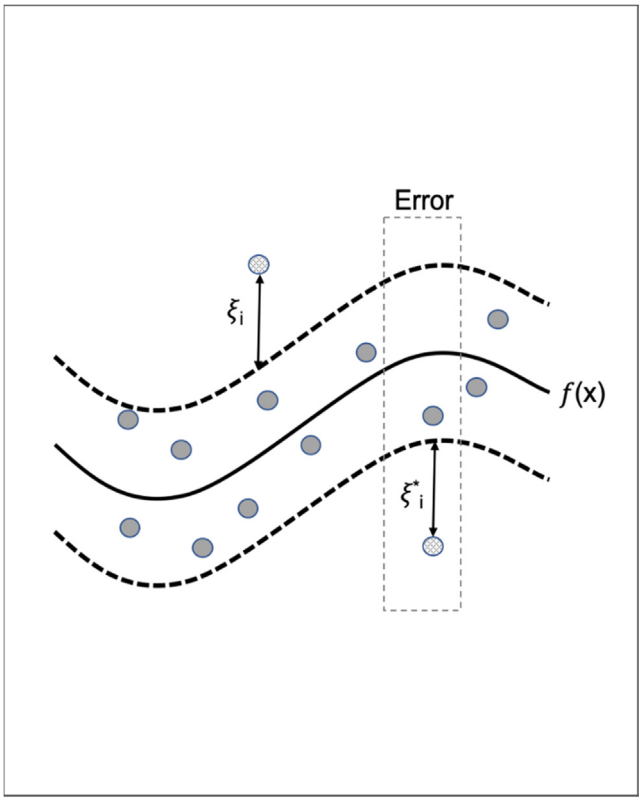

Fig. 3. Nonlinear Support Vector Regression. Errors are acceptable as long as they are less than ε.

SVM is used in regression it is called support vector regression (SVR) [20]. Fig. 3 shows how SVR fits a continuous-valued function to data. In a regression problem model inputs are mapped to a higher dimensional space, and the SVR is trained by a structural risk minimization (SRM) principle [35]. The mapping of inputs to a higher dimensional space is done by a kernel function.

SVR main concept is as follows: consider this dataset as the training data $\{(x_1, t_1), \ldots, (x_n, t_n)\}$, x is the input for the model, and t is the target. The SVR algorithm tries to estimate a function $f(x)$ that has less than ε deviation from the observed target y_i for all the input data values [12]. The SVR function in the form

$$y = w\varphi(x) + b \quad (3)$$

Where the term $\varphi(x)$ represents the non-linearity mapping, w represents a hyperplane, and the term b represent an offset. In addition, SVR uses a penalty function as follows:

$$\begin{cases} |t_i - y_i| \le \varepsilon, notallocatingapenalty \\ |t_i - y_i| > \varepsilon, alloatingapenalty \end{cases} \quad (4)$$

When the predicted value is inside the ε tube, the loss values equals zero. The regression function's parameters can be obtained by minimizing this objective function

$$min\left[\frac{1}{2}\|w\|^2 + c\sum_{i}^{n} L^{\varepsilon}(y_i, t_i)\right] \quad (5)$$

$$L^{\varepsilon}(y_i, t_i) = \max(o, |y_i - t_i| - \varepsilon \quad (6)$$

where c is the regularized constant. When the value of c increases then the relative importance of the empirical risk increases [24].

The assumption through Equation (4), that there is a function f that is able approximate all the data pairs with ε precision. How-

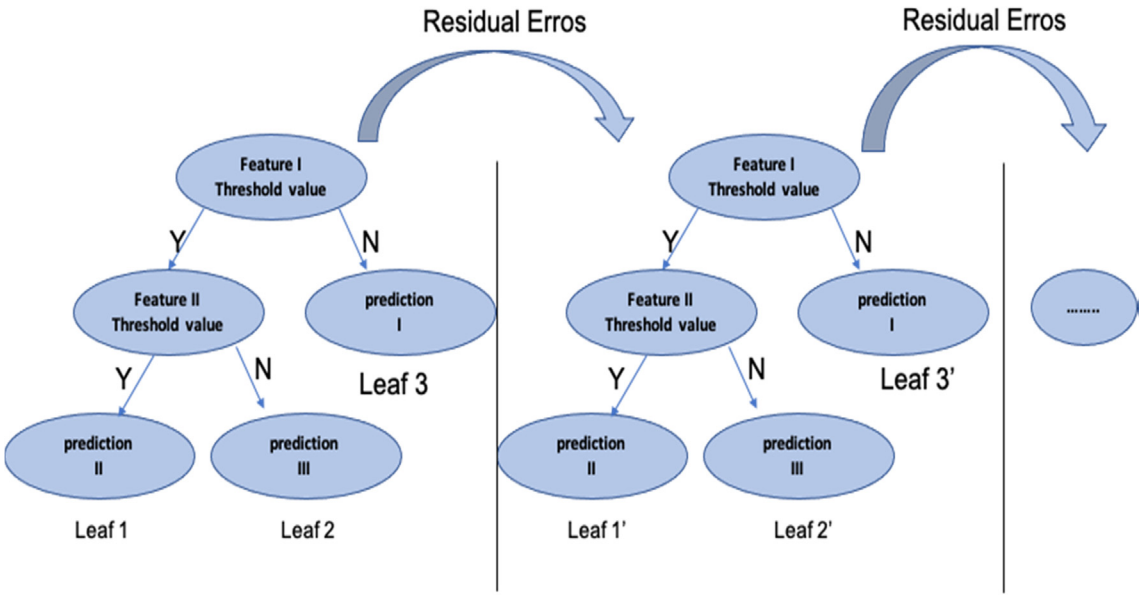

Fig. 4. schematic of Xgboost trees [41].

ever, when this in not possible, slack variables ξ_i, ξ_i^* are introduced [31].

$$min\left[\frac{1}{2}\|w\|^2 + c\sum_{i}^{n}(\xi_i, \xi_i^*)\right] \tag{7}$$

Subject to: $(w\varphi(x_i) + b) - t_i \le \varepsilon + \xi_i$

$$t_i - \left(w^t\varphi(x_i) + b\right) - t_i \le \varepsilon + \xi_i^*$$

$$\xi_i > 0, \xi_i^* > 0$$

the Lagrange multipliers is then used to solve the above optimization problem as follows:

$$max\left[\sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i - \varepsilon\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) - \frac{1}{2}\sum_{i=1,j=1}^{n}(\alpha_i - \alpha_i^*)\left(\alpha_i - \alpha_j^*\right)K(x_i, x_j)\right] \tag{8}$$

Subject to: $0 \le \alpha_i \le c$

$$0 \le \alpha_i^* \le c$$

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0$$

$$K(x_i, x_j) = \varphi(x_i)^T.\varphi(x_j) \tag{9}$$

$K(x_i, x_j)$ represents a nonlinear kernel function, which maps the input from a lower dimension feature space into a higher dimension.

In this study the SVR model has been built using python programming language, and the radial basis function (RBF) kernel is used to map the input. input variables for this model were organized by scenarios, with different combinations. Furthermore, different values of the SVR hyperparameters such as, the epsilon value (ε), cost constant (C) were used to find the best predictive model.

Table 3
Cross-correlation between the input variables and groundwater levels.

Rainfall	r	Temperature	r	Evaporation	r
R(t-1)	0.49	T(t-1)	−0.42	E(t-1)	−0.15
R(t-2)	0.31	T(t-2)	−0.3	E(t-2)	−0.09
R(t-3)	0.24	T(t-3)	−0.26	E(t-3)	−0.046
R(t-4)	0.25	T(t-4)	−0.27	E(t-4)	−0.03
R(t-5)	0.2	T(t-5)	−0.25	E(t-5)	−0.09

Table 4
Selected scenarios for the models built.

	combination
Parameters scenario (1)	
R (t-1)	(1)
R (t-1), R (t-2)	(2)
R (t-1), R (t-2), R (t-3)	(3)
Parameters scenario (2)	
R (t-1), T (t-1)	(1)
R (t-1), E (t-1)	(2)
R (t-1), T (t-1), E (t-1)	(3)
Parameters scenario (3)	
R (t-1), T (t-1), GWL (t-1)	(1)
R (t-1), E (t-1), GWL (t-1)	(2)
R (t-1), T (t-1), E (t-1), GWL(t-1)	(3)

Table 5
The performance of the models in Scenario 1.

combination	ANN	SVR	Xgboost
Scenario 1			
Train(MAE)			
1	0.420	0.307	0.323
2	0.529	0.330	0.333
3	0.587	0.356	0.349
Test(MAE)			
1	0.401	0.382	0.357
2	0.549	0.403	0.365
3	0.571	0.393	0.374
Train(RMSE)			
1	0.815	0.395	0.387
2	0.858	0.417	0.401
3	0.984	0.434	0.418
Test(RMSE)			
1	0.452	0.450	0.433
2	0.681	0.472	0.435
3	0.755	0.482	0.447

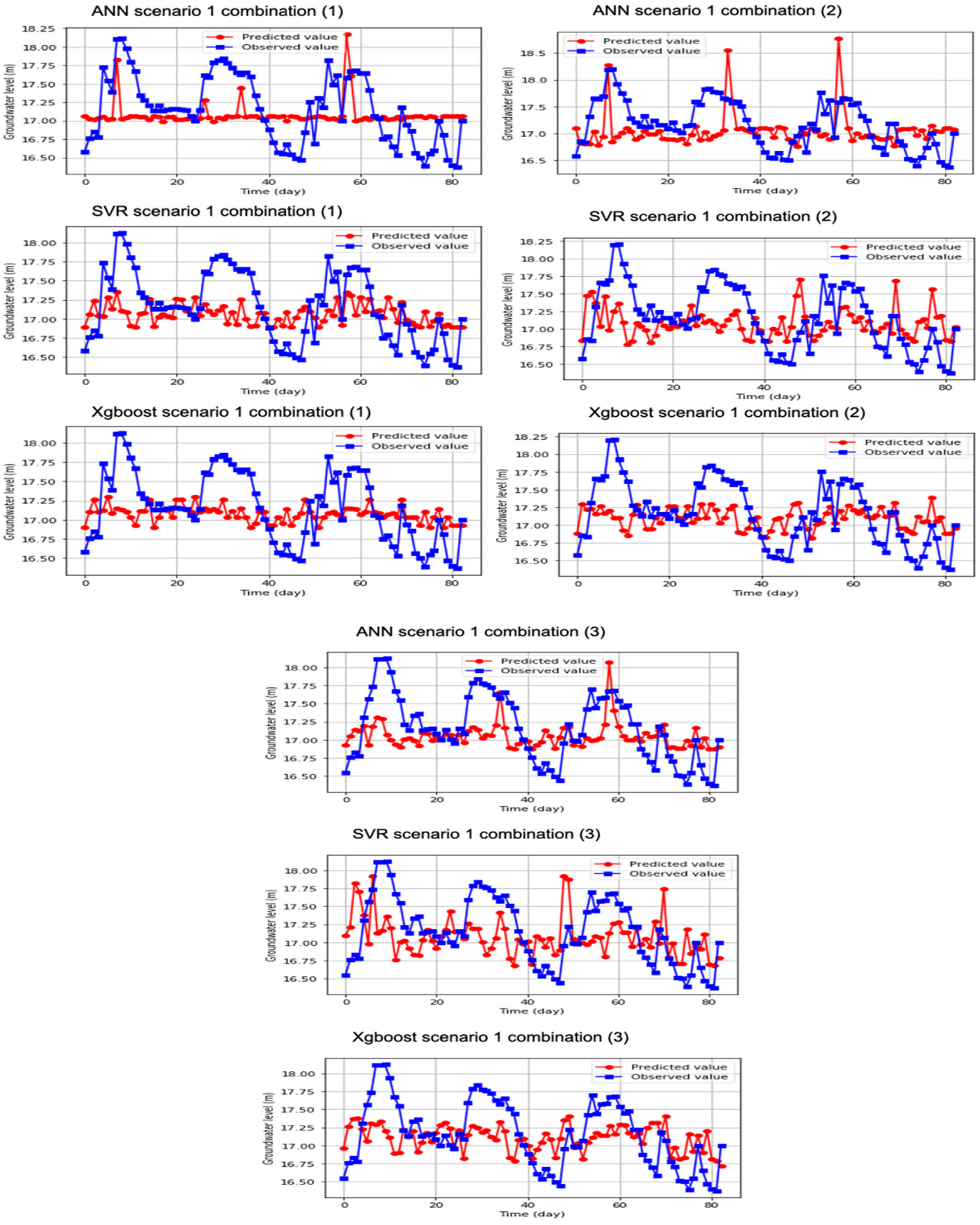

Fig. 5. First input scenario fitting curve of the observed and predicted values obtained using ANN, SVR, and Xgboost models.

2.5. Extreme gradient boosting (Xgboost)

Xgboost is one of the implementations of gradient boosting machines (gbm) which is known as one of the best performing algorithms utilized for supervised learning. It can be used for both regression and classification problems. Xgboost preferred by data scientists because its high execution speed out of core computation [8]. The way the Xgboost works is as follow: If we have for example a dataset DS that has $\boldsymbol{m}$ features and an $\boldsymbol{n}$ number of examples $DS = \{(x_i, y_i) : i = 1 \cdots ..n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$. Let $\hat{y}_i$ be the predicted output of an ensemble tree model generated from the following equations:

$$\text{Å}_{\cdot i} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathscr{F} \tag{10}$$

Where K represents the number of trees in the model as shown in Fig. 4, f_k represents the (k-th tree), to solve the above equation, we need to find the best set of functions by minimizing the loss and regularization objective.

$$\mathscr{L}(\phi) = \sum_i l(y_i, \text{Å}_{\cdot i}) + \sum_k \Omega(f_k) \tag{11}$$

Where l represents the loss function which is the difference between the predicted output $\hat{y}_i$ and the actual output y_i. while Ω is a measure of how complex the model is, this assists in avoiding over-fitting of the model. and it is calculated using:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{12}$$

T, in the above equation represents the number of leaves of the tree, w is the weight of each leaf.

In decision trees to minimize the objective function boosting is used in training the model, which works by adding a new function f as the model keeps training. So, in the t-th iteration a new function (tree) is added as follow:

$$\mathscr{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \text{Å}_{\cdot i}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{13}$$

$$\mathscr{L}_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{14}$$

$$g_i = \partial_{\text{Å}^{\cdot(t-1)}} l\left(y_i, \text{Å}^{\cdot(t-1)}\right)$$

$$h_i = \partial^2_{\text{Å}^{\cdot(t-1)}} l\left(y_i, \text{Å}^{\cdot(t-1)}\right)$$

2.6. Performance measures

Performance of machine learning models needs to be evaluated. To do that, the output generated by the models is compared with the actual values of the groundwater levels, this is done by using the following performance indices:

i) Mean absolute error (MAE), which measures how accurate the predict values compared to the observed ones.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{15}$$

ii) Root mean squared error (RMSE), this index measures the error between the model's output ($\hat{y}$) and the target values (y).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{16}$$

3. Results and discussion

The model inputs were precipitation, evaporation, temperature, and groundwater level with certain selected delays. Cross-correlation between the climatic variables and groundwater levels were found to determine the appropriate delays to be used. Table 3 shows the results of the cross-correlation, after which 3 scenarios in Table 4 were determined to construct the prediction models. We used the mean Absolut error (MAE), root mean squared error (RMSE), coefficient of determination (R^2), and Willmott index (WI) to evaluate the models constructed

3.1. Analysis of first scenario results

Based on the results in Table 5, which shows the statistical measures of different models' performance in predicting the groundwater level using rainfall data as an input (scenario 1), the results obtained are as follows:

1. Comparing the three models, Xgboost performance was better than the performance of the other models. When considering models' performance in the testing stage, then Xgboost had the least MAE based on the first combination (using rainfall with delay of 1 day as an input for the models), which is 11% and 7% less than that of ANN and SVR models respectively.
2. When comparing the performance of the models based on all three input combinations (using different rainfall delays in Table 4) in scenario 1, the first combination (using rainfall delay of 1 day) had the best performance in both training and testing stages for all the models. While models using combination 3 (rainfall delay of 1, 2 and 3 days) as an input preformed the worst.
3. The Xgboost model had better performance than SVR when compared. For example, the RMSE of the Xgboost model considering the testing stage and based on the first input combination (using rainfall with delay of 1 day as an input for the models) was 0.433 and 0.450 for the SVR model.
4. The increase in delay of rainfall resulted in the decrease of models' performance. Fig. 5 depicts the predicted water level for the three input combinations in scenario 1.

3.2. Analysis of the results in scenario 2

Based on the reported results in Table 6, which contains the statistical measures of different models' performance in predicting

Table 6
The performance of the models in Scenario 2.

combination	ANN	SVR	Xgboost
Scenario 2			
Train(MAE)			
1	0.569	0.350	0.332
2	0.555	0.336	0.334
3	0.465	0.320	0.318
Test(MAE)			
1	0.528	0.396	0.371
2	0.663	0.391	0.370
3	0.490	0.363	0.344
Train(RMSE)			
1	1.168	0.429	0.392
2	0.912	0.427	0.405
3	0.735	0.420	0.380
Test(RMSE)			
1	0.682	0.475	0.448
2	0.990	0.490	0.443
3	0.665	0.458	0.435

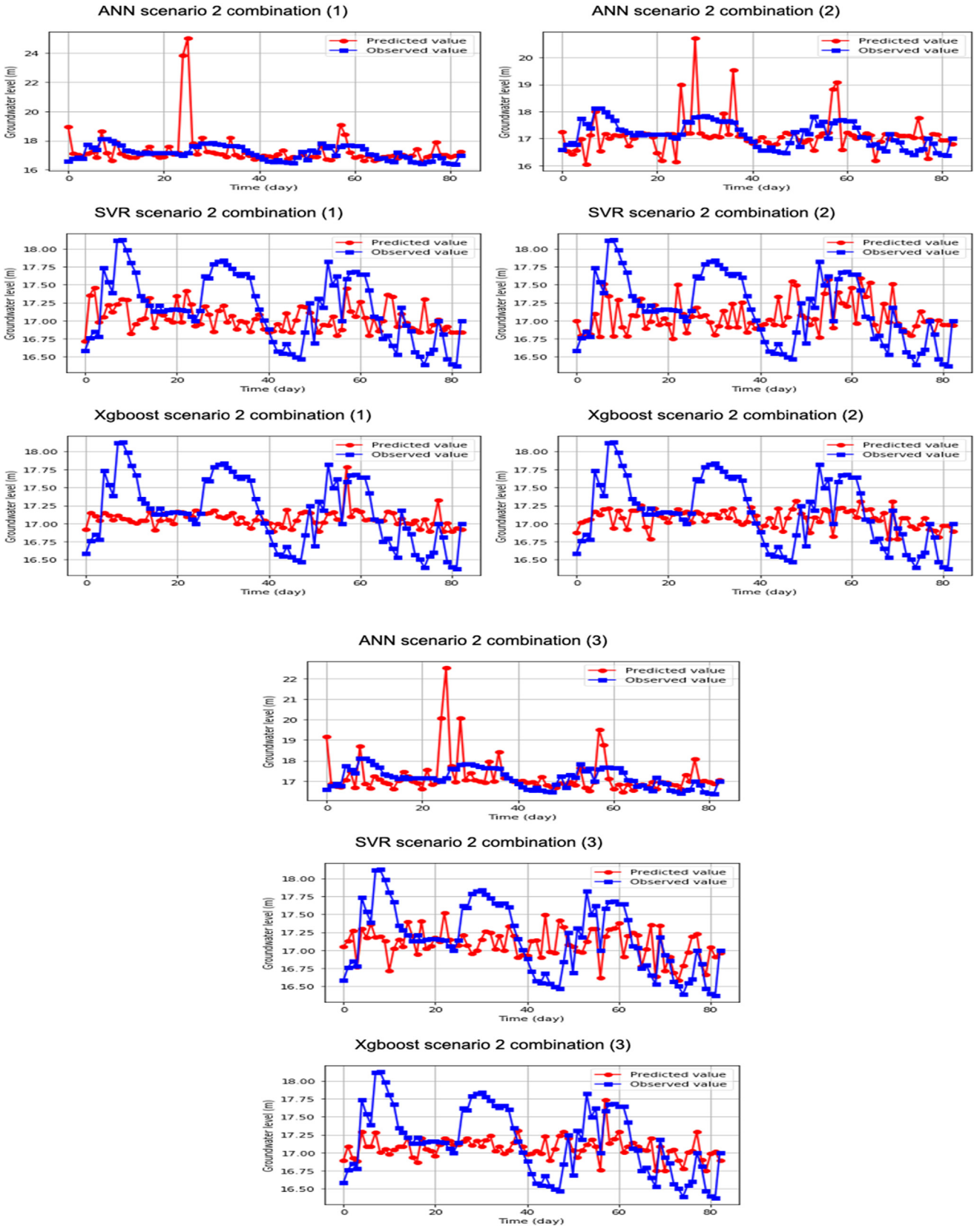

Fig. 6. Second input scenario fitting curve of the observed and predicted values obtained using ANN, SVR, and Xgboost models.

Table 7
The performance of the models in Scenario 3.

combination	ANN	SVR	Xgboost
Scenario 3			
Train(MAE)			
1	0.346	0.080	0.062
2	0.405	0.075	0.060
3	0.175	0.072	0.060
Test(MAE)			
1	0.325	0.118	0.087
2	0.434	0.108	0.089
3	0.254	0.111	0.086
Train(RMSE)			
1	0.625	0.101	0.090
2	0.616	0.097	0.090
3	0.232	0.094	0.087
Test(RMSE)			
1	0.451	0.166	0.138
2	0.704	0.166	0.139
3	0.374	0.162	0.137

the groundwater level using input combinations in (scenario 2), the results obtained are as follows:

Using rainfall of 1 day, temperature of 1 day and evaporation of 1 day (the third combination in scenario 2), resulted in the best performance compared to the other input combinations of in the scenario.
Xgboost model performance was superior when compared to how the other two models preformed. For instance, when considering the performance in the testing stage, Xgboost model result based on the RMSE index was 0.435, which was 35% and 5% better than the ANN and SVR models. Fig. 6 shows the simulation of daily predicted water level using the three models.
ANN model had the worst performance compared to the other two models, since it had the highest values of MAE and RMSE in all three input combinations (using 1 day delay of rainfall with 1 day delay of temperature, 1 day delay of rainfall and 1 day delay of evaporation, 1 day delay rainfall with 1 day delay of temperature and 1 day delay of evaporation) in both training and testing.

3.3. Analysis of the results in scenario 3

The input variables for models in scenario 3 uses the same input combinations as in scenario 2 with 1-day delay of groundwater level added to each combination. Results of the models in this scenario are as shown in Table 7. Based on the third combination, the models performed better in training than in testing stage. As in scenario 2, the third input combination in scenario 3 showed a superiority in performance compared to the other input combinations. For instance, MAE of the Xgboost model in the testing stage using the third input combination was 0.086 compared to 0.254 and 0.111 for the ANN and SVR models respectively.

When comparing the models based on RMSE index, the Xgboost model had the best performance in both training and testing with the a RMSE equaled to 0.137 which slightly better than the SVR, but 63.4% better than the ANN model. based on Fig. 7 it can clearly be seen that the SVR, and Xgboost models are significantly more accurate at predicting the water level compared to the ANN model in this scenario.

3.4. Models' performance based on R^2 coefficient

The Xgboost model had the best performance, then the SVR model and finally ANN model. However, the best input combination varied among the 3 scenarios. For example, in scenario 1, the best input combination was the first combination, while in the second and third scenarios the best input combination was combination 3. Therefore, R^2 were found for all three models based on the best input combination in each scenario.

The R^2 for the Xgboost model was greater than the other models in scenario 1 and 3. While in scenario 2 the SVR model had greater R^2 than that of ANN and Xgboost models. In addition, the R^2 obtained for the Xgboost model in scenario 3 had the greatest R^2 among all the other models in all scenarios as shown in Fig. 8.

4. Conclusions

In this paper, an ensemble model using Xgboost algorithm was constructed to predict the groundwater level. The Xgboost model was compared to standalone ANN, and SVR models for predicting the daily groundwater level in southern Selangor, Malaysia. The data of four influencing factors, namely, rainfall, temperature, water levels of previous days and evaporation, obtained from October 2017 to July 2018 were used as input data for the models. The performance of the prediction models was determined by training and testing the ANN, SVR, and Xgboost models on the sample data, using nine different input combinations. The results showed that the MAE and RMSE values in the optimal and worst training of the ANN, and SVR models were lower than that of the Xgboost model. In addition, the Xgboost model resulted more consistent values and smaller RMSE values in all input combinations, compared with the ANN, and SVR models. Moreover, the Xgboost model presented more accurate prediction results with the highest R^2, and smaller MAE, and RMSE values than the ANN, and SVR models. The prediction of daily groundwater levels using the Xgboost model in an urban area in the southern Part of Selangor, Malaysia was realized in this study. The research findings can guide the scientific utilization of groundwater resources and provide a new approach for similar research in other urban areas with the same characteristics.

Acknowledgments

The authors would like to appreciate the financial support received from Institute of Postgraduate Studies and Research (IPSR) of Universiti Tunku Abdul Rahman, Malaysia for covering the APC. In addition to that, the authors would like to acknowledge the Innovation & Research Management Center (iRMC) of Universiti Tenaga Nasional for their technical and financial support provided under the grant coded RJO10517844/088 by Innovation & Research Management Center (iRMC), Universiti Tenaga Nasional.The authors also would like to thank National Water Research Institute of Malaysia (NAHRIM) for providing the data to conduct this study.

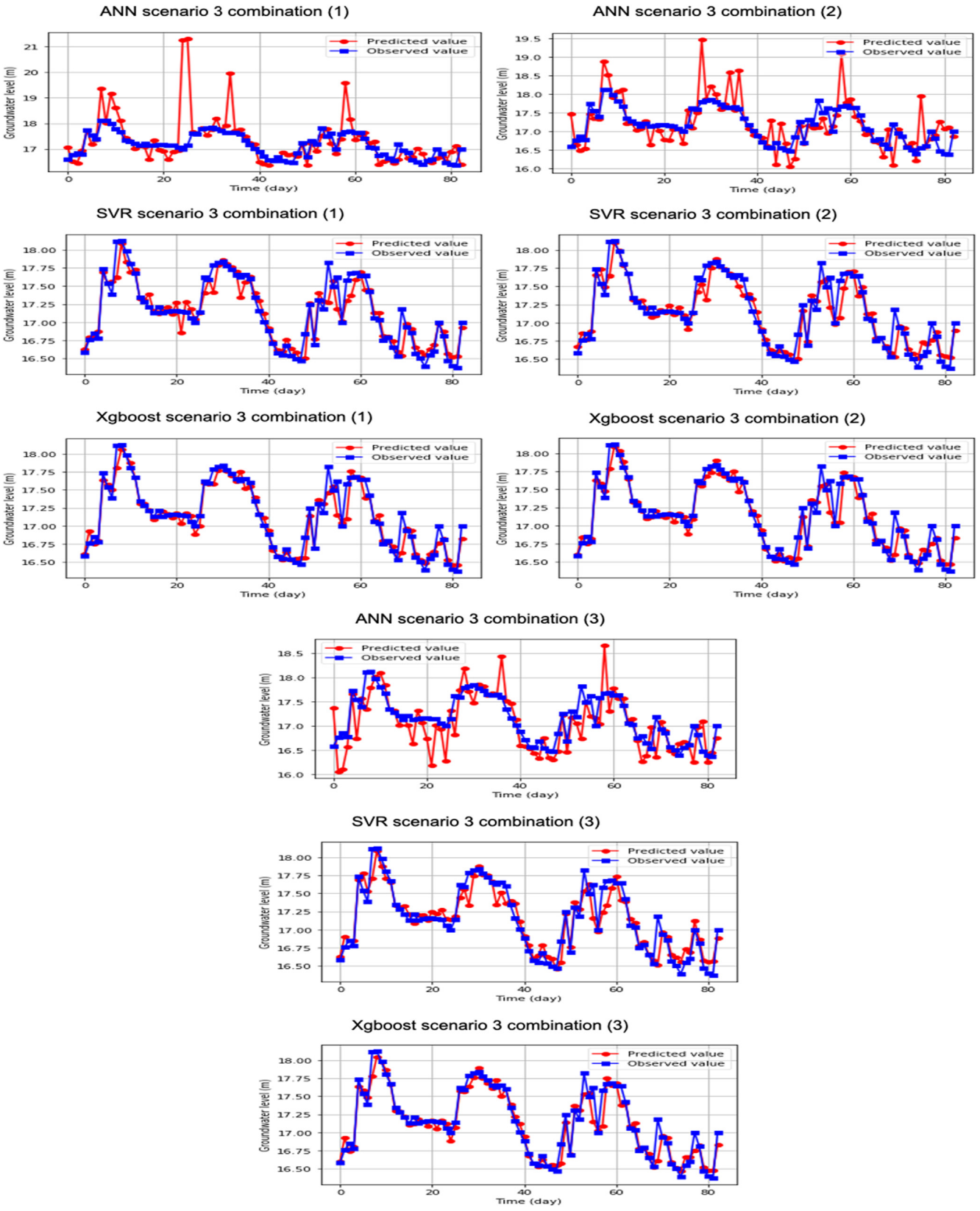

Fig. 7. Third input scenario fitting curve of the observed and predicted values obtained using ANN, SVR, and Xgboost models.

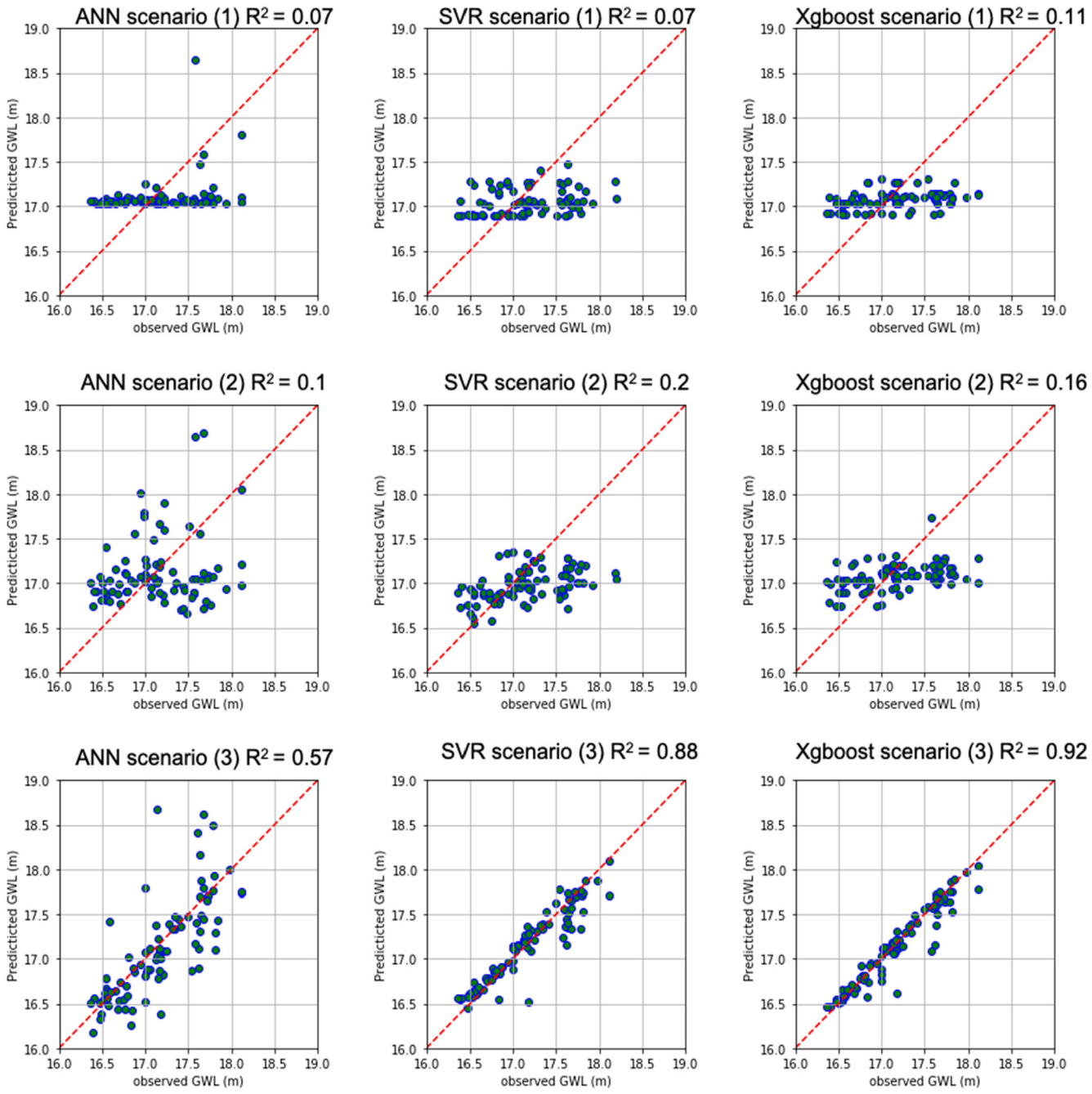

Fig. 8. Comparison of R^2 among all the best preforming models in each scenario.

References

[1] Adamowski J, Chan HF. A wavelet neural network conjunction model for groundwater level forecasting. J. Hydrol. 2011;407:28–40.
[2] AK, L., G, K., 2015. Groundwater Level Simulation Using Artificial Neural Network in Southeast, Punjab, India. J. Geol. Geophys. 04, 206.
[3] Ali I, Alharbi OML, Alothman ZA, Badjah AY, Alwarthan A, Basheer AA. Artificial neural network modelling of amido black dye sorption on iron composite nano material: Kinetics and thermodynamics studies. J. Mol. Liq. 2018;250:1–8.
[4] Alizamir M, Sobhanardakani S, Taghavi L. Modeling of Groundwater Resources Heavy Metals Concentration Using Soft Computing Methods: Application of Different Types of Artificial Neural Networks. Risks: J. Chem. Heal; 2017. p. 7.
[5] Barzegar R, Fijani E, Asghari Moghaddam A, Tziritis E. Forecasting of groundwater level fluctuations using ensemble hybrid multi-wavelet neural network-based models. Sci. Total Environ. 2017;599–600:20–31.
[6] Boelee L, Bahrom R, Amer H, Sondor NZ, Brown E, Ahmad F, et al. Operational decision support system for sustainable water resource management for Sungai Selangor. 37th IAHR. World Congr; 2017. p. 1–16.
[7] Chang FJ, Chang LC, Huang CW, Kao IF. Prediction of monthly regional groundwater levels through hybrid soft-computing techniques. J. Hydrol. 2016;541:965–76.
[8] Chen T, Guestrin C. In: XGBoost: A scalable tree boosting system. Association for Computing Machinery; 2016. p. 785–94.
[9] Craig CA, Feng S, Gilbertz S. Water crisis, drought, and climate change in the southeast United States. Land use policy 2019;88:104110.
[10] Department of Statistics Malaysia, 2010. Population Distribution and Basic Demographic Characteristic Report 2010.
[11] Goodarzi M. Application and performance evaluation of time series, neural networks and HARTT models in predicting groundwater level changes, Najafabad Plain. Iran. Sustain. Water Resour. Manag. 2020;6:1–10.
[12] Granata F, Gargano R, de Marinis G. Support Vector Regression for Rainfall-Runoff Modeling in Urban Drainage: A Comparison with the EPA's Storm Water Management Model. Water 2016;8:69.
[13] Haykin, S., 2009. Neural networks and learning Third Edition, Institute of Physics Conference Series.
[14] Hosseini Z, Nakhaie M. Estimation of groundwater level using a hybrid genetic algorithm-neural network. Pollution 2015;1:9–21.

[15] Huang, M., Tian, Y., 2015. Prediction of Groundwater Level for Sustainable Water Management in an Arid Basin Using Data-driven Models 134–137.
[16] International Decade for Action "Water for Life" 2005-2015. Focus Areas: Water scarcity, n.d.
[17] Iqbal M, Ali Naeem U, Ahmad A, Rehman H ur, Ghani U, Farid T. Relating groundwater levels with meteorological parameters using ANN technique. Meas. J. Int. Meas. Confed. 2020;166:108163.
[18] Jagtap, A.S., Kavitha, K.V.N., Hussein, A.D., 2019. Monitoring of Groundwater level and Development of Control Mechanism based on Machine Learning Algorithm. In: Proceedings - International Conference on Vision Towards Emerging Trends in Communication and Networking, ViTECoN 2019. Institute of Electrical and Electronics Engineers Inc.
[19] Kasiviswanathan KS, Saravanan S, Balamurugan M, Saravanan K. Genetic programming based monthly groundwater level forecast models with uncertainty quantification. Model. Earth Syst. Environ. 2016;2:1–11.
[20] Lai V, Ahmed AN, Malek MA, Afan HA, Ibrahim RK, El-Shafie Ahmed, et al. Modeling the Nonlinearity of Sea Level Oscillations in the Malaysian Coastal Areas Using Machine Learning Algorithms. Sustain. 2019;11.
[21] Le Brocque AF, Kath J, Reardon-Smith K. Chronic groundwater decline: A multi-decadal analysis of groundwater trends under extreme climate cycles. J. Hydrol. 2018;561:976–86.
[22] Li H, Lu Y, Zheng C, Yang M, Li S. Ground water level prediction for the arid oasis of Northwest China based on the artificial bee colony algorithm and a back-propagation neural network with double hidden layers. Water (Switzerland) 2019;11:1–20.
[23] Lorenzo-Lacruz J, Garcia C, Morán-Tejeda E. Groundwater level responses to precipitation variability in Mediterranean insular aquifers. J. Hydrol. 2017;552:516–31.
[24] Luo X, Yuan X, Zhu S, Xu Z, Meng L, Peng J. A hybrid support vector regression framework for streamflow forecast. J. Hydrol. 2019;568:184–93.
[25] Mirarabi A, Nassery HR, Nakhaei M, Adamowski J, Akbarzadeh AH, Alijani F. Evaluation of data-driven models (SVR and ANN) for groundwater-level prediction in confined and unconfined systems. Environ. Earth Sci. 2019;78:1–15.
[26] Moghaddam Hamid Kardan, Moghaddam Hossein Kardan, Kivi ZR, Bahreinimotlagh M, Alizadeh MJ. Developing comparative mathematic models, BN and ANN for forecasting of groundwater levels. Groundw. Sustain. Dev. 2019;9:100237.
[27] Mohanty S, Jha MK, Kumar A, Panda DK. Comparative evaluation of numerical model and artificial neural network for simulating groundwater flow in Kathajodi-Surua Inter-basin of Odisha. India. J. Hydrol. 2013;495:38–51.
[28] Porte P, Kumar Isaac R, Singh Mahilang KK, Sonboier K, Minj P. Groundwater Level Prediction Using Artificial Neural Network Model. Int. J. Curr. Microbiol. Appl. Sci. 2018;7:2947–54.
[29] Reinecke R, Wachholz A, Mehl S, Foglia L, Niemann C, Döll P. Importance of Spatial Resolution in Global Groundwater Modeling. Groundwater 2020;58:363–76.
[30] Sahoo S, Russo TA, Elliott J, Foster I. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. Water Resour. Res. 2017;53:3878–95.
[31] Sattari MT, Mirabbasi R, Sushab RS, Abraham J. Prediction of Groundwater Level in Ardebil Plain Using Support Vector Regression and M5 Tree Model. Groundwater 2018;56:636–46.
[32] Tubau I, Vázquez-Suñé E, Carrera J, Valhondo C, Criollo R. Quantification of groundwater recharge in urban environments. Sci. Total Environ. 2017;592:391–402.
[33] Wanto A, Windarto AP, Hartama D, Parlina I. Beck'scher Studienfuhrer Jura : Universitaten, Literatur, Tipps, Adressen.. IJISTECH (International J. Inf. Syst. Technol. 2017;1:43–54.
[34] Wei Z lei, Wang D Fei, Sun H Yue, Yan X. Comparison of a physical model and phenomenological model to forecast groundwater levels in a rainfall-induced deep-seated landslide. J. Hydrol. 2020;586:124894.
[35] Wu J, Liu H, Wei G, Song T, Zhang C, Zhou H. Flash Flood Forecasting Using Support Vector Regression Model in a Small Mountainous Catchment. Water 2019;11:1327.
[36] Yadav B, Ch S, Mathur S, Adamowski J. Assessing the suitability of extreme learning machines (ELM) for groundwater level prediction. J. Water L. Dev. 2017;32:103–12.
[37] Yoon H, Hyun Y, Ha K, Lee KK, Kim GB. A method to improve the stability and accuracy of ANN- and SVM-based time series models for long-term groundwater level predictions. Comput. Geosci. 2016;90:144–55.
[38] Yoon H, Jun SC, Hyun Y, Bae GO, Lee KK. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. J. Hydrol. 2011;396:128–38.
[39] Zhou T, Wang F, Yang Z. Comparative analysis of ANN and SVM models combined with wavelet preprocess for groundwater depth prediction. Water (Switzerland) 2017;9:781.
[41] Dong W, Huang Y, Lehane B, Ma G. XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring. Automation in Construction 2020;114:103155.