Studying Characteristics of Huay Saneng and Predicting The Water Levels by Comparison with Other Regression Models

Sakchan Luangmaneerote* and Boonlueo Nabumroong*

Received : April 19, 2021
Revised : June 10, 2021
Accepted : July 1, 2021

Abstract

The drought problem in Surin Province has become complex issue over the course of the last five years. Water level in reservoir is directly associated with drought. Therefore, understanding the root cause of the problem is of utmost importance. This research investigated which attributes were subject to drought in a reservoir called Huay Saneng and which machine learning regression models showed the best accuracy using data from the last 5 years; with 1,825 records derived from Royal Irrigation Department operation on Huay Saneng area. The research finding established that Decision tree regression showed satisfactory results with 94.7% reliability; better than Multiple linear regression, Polynomial regression, and Random forest regression. The use of four attributes consisting of date, evaporation, water levels and rain can assist to make prediction with satisfactory results. In addition, the research found that the levels of evaporation of water is an important factor in assisting prediction through the process of feature selection. The research discussed how conservation of water levels in the past five years tended to be less inclined and how future research may be improved in order to predict drought accurately in the future.

Keywords: Drought, Regression models, Huay Saneng.

1. Introduction

Water scarcity has become the norm these days, especially in Thailand. Several reasons for this are due to the high demand for water for agriculture, excessive consumption, or climate change [1]. Drought affects both surface and groundwater and can lead to a reduced water supply. These lead to crop failure and unexpected damages. Understanding drought is a key component of interest to hydrologists meteorologists and data scientists. Managing the problem of drought lies mainly in planning and managing water resources [2].

Drought forecasting is a key component of drought hydrology, which plays a crucial role in drought risk management, preparedness, and mitigation. Nowadays, there are many drought modeling such as identifying drought and predicting the duration of severity.

Thus, prediction of drought is a more challenging problem, but in the past five years, Surin has faced more frequent drought [3]. The largest and most intense drought was in 2019. It was extremely difficult for Surin residents. Many hotels lacked water to supply their guests. The hospitals' reserves were too depleted to care for patients. Various houses had deficient in water for consumption. These disasters inflict immense damage to the economy and the suffering of the people living in Surin Province.

A machine learning model is a file that has been trained to understand some types of patterns. A trained model can be used to make predictions and made decision-based on logic without programming [4], [5]. Machine learning algorithms have ability to learn both linear and non-linear such as rain prediction, drought prediction and heat wave prediction. Many machine learning algorithms, such as Random Forests, Support Vector Machine, k-Nearest neighbors, Artificial Neural Network, Decision Tree, are capable of using in complex patterns.

* *Department of Computer Technology, Faculty of Agriculture and Technology, Rajamangala University of Technology Isan, Surin Campus.*

Currently, recent research studies have found that the machine learning model has been implemented in various fields such as forecasting of daily inflow to large reservoir [6], forecasting reservoir inflow [7], daily rain forecast [8], estimation of missing rainfall data [9], flood forecast [10][11], forecasting of river flow rates [12], and prediction of groundwater[13].

In terms of applying models, multiple linear regression models can be used to predict various ways such as drought weather variables, groundwater level [14], or number of wet days [15], forecasting bicycling rental [16] as well as polynomial regression that fits a nonlinear data.

Artificial neural network (ANN)-based techniques and the support vector machine (SVM) are considered as standard nonlinear estimators [17], [18], [19], [20]. However, many research suggest that the two models have many drawbacks such as taking more time [21], [22] affecting outliers and redundant data [23]. According to several papers [24], [25], [26] they suggested that random forest regression and decision tree showed satisfactory results. Thus, this research used the two models to rescue. The research required the knowledge of when to apply two models for the real problem and whether these models can be used effectively by comparing with other regression models.

This paper investigated four regression models, consisting of multiple linear regression, polynomial regression, decision tree regression, and random forest regression, with the purpose of finding the optimal model to predict the water level of Huay Saneng. In addition, this research revealed the real situation of water level over the last 5 years, between 2016 and 2020, in order to apply the results of the study in preventing drought in the future. If the water level can be predicted accurately, this can assist the local officers to notify supervisors and prevent the drought early. Besides, this study will extend the scope of which key factors contribute to drought and water management.

2. Theoretical background and related researches

2.1 Background of Huay Saneng

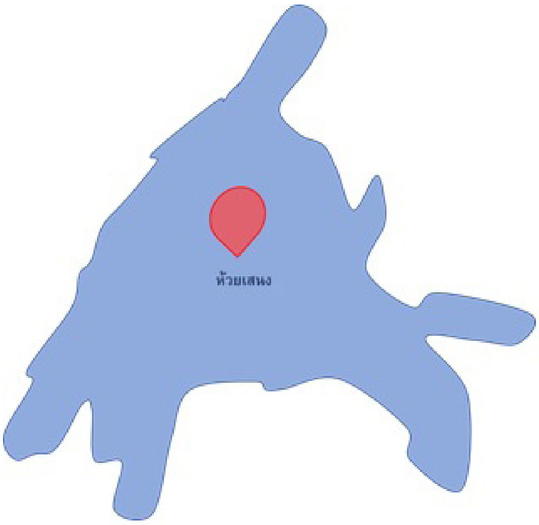


Figure 1. *Shape of Huay Saneng.*

The Huay Saneng reservoir in Surin province is a symbol of the recreational tourist area, where is a source of water for production. Huay Seneng is an important reservoir to the people of Surin Province, starting from the upstream and sub-branches that flow together to form Huay Saneng. Another important source of water comes from the Phanom Dong Reg mountain range which the local people call Phanom Sor. Huay Saneng has a capacity of 20 million cubic meters in the Niang Subdistrict. There is also a twin reservoir, an Ampuen reservoir with a capacity of 22 million cubic meters in the Tenmey Subdistrict, which is the source of reserve for the reservoir at Huay Saneng. The shape of Huay Saneng reservoir is shown in figure 1.

2.1 Attributes leading to drought

Drought is an environmental and economic threat. It can occur in almost all climatic areas. Predicting drought used many attributes. Many research suggested that changing temperature and precipitation resulted in drought because increasing temperatures can be implied to increase evaporation [27], [28] while some suggested that soil moisture lead to inability to maintain water [29]. The diverse climateconditions can cause droughts which the system should adapt needs of specific regions [30]. However, the cause of drought may come from unexpected attributes such as the expansion of the increasing population, the agricultural sector, industrial plants in local area or climate change. Unfortunately, these attributes are not stored in Excel file of Royal Irrigation Department operation on Huay Saneng area. Therefore, this research used only attributes kept by Royal Irrigation Department.

2.2 Multiple linear regression

Multiple linear regression is a technique to predict the outcome based on two or more variables, extended from linear regression [31]. The dependent variable are the predicted variables while variables used to predict are called independent variables. The equation for multiple linear regression model is shown as equation 1.

$$y = \beta_0 + \beta_1 + \ldots\ldots + \beta_n X_n \qquad (1)$$

Given equation 1, y is a dependable variable. β_n and β_0 , β_1 are regression coefficient while X_1 and X_n are independent variable. In general, multiple linear regression is used in various ways such as predicting blood pressure [32], predicting effect of fertilizer and water on crop yields [33], or the effect of different training regimens have on player performance [34].

2.3 Polynomial regression

Polynomial regression is a special case of multiple linear regression which is designed with the data with a curvilinear relationship between the target variable and the independent variables [35]. The polynomial regression is suitable for the value of the target variable that changes in case of a non-uniform manner relating to the predictor. Generally, the equation for the polynomial linear regression model is shown as equation 2.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^n + \ldots\ldots + \beta_n X_i^n + (i = 1, 2, \ldots. n) \qquad (2)$$

In general, the equation of polynomial in equation 2 is relatively similar to the equation of multiple linear regression while *n* is degree of polynomial. Normally, increasing the degree in the polynomial model results in increasing the performance of the model and creating more complicated algorithm. However, one should be aware of the risk of model overfitting and underfitting when adjusting more and more degree. There are many research studies which have used polynomial regression, such as predicting longitudinal dispersion coefficient in rivers [36], the trend of water quality [37], interpolation of rainfall through polynomial regression [38] ,and water supply-demand [39].

2.4 Decision tree

Decision tree was proposed by Quinlan [40] which he extended the principle of Occam's razor in attempting to build the smallest decision tree. Decision tree forms tree structure which breaks down the dataset into smaller subsets while slowly creating a related decision tree. The decision tree is classified as a supervised learning algorithm. It can function with both continuous and categorical output variables. The branches and edges represent the outcome nodes and the nodes have either: Conditions (Decision Nodes) or Result (End Nodes). Decision tree regression trains a model in a tree's structure to make a prediction. Decision tree has been used in various ways such as, predicting flooding in China [41] and forecast drought [42].

2.5 Random forest

Random forest is a continued development of the bagging method [43] as it combines both bagging and feature randomness to build an unassociated forest of decision trees. The Random Forest algorithm is one of the most popular algorithms which can be used in both classification and regression. The concept of random forest is comprised of various decision trees inside an increased number of decision tree resulting in a more complex algorithm. The algorithm chooses the best vote of decision tree from all of them. Unlike a decision tree, a random forest classifier chooses the features randomly to build several decision trees in order to fix the overfitting problem. There are several research studies suggesting to use random forest algorithm with various reasons such as taking less time [21], [22] unaffected to outliers, and redundant data [23].

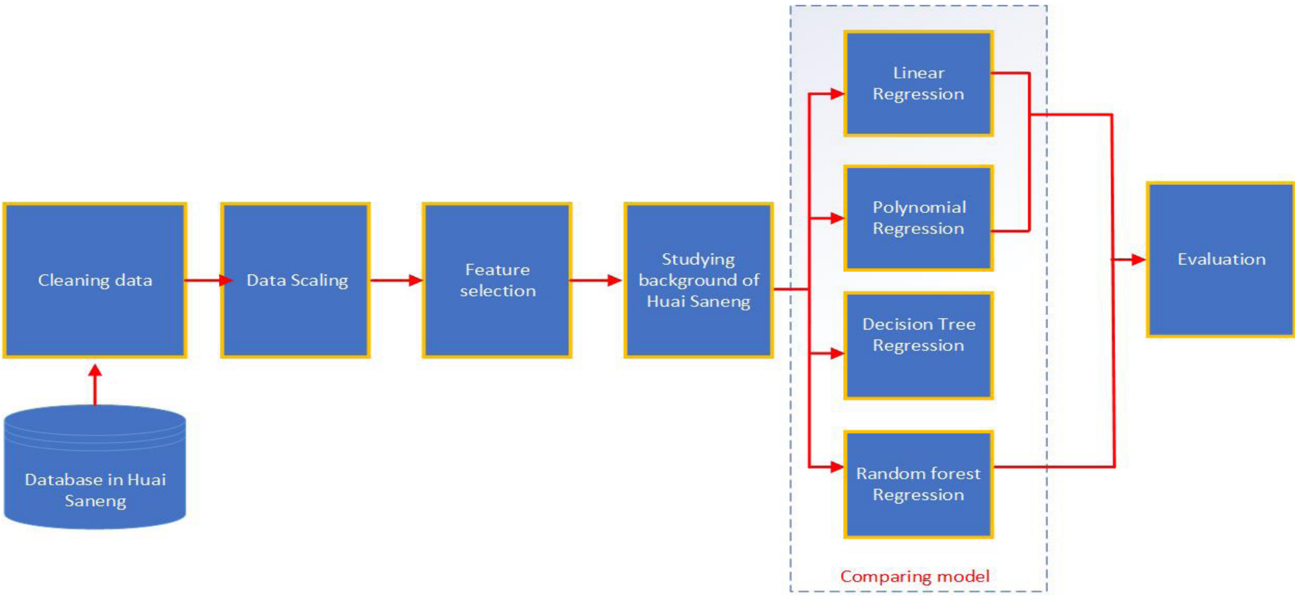


Figure 2. *Steps of research methodology.*

3. Research Methodology

The stored data in Royal Irrigation Department operating on Huay Saneng area between 2016 and 2020 was utilized for this research. The data collected had 1,825 records with 4 attributes, consisting of date, evaporation of water in Huay Saneng, water levels, and rainfall. The scikit-learn in Python was used for the analysis of data, while other packages were used for visualization. The steps of the research are shown in figure 2.

3.1 Data cleaning

Usually, obtained data cannot be used with the machine learning model instantly [44]. Thus, data cleaning is an imperative process for preparing data. This step Saneng. The data was comprised of 4 attributes, consisting of date, evaporation of water, water levels and rainfall. All data was relatively clean except for rainfall, which had been replaced some of null value with previous value. All data after cleaning were still maintained at 1,825 records, which was data from 2016 to 2020.

3.2 Data scaling

Data scaling is a particular method used to standardize a range of data because sometimes the range of data values can be varied [45]. This step is necessary in preprocessing data before putting the data into machine model. For this research, the obtained data possessed the different units, such as evaporation of water, rainfall, and water levels. Therefore, it was imperative to scale the three attributes before the data was used for further processing.

3.3 Feature selection

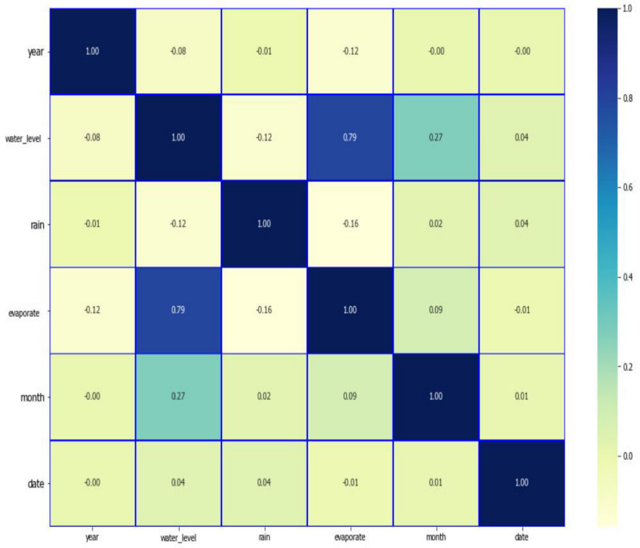


Figure 3. *Correlation heatmap of attributes.*

There are various ways to strengthen a model. Feature selection is the one of necessary step of the developmental process. The main goal is to decrease the number of input variables and study characteristic of obtained attributes

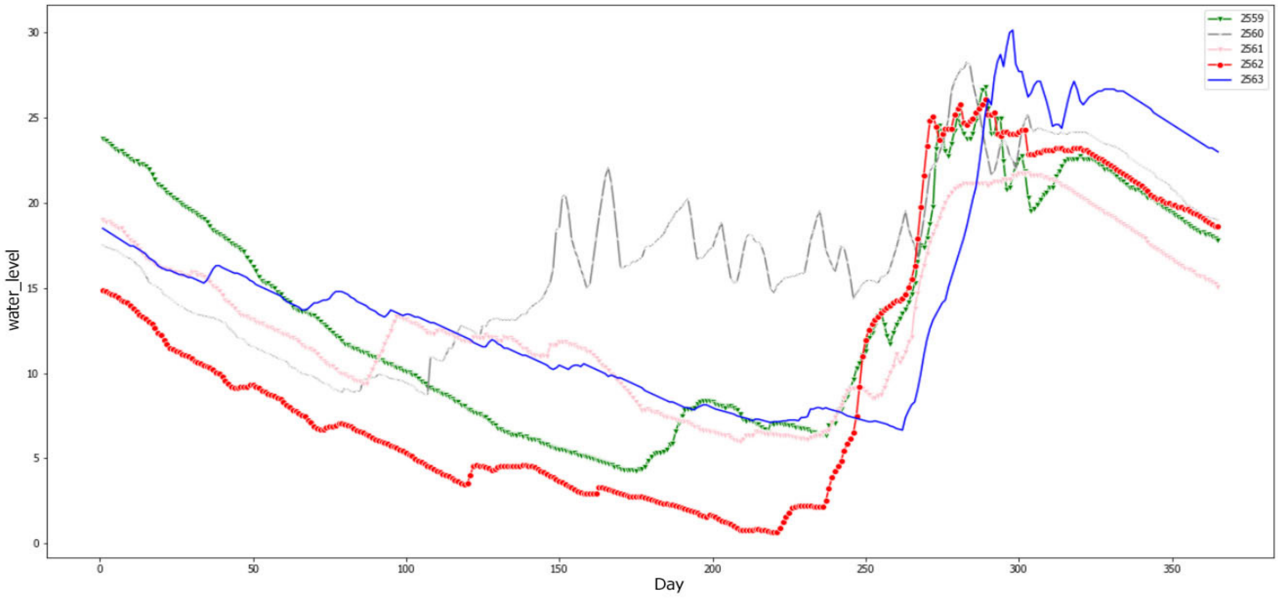
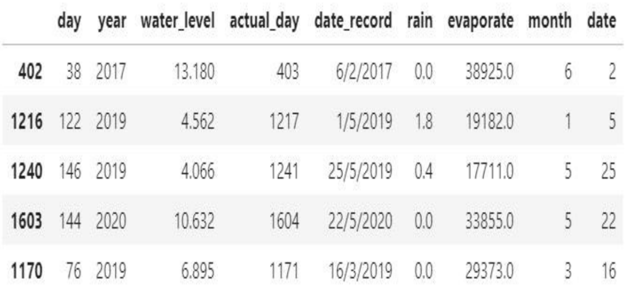


Figure 4. *Overall trend of water level between 2016 and 2020.*

before building a predictive model. Understanding the characteristics of attributes will improve the performance of the predictive model. Feature selection can be classified into three types: filter method, wrapper method, and embedded method. Using correlation matrix with heatmap is classified as a type of filtering method. The correlation coefficient with heatmap normally yields values between -1 and 1. Any value closer to 1 shows a stronger positive correlation, while a value closer to 0 shows a weaker correlation as well as a stronger negative correlation yields a value closer to -1. This step is required to deeply understand which attributes interact with each other, as the obtained data is necessary to study which attributes to use in the model. Given figure 3, it can be seen that the evaporation of water were considered to have the most positive correlation with water level, followed by month. Surprisingly, rainfall did not show a strong correlation with water level. Even though some attributes showed a weak correlation, the research had decided to put all attributes into a predictive model while date will be put into an index of a dataframe. The sample data used to create predictive model as shown in figure 5.



	day	year	water_level	actual_day	date_record	rain	evaporate	month	date
402	38	2017	13.180	403	6/2/2017	0.0	38925.0	6	2
1216	122	2019	4.562	1217	1/5/2019	1.8	19182.0	1	5
1240	146	2019	4.066	1241	25/5/2019	0.4	17711.0	5	25
1603	144	2020	10.632	1604	22/5/2020	0.0	33855.0	5	22
1170	76	2019	6.895	1171	16/3/2019	0.0	29373.0	3	16

Figure 5. *Sample data used for making predictive model.*

3.4 Studying background of Huay Saneng

At this stage, it is necessary to conduct an extensive survey of the nature of Huay Saneng between 2016 and 2020, with the aim of understanding characteristics of rainfall and the factors affecting the water level in Huay Saneng, through the obtained data. This stage assisted to create a straightforward view and broader picture of potential trends. The results of the survey will benefit the sustainable water management of Huay Saneng in the future. This stage used Seaborn and Matplotlib which are packages used in Python for virtualization.

3.5 Modeling

Modeling consisted of three steps: selecting modelling technique, generating a test design, and building the model [46].

This step selected four algorithms consisting of multiple liner regression, polynomial regression, decision tree regression, and random forest regression, and then separated data into test set and training set with a proportion of 30% for test set and 70% for training set. Training set would be used to train model while test set was used for testing the model.

3.6 Evaluation

At the end of this stage, a decision was made regarding the obtained results. Four machine learning models were evaluated with coefficient of determination for both training set and test set, and line graph was used to show the overall result of prediction. The main objective of the evaluation was to identify which model yielded the best result in order to adjust the model for use in the future.

4. Results of experiments

This section is split into three main parts. The first part revealed the results of the background data of Huay Saneng over the last five years while the second part showed the results of the four models. The third part discussed experiments and future research.

4.1 Results of studying background of Huay Saneng

This part is comprised of three parts. The histogram graph was used in the first part with the specific purpose of illustrating the trend of the obtained data. The second part illustrated a pair plot graph which showed the relationship of the general characteristics of factors. Finally, a line graph was used to show the overall trend of water levels over the last 5 years.

1) Results of overall trend of water levels

The line graph in Figure 4 shows changes in water level in Huay Saneng from 2016 to 2020. The x-axis shows the number of days within 1 year, while the y-axis shows the water level measured manually by the local authority of Huay Saneng.

According to the graph, overall, it is clear that the water level gradually declined early in the year and rose again between July and September. The water level set to decline again after September.

It can be noticed that the water level showed the lowest level in 2012 from early in the year and fell sharply until reaching the lowest point around June. This was the year that Surin experienced an unprecedented drought. However, this pattern of water levels was not the same every year. The water level rose rapidly in 2017 around April, which did not occur frequently.

2) Results of relationship of attributes

For this section, the obtained pair plots were derived from the seaborn virtualization library in Python. The graph in figure 6 shows the relationship of the three factors, consisting of evaporation of water, water levels, and rainfall. Using pair plots assisted to reveal the relationship of data between two variables.

This study showed that amount of evaporation implied that the water level was related to the evaporation of water. If the evaporation of water was high, the water level stood at a good stage. Surprisingly, the rainfall amount did not contribute to the increase of water level as expected. It can be seen that Huay Saneng has several methods of retaining water besides relying only on rainfall.

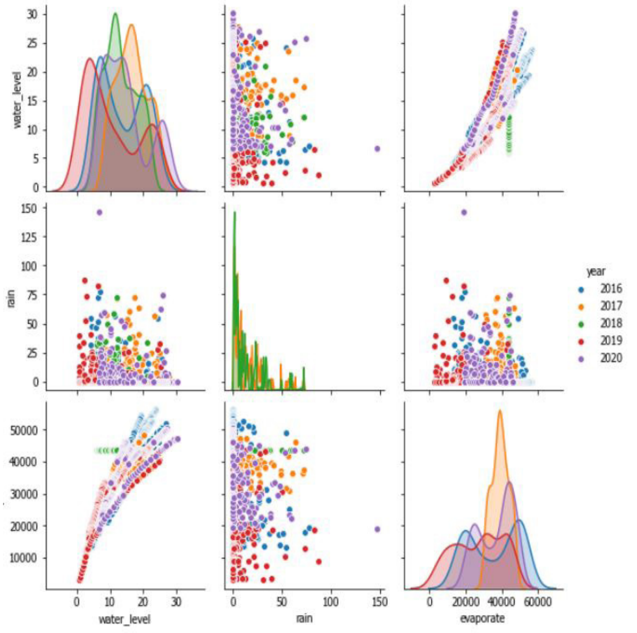


Figure 6. Relationship of water level, rainfall, and evaporation.

3) Results of distribution of the three factors.

Regarding histogram graph from figure 7, figure 8, and figure 9 these findings suggested that most of the water retained in Huay Saneng had remained relatively low level over the past five years in regard to figure 7 showing a right-skewed distribution. Given figure 8, the evaporation of water showed a right-skewed distribution. It can be assumed that the temperature was probably high in Surin area. This assumption would be proved if the department operating on Huay Saneng area collect the highest and lowest temperature in the future. In the part of rainfall in figure 9, the rainfall amount was relatively row over the last 5 years. It can be proved that Huay Saneng area can collect water from other sources which need to be investigated further.

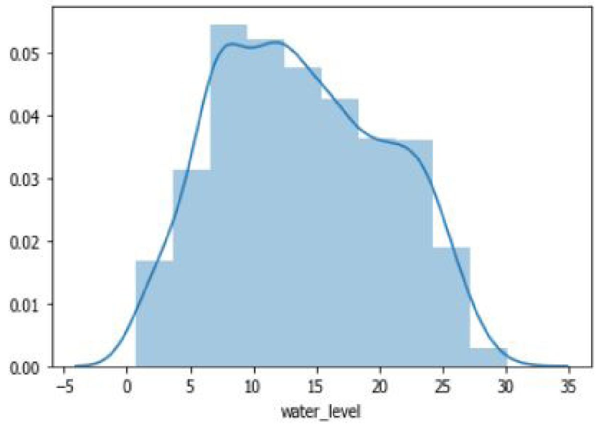


Figure 7. *Distribution of water level over the last 5 years.*

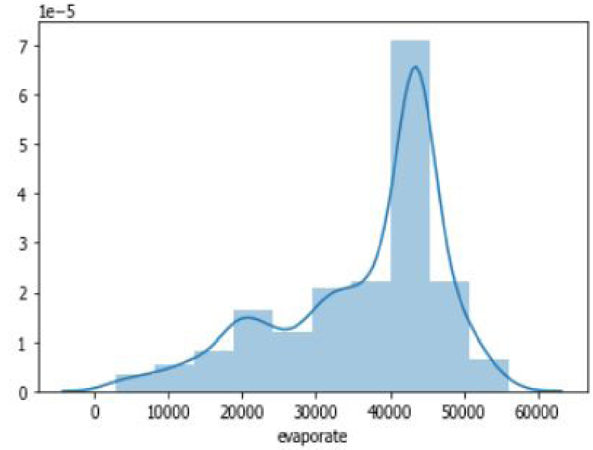


Figure 8. *Distribution of water evaporation over the last 5 years.*

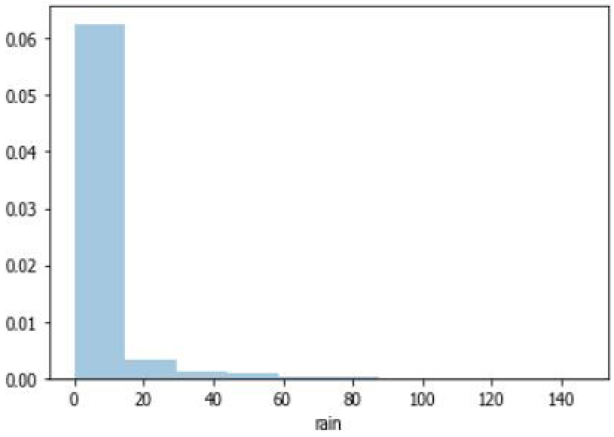


Figure 9. *Distribution of rainfall over the last 5 years.*

4.2 The evaluation of four machine learning models

This research used four machine learning models, consisting of multiple linear regression, polynomial regression, decision tree regression, and random forest regression, to compare the efficiency of each model. The four models yielded the R squared, known as the coefficient of determination, mean squared error regression loss (MSE), and used a line graph to virtualize in order to view the overall trend of the actual data and predicted value through the selected machine learning model.

Normally, the coefficient of determination reflects how much variability of one factor contributes to a relationship to another related factor. The goodness of fit is represented as a value from 0.0 to 1.0. A value of 1.0 shows a perfect fit, meaning a highly reliable model, while a value of 0.0 would represent that the machine learning model is not adequate for implementation. The equation of the coefficient of determination shown as equation 3. SS_{RES} is the sum of squares of residuals, while SS_{ToT} is the total sum of square. $\bar{y}_i$ is the mean of y-axis and y_i is the observed value and $\hat{y}_i$ is the predicted data.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{ToT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \quad (3)$$

The mean squared error regression loss (MSE) is computed by the equation 4 where is the observed value and $\hat{y}_i$ is the predicted value.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (4)$$

Table 1 shows that Random forest regression exhibited higher value of the coefficient of determination more than the other models, showing a value of 0.950, followed by Decision tree regression showing a value of 0.947. The machine learning model most unsuitable for implementation is multiple linear regression, showing a value of 0.662. Using polynomial regression showed a significant improvement when increase in the degree. However, the polynomial regression at 6 degree showed the worst performance due to the overfitting of model, considering the significant difference of value of the coefficient of determination between training set and test set at 6 degree. It can be noticed that polynomial regression showed a perfect fit gradually in the case of increased degree but the degree should not be exceedingly increased since this leads to overfitting. In terms of coefficient of determination of training set, all machine learning models showed insignificant difference of coefficient of determination between test set and training set.

Table 2 shows the mean squared error regression loss (MSE) of the tested machine learning models. The main reason for this is in order to recheck in case t the dataset has an outlier. In the case of mean squared error, the perfect value needs to show a near-zero value. The decision tree showed the best number, yielding 0.167. The second best model was the random forest regression which showed an MSE value of 0.335. The worst model was multiple linear regression, showing a value of 2.845. The results of MSE in the case of polynomial regression are consistent in Table 2. The figure shows significant improvement with an increase in the degree. However, this research did not mention discuss the result of polynomial regression at 6 degree which showed a value of 3.2125 due to overfitting.

Table 1. *Results of coefficient of determination.*

Model	**Coefficient of determination of training set (R2)**	**Coefficient of determination of test set (R2)**
Multiple linear regression	0.6715	0.6628
Polynomial		
Polynomial regression (degree=2)	0.7389	0.7231
Polynomial regression (degree=3)	0.7881	0.7738
Polynomial regression (degree=4)	0.8944	0.8655
Polynomial regression (degree=5)	0.9338	0.8793
Polynomial regression (degree=6)	0.8579	-10.9668
Decision tree regression	1.0000	0.9470
Random forest regression	0.9968	0.9506

Table 2. *Mean of MSE of four machine learning models.*

Model	**Mean squared error regression loss (MSE)**
Multiple linear regression	2.8454
Polynomial	
Polynomial regression (degree=2)	2.5420
Polynomial regression (degree=3)	2.1790
Polynomial regression (degree=4)	1.5195
Polynomial regression (degree=5)	1.2359
Polynomial regression (degree=6)	3.2125
Decision tree regression	0.1672
Random forest regression	0.3353

Figures 10 to 17 show the overall trend of actual data and predicted data through testing machine learning models. The research established that multiple linear regression in figure 10 was unsuitable for data like a sine wave. Applying linear regression is suited for the data with characteristics similar to a straight line. Polynomial regression from figure 11 to figure 15 showed the worst performance in case of adjusting at a low degree while an increase in the degree could increase the performance of the model. However, it should be cautioned that increasing degree arbitrarily may result in model's overfitting.

In terms of both random forest regression and decision tree regression in figures 16 and 17, the finding of the study suggested that they are suitable for data with the characteristics of a sine wave. Both the two models yielded a very similar value of best figure throughout their prediction. Even though random forest showed the best figure of coefficient of determination, its MSE showed a relatively lower value than decision tree regression. Selecting decision tree regression to use was likely to suit the characteristics of the data in that way.

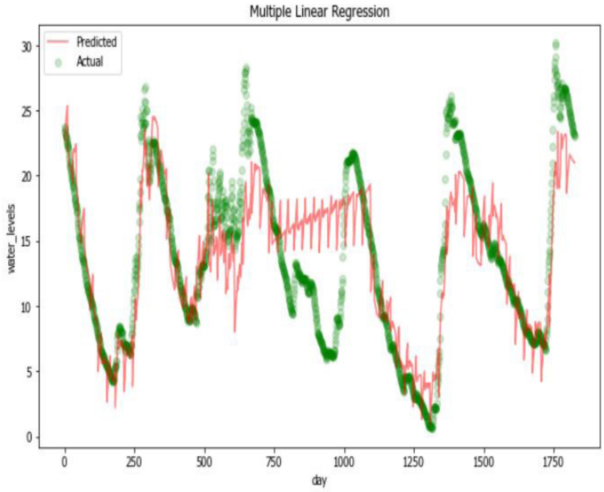


Figure 10. *Overall trend through multiple linear regression.*

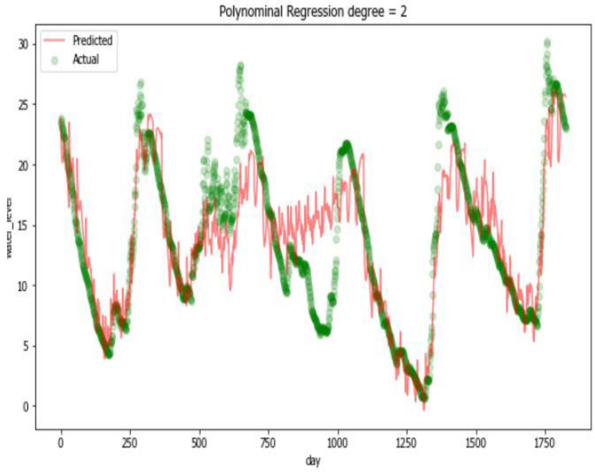


Figure 11. *Overall trend through polynomial regression at 2 degree.*

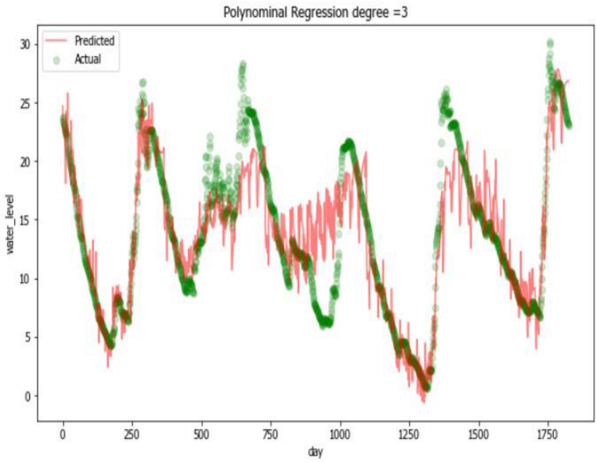


Figure 12. *Overall trend through polynomial regression at 3 degree.*

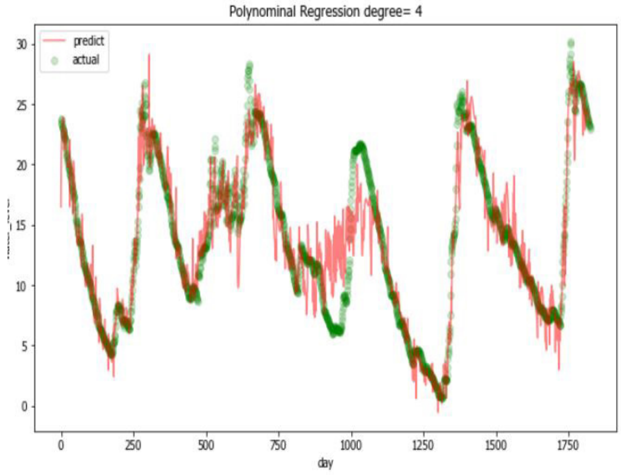


Figure 13. *Overall trend through polynomial regression at 4 degree.*

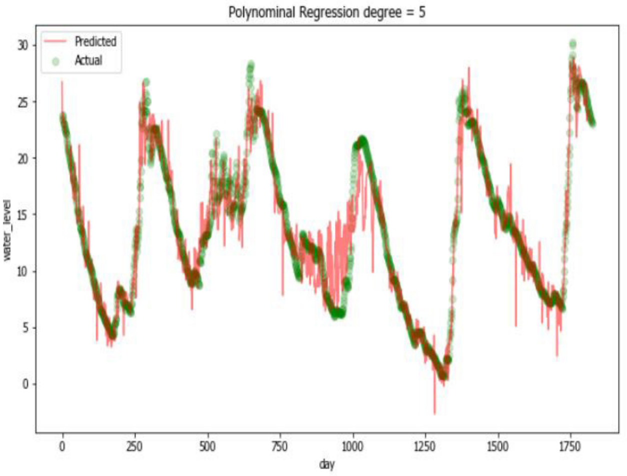


Figure 14. *Overall trend through polynomial regression at 5 degree.*

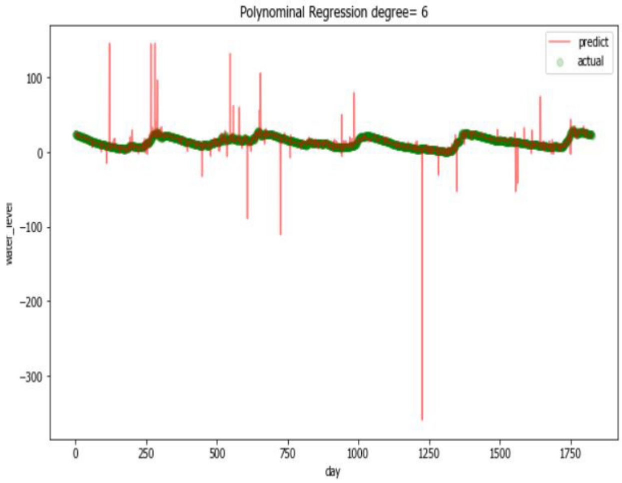


Figure 15. *Overall trend through polynomial regression at 6 degree.*

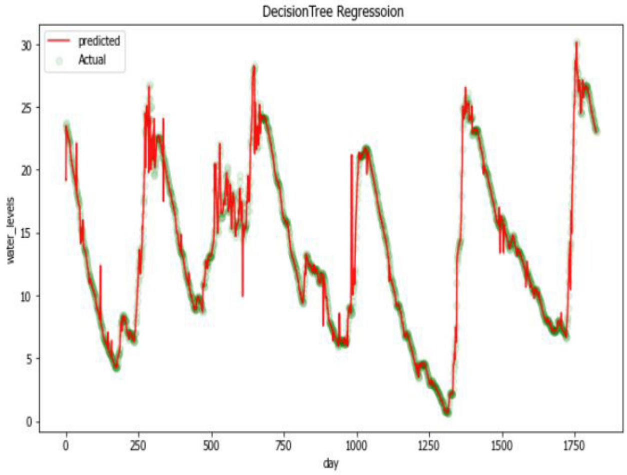


Figure 16. *Overall trend through decision tree regression.*

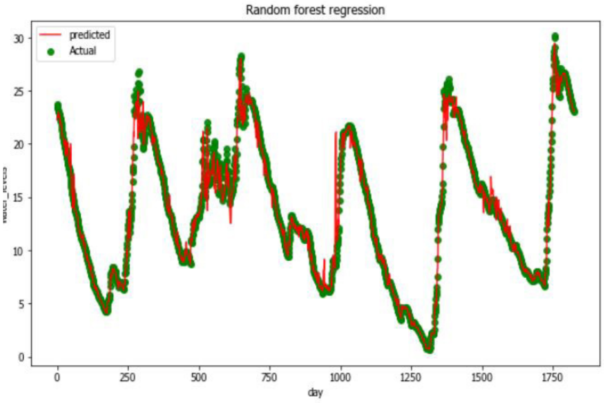


Figure 17. *Overall trend through random forest regression.*

4.3 Experimental discussion and future work

The results of this experiments were as expected. The multiple liner regression model was proved to be relatively inappropriate for the non-linear data which is consistent with Feng [16]. While polynomial showed a significant improvement with an increase in the degree, it should be cautioned that model overfitting may occur when the degree is arbitrarily increased. However, using polynomial model could not beat the decision tree regression and random forest regression. It was established that both models are suitable for data with characteristics relatively similar to sine wave, compared with multiple linear regression and polynomial regression, which obtained results according to [24], [25], [26]. In the part of background data of Huay Saneng, this needs to be investigated further, especially flows of water accumulating in Huay Saneng. This includes humidity, temperature, and other water sources that could contribute to the water level of Huay Saneng.

5. Conclusions

This research studied how to predict the water levels in order to alleviate drought problem. To achieve this goal, understanding the basic background of Huay Saneng was compulsory. The research findings showed that the evaporation of water and water level had strong correlation over the last five years while rainfall paid a small contribution to the water level. An implication of this is the possibility that Huay Saneng may have various ways of gathering its water from other sources whereby this requires a further investigation in the future. The decision tree was considered the best model to predict the drought, considering its coefficient of determination and the mean squared error regression loss (MSE) as compared amongst the four machine learning models. The multiple linear regression was proved that it could not apply when data was so complex and non-linear. In the future, it may be necessary to collect more data related to drought such as weather, water flowing from other sources, draining water for consumption and agriculture in order to improve the accuracy of prediction.

6. Acknowledgement

This research would not have been possible without regular support from Royal Irrigation Department operation on Huay Saneng area. Finally, I would like to thank Faculty of Agriculture and Technology, Rajamangala Isan, Surin Campus always supporting my research.

7. References

[1] B. C. Bates, Z. W. Kundzewicz, S. Wu, and J. P. Palutikof, “Climate Change and Water. Technical Paper of the Intergovernmental Panel on Climate Change, IPCC Secretariat: Geneva, Switzerland,” *The American Midland Naturalist*, Vol. 168, No. 1, 2008.

[2] A. K. Mishra and V. P. Singh, “A review of drought concepts,” *Journal of hydrology*, Vol. 391, No. 1–2, pp. 202–216, 2010.

[3] bangkokpost, *Drought-*s*truck Surin Hospital declares emergency*,. Available Online at. https://www.bangkokpost.com/thailand/general/1727555/drought-struck-surin-hospital-declares-emergency, accessed on 10 Mar 2021.

[4] B. Lantz, *Machine learning* with R. Packt publishing ltd, 2013.

[5] D. A. Sachindra and S. Kanae, “Machine learning for downscaling: the use of parallel multiple populations in genetic programming,” *Stochastic Environmental Research and Risk Assessment*, Vol. 33, No. 8, pp. 1497–1533, 2019.

[6] P. Sukka, *Forecasting of Daily Inflow to Large Reservoir in Upper Ping River Basin Using Artificial Neural Network*, Master of Engineering (Civil Engineering), Chiang Mai University, Chiang Mai.(in Thai), 2005.

[7] S. Amloy, *Reservoir Inflow Forecasting by Decision Tree Model: A Case Study of Huay Nam Sai Reservoir Nakhon Si Thammarat Province,* Master of Engineering (Water Resources Engineering), Kasetsart University, Bangkok.(in Thai), 2009.

[8] H. D. P. Weerasinghe, H. L. Premaratne, and D. U. J. Sonnadara, “Performance of neural networks in forecasting daily precipitation using multiple sources,” *Journal of the National Science Foundation of Sri Lanka*, Vol. 38, No. 3, 2010.

[9] N. Z. Che Ghani, Z. Abu Hasan, and L. Tze Liang, “Estimation of missing rainfall data using GEP: case study of raja river, Alor Setar, Kedah,” *Advances in Artificial Intelligence*, Vol. 2014, 2014.

[10] L. Mediero, L. Garrote, and A. Chavez-Jimenez, “Improving probabilistic flood forecasting through a data assimilation scheme based on genetic programming,” *Natural Hazards and Earth System Sciences,* Vol. 12, No. 12, pp. 3719–3732, 2012.

[11] G. Artigue, A. Johannet, V. Borrell, and S. Pistre, “Flash flood forecasting in poorly gauged basins using neural networks: case study of the Gardon de Mialet basin (southern France),” *Natural Hazards and Earth System Sciences*, Vol. 12, No. 11, pp. 3307–3324, 2012.

[12] N. Watanabe, K. Fukami, H. Imamura, K. Sonoda, and S. Yamane, “Flood forecasting technology with radar-derived rainfall data using genetic programming,” *2009 International Joint Conference on Neural Networks,* Atlanta, Georgia, USA,

pp. 3311–3318, 2009.

[13] R. Obiedat, M. Alkasassbeh, H. Faris, and O. Harfoushi, “Customer churn prediction using a hybrid genetic programming approach,” *Scientific Research and Essays*, Vol. 8, No. 27, pp. 1289–1295, 2013.

[14] H. Ebrahimi and T. Rajaee, “Simulation of groundwater level variations using wavelet combined with neural network, linear regression and support vector machine,” *Global and Planetary Change*, Vol. 148, pp. 181–191, 2017.

[15] U. S. Panu and T. C. Sharma, “Challenges in drought research: some perspectives and future directions,” *Hydrological Sciences Journal*, Vol. 47, No. S1, pp. S19–S30, 2002.

[16] Y. Feng and S. Wang, “A forecast for bicycle rental demand based on random forests and multiple linear regression,” 2*017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, Wuhan, China, pp. 101–105, 2017.

[17] S. Barua, A. W. M. Ng, and B. J. C. Perera, “Artificial neural network–based drought forecasting using a nonlinear aggregated drought index,” *Journal of Hydrologic Engineering,* Vol. 17, No. 12, pp. 1408–1413, 2012.

[18] S. Ghimire, R. C. Deo, N. J. Downs, and N. Raj, “Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia,” *Journal of cleaner production,* Vol. 216, pp. 288–310, 2019.

[19] T. Yang, X. Zhou, Z. Yu, V. Krysanova, and B. Wang, “Drought projection based on a hybrid drought index using Artificial Neural Networks,” *Hydrological Processes,* Vol. 29, No. 11, pp. 2635–2648, 2015.

[20] M. Mokhtarzad, F. Eskandari, N. J. Vanjani, and A. Arabasadi, “Drought forecasting by ANN, ANFIS, and SVM and comparison of the models,” *Environmental earth sciences,* Vol. 76, No. 21, pp. 1–10, 2017.

[21] P. C. Deka, “Support vector machine applications in the field of hydrology: a review,” *Applied soft computing*, Vol. 19, pp. 372–386, 2014.

[22] X. Wang, T. Liu, X. Zheng, H. Peng, J. Xin, and B. Zhang, “Short-term prediction of groundwater level using improved random forest regression with a combination of random features,” *Applied Water Science*, Vol. 8, No. 5, pp. 1–12, 2018.

[23] V. Rodriguez-Galiano, M. P. Mendes, M. J. Garcia-Soldado, M. Chica-Olmo, and L. Ribeiro, “Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain),” *Science of the Total Environment*, Vol. 476, pp. 189–206, 2014.

[24] M. Hamza and D. Larocque, “An empirical comparison of ensemble methods based on classification trees,” *Journal of Statistical Computation and Simulation*, Vol. 75, No. 8, pp. 629–643, 2005.

[25] H. Liao and W. Sun, “Forecasting and evaluating water quality of Chao Lake based on an improved decision tree method,” *Procedia Environmental Sciences*, Vol. 2, pp. 970–979, 2010.

[26] W. Moudani, “Dynamic features selection for heart disease classification,” *International Journal of Health and Medical Engineering,* Vol. 7, No. 2, pp. 105–110, 2013.

[27] D. PANAGOULIA, “Impacts of GISS-modelled climate changes on catchment hydrology,” *Hydrological Sciences Journal*, Vol. 37, No. 2, pp. 141–163, 1992.

[28] D. P. Lettenmaier and T. Y. Gan, “Hydrologic sensitivities of the Sacramento-San Joaquin River basin, California, to global warming,” *Water Resources Research*, Vol. 26, No. 1, pp. 69–86, 1990.

[29] M. B. Richman and L. M. Leslie, “Machine Learning for Attribution of Heat and Drought in Southwestern Australia,” *Procedia Computer Science*, Vol. 168,

pp. 3–10, 2020.

[30] H. Balti, A. Ben Abbes, N. Mellouli, I. R. Farah, Y. Sang, and M. Lamolle, "A review of drought monitoring with big data: Issues, methods, challenges and research directions," *Ecological Informatics*, Vol. 60, pp. 101136, 2020.

[31] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*: John Wiley & Sons, 2021.

[32] D. E. Kandzari, D. L. Bhatt, S. Brar, C. M. Devireddy, M. Esler, M. Fahy, J. M. Flack, B. T. Katzen, J. Lea, and D. P. Lee, "Predictors of blood pressure response in the SYMPLICITY HTN-3 trial," *European heart journal*, Vol. 36, No. 4, pp. 219–227, 2015.

[33] D. B. Lobell, J. I. Ortiz-Monasterio, G. P. Asner, R. L. Naylor, and W. P. Falcon, "Combining field surveys, remote sensing, and regression trees to understand yield variations in an irrigated wheat landscape," *Agronomy Journal*, Vol. 97, No. 1, pp. 241–249, 2005.

[34] J. A. Vitale, V. Povìa, N. D. Vitale, T. Bassani, G. Lombardi, L. Giacomelli, G. Banfi, and A. La Torre, "The effect of two different speed endurance training protocols on a multiple shuttle run performance in young elite male soccer players," *Research in Sports Medicine*, Vol. 26, No. 4, pp. 436–449, 2018.

[35] A. Gelman, J. Hill, and A. Vehtari, *Regression and other stories.* Cambridge University Press, 2020.

[36] M. R. Balf, R. Noori, R. Berndtsson, A. Ghaemi, and B. Ghiasi, "Evolutionary polynomial regression approach to predict longitudinal dispersion coefficient in rivers," *Journal of Water Supply: Research and Technology-Aqua*, Vol. 67, No. 5, pp. 447–457, 2018.

[37] H. Huang, Z. Wang, F. Xia, X. Shang, Y. Liu, M. Zhang, R. A. Dahlgren, and K. Mei, "Water quality trend and change-point analyses using integration of locally weighted polynomial regression and segmented regression," *Environmental Science and Pollution Research*, Vol. 24, No. 18, pp. 15827–15837, 2017.

[38] M. Gentilucci, C. Bisci, P. Burt, M. Fazzini, and C. Vaccaro, "Interpolation of rainfall through polynomial regression in the Marche region (Central Italy)," *The Annual International Conference on Geographic Information Science*, Lund, Sweden, pp. 55–73, 2018.

[39] X. Jing, "An integrated prediction model for water supply-demand ability," *2016 4th International Conference on Advanced Materials and Information Technology Processing (AMITP 2016)*, Guilin, China, pp. 528–531, 2016.

[40] J. R. Quinlan, C4. 5: *programs for machine learning*. Elsevier, 2014.

[41] S. Ragettli, J. Zhou, H. Wang, C. Liu, and L. Guo, "Modeling flash floods in ungauged mountain catchments of China: A decision tree learning approach for parameter regionalization," *Journal of Hydrology,* Vol. 555, pp. 330–346, 2017.

[42] E. Sutoyo and A. Musnansyah, "A Hybrid of Seasonal Autoregressive Integrated Moving Average (SARIMA) and Decision Tree for Drought Forecasting," *Proceedings of the International Conference on Engineering and Information Technology for Sustainable Industry,* Tangerang, Indonesia, pp. 1–6, 2020.

[43] L. Breiman, "Bagging predictors," *Machine learning*, Vol. 24, No. 2, pp. 123–140, 1996.

[44] J. W. Osborne, *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Sage, 2013.

[45] S. Raschka, *Python machine learning:* Packt publishing ltd, 2015.

[46] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* O'Reilly Media, 2019.