International Journal of Environment, Agriculture and Biotechnology
Vol-9, Issue-4; Jul-Aug, 2024
Peer-Reviewed International Journal
Journal Home Page Available: https://ijeab.com/
Journal DOI: 10.22161/ijeab

Forecasting of Dam Lake Water Level Using M5 Decision Tree and Anfis Models

Özden Nur Şentürk*, Fatih Üneş, Mustafa Demirci, Bestami Taşar

Department of Civil Engineering, Iskenderun Technical University, Türkiye
*Email: ozdennursenturk@gmail.com

Received: 25 Jun 2024; Received in revised form: 20 Jul 2024; Accepted: 29 Jul 2024; Available online: 07 Aug 2024

Abstract—*Dam reservoir level prediction is important for dam construction, operation, design and safety. In this study, dam reservoir level change predictions were investigated using the M5 Decision Tree (M5 Tree) and Adaptive Neural Fuzzy Inference System (ANFIS) models. For modeling the daily dam reservoir water level (t), the lagged time of reservoir water level (t-1), stream flow (t) and precipitation heights in the dam basin (t) were used. The model results were compared with the results of conventional multiple linear regression (MLR) models. The models were analyzed with graphical and statistical results. The coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE) performance criteria were taken into account when comparing the prediction models. The results showed that M5 Tree and Anfis model results gave a better performance in predicting the dam reservoir level change.*

Keywords—***Dam Reservoir Level, Fuzzy, Modelling, Prediction, Regression.***

I. INTRODUCTION

The contents of each section may be provided to understand easily about the paper. Reservoirs and dams are essential to the management of water resources. In addition to providing water to cities, they are also employed in the production of hydroelectric power, flood control, and agricultural irrigation. A multipurpose water storage facility must have its reservoir or dam level regularly monitored in order to make the necessary modifications on time and to ensure maximum performance. In the field of water supply management, one of the most difficult jobs for planners and operators is forecasting water levels.

Control of water volume in the dam reservoir is achieved by accumulating and distributing water at the right time. Due to the precautions not taken in time and water-related problems, there may be loss of life and property. Therefore, proper dam reservoir management is a necessity not only in terms of freshwater supply but also in terms of preventing possible damages. One of the basic conditions for the most effective management of dam reservoirs is to determine the dam reservoir water volume and to be able to predict the ups and downs in this volume.

The first studies to determine the dam reservoir capacity were made by Ripple [1] and Sudler [2]. Since those studies, many researchers have used classical and traditional methods in dam reservoir studies. Sudheer and Jain [3] tried to explain the internal behavior of artificial neural networks with river flow models. Sudheer [4] tried to create river models with information extracted from trained neural networks. Üneş [5] and Unes et al [6] tried to determine the dam reservoir level changes with artificial intelligence techniques. In these methods, the reservoir volume is defined as the conservation of mass (continuity equation) at the macro scale in hydraulic research systems. In past studies on the water level and volume in lakes, the stability of the annual level of water was generally used by considering the mass-volume methods and statistical methods.

An earlier study used artificial neural networks (ANN) in conjunction with tree-based models, including decision trees (M5T), random forests (RF), and gradient-boosted trees (GB), to predict the dam intake into the Soyang River Dam in South Korea [7]. Research showed that an ensemble method, which merges the RF/GB forecasts with a multilayer perceptron (MLP), might outperform the use of

a single individual model. The Upo wetland in South Korea serves as another example of the predictive power of tree-based approaches. In comparison to ANNs, DTs, and support vector machines (SVM), RF was found to have the best forecast accuracy [8].

To estimate the water level of Lake Erie, other techniques such as the Gaussian process (GP), multiple linear regression (MLR), and k-nearest neighbor (KNN) have also been compared to tree-based and ANN models [9]. Their findings demonstrate how machine learning techniques, particularly the MLR and M5P model tree, outperformed the process-based advanced hydrologic prediction system (AHPS) in terms of accuracy and training speed [10].

In this study, forecasting models were developed for the dam lake water level. In the forecasting models, stream flow, precipitation amount falling in the basin and shifted lake water level were used as independent variables. M5 Decision Tree (M5 Tree), which is one of the machine learning techniques that show superior performance in nonlinear problems, and Adaptive Neuro Fuzzy Logic (ANFIS) models, which is a hybrid method working with Fuzzy logic algorithm, were used.

II. MATERIAL AND METHODS

Study Area

The study location is Lake Tuscaloosa, which is located close to Tuscaloosa, Alabama, USA. (Figure 1) By damming the North River, a reservoir known as Lake Tuscaloosa was formed in west-central Alabama. Thornton Jones built it to supply water to Tuscaloosa citizens as well as for industrial purposes. At a cost of around $7,725,000, it was finished in 1970. The lake is a popular spot for outdoor enjoyment because it's close to Northport and Tuscaloosa. When Tuscaloosa's population grew and its two existing reservoirs, Harris Lake and Lake Nicol, could no longer hold enough water, the city built Lake Tuscaloosa. By building a dam on the North River, the region that would eventually become Lake Tuscaloosa was flooded.

Fig.1: Study area

The data used in this study were obtained by the United States Geological Survey (USGS). Streamflow (Q, m3/s), precipitation height in the basin (P, cm) and Lake Water Level (LWL, m) variables were used in the estimation models. The daily change in the LWL variable of the Tuscaloosa reservoir between 2018-2021 is given in Figure 2.

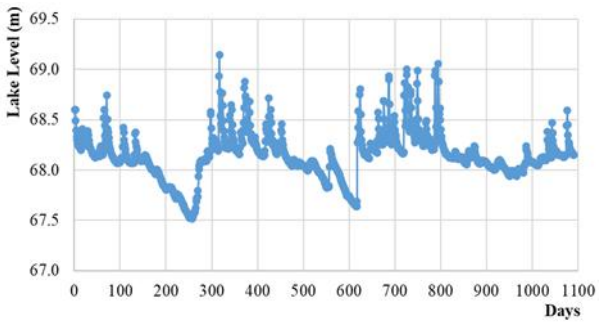

Fig.2: Daily lake water level change

Methods

Multi Linear Regression (MLR)

Multiple Linear Regression analyses are among the methods used to model the relationship between two or more variables according to the cause-effect relationship. If a single independent variable is used as an input in the model established to estimate the dependent variable, it is called single regression, and if more than one independent variable is used, it is called multiple regression analysis. In the MLR method, the effect of independent variables on dependent variables is expressed with the regression coefficient in the equation. This coefficient shows the degree of effect of independent variables on the dependent variable in the regression equation. Multiple Linear Regression is given in Equation 1

$$Y_i = (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n) + \varepsilon_i \qquad (1)$$

This equation contains linear expressions. In this equation, Xi (i = 1, …, n) independent variables, Yi dependent variable, β regression coefficient and ε represents the error.

M5 Decision Tree (M5 Tree)

M5 Tree was first proposed by Quinlan [11] This method results in the estimated value of the dependent variable in a fast, practical and understandable way. It is a versatile logical model. It is a guide on how to deal with numerical data and missing data values. It is quite fast and produces understandable outputs that are very accurate at very high rates. This situation is explained by the robust and versatile operation of decision tree learning that can cope with the demands of real-world data sets. (Witten et al. [12]). The M5T algorithm creates a regression series by repeatedly dividing the sample space using tests on a single feature that maximizes the variance in the target space. The mathematical equation for calculating the standard deviation reduction (SDR) is given in Equation 2

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} sd(T_i) \quad (2)$$

Adaptive Neuro Fuzzy Inference System (ANFIS)

An adaptive network-based fuzzy inference system (ANFIS) is used as an artificial neural network method based on a fuzzy inference system. ANFIS model was developed by Jang since the early 1990s and is used in modeling nonlinear functions and estimating chaotic time series [13-14]. ANFIS consists of nodes directly connected and each node represents a processing unit [15]. Since ANFIS uses both artificial neural networks and fuzzy logic inference methods, it uses a hybrid learning algorithm [16]. There are two approaches to fuzzy inference systems. These approaches are the approach of Mamdani and Assilian, Takagi and Sugeno [17]. To apply an adaptive neuro-fuzzy inference system (ANFIS), data sets with input and output are generally needed. The ANFIS method finds the best values for the membership functions of fuzzy sets by training the model with the principle of reducing errors. It also creates fuzzy rules for FIS. The structure of the Adaptive Neural Inference System (ANFIS) is shown in Figure 2. Here; "x, y, z, t" are our independent variables, "a1, a2, b1, b2, c1, c2, d1, d2" are the input parameters, "∏ (pi)" are the membership functions, "N" are the rules and "wi" are the weights of the parameters.

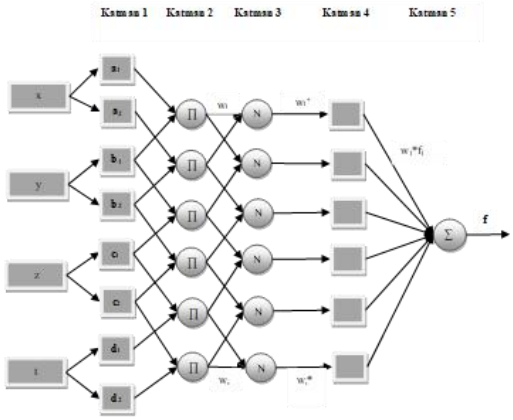

Fig.4: ANFIS model with four inputs and one output.

In Figure 4; in the 1st layer, the membership function is selected, and the membership levels of the linguistic variables are determined. In the ANFIS model of this study, the number of membership functions is two for each independent variable. In the 2nd layer, all nodes in the second layer are fixed nodes indicated by the symbol "∏". The products of the outputs of the first layer represent the resulting fuzzy rules. In the 3rd layer, here too, the nodes in the layer are fixed nodes and indicated by the symbol "N". ANFIS normalizes the values in the network structure. These values are taken as output. In the 4th layer, all nodes in this layer are normalized nodes and the weight values (w) coming from the third layer are multiplied by the first-degree polynomial equation. "w1*f1" is the layer output. In the 5th layer, there is only one fixed node in this layer. It gives the total result of all the operations coming as "Σ".

III. RESULTS

In the model analysis, the first 75% of the total data set (1091) was used as training data and the last 25% (273) as test data. For the 273-day test data; MLR, M5 Tree and Anfis model's performances were evaluated using statistical criteria (RMSE, MAE and R^2). For each model, mean absolute error (MAE), root mean square error (RMSE), and coefficients of determination (R^2) between model predictions and measured values were used and the statistical criteria used are given in the equations below. Table 1 shows model performance comparisons as a result of the analysis.

$$RMSE = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N} LWL_{measurement} - LWL_{prediction}\right)^2} \quad (3)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|LWL_{measurement} - LWL_{prediction}\right| \quad (4)$$

Table 1. Error information and correlation changes of the models.

Model	Model Inputs	MAE (m)	RMSE (m)	R^2
MLR	Q(t), P(t) ve GSS(t-1)	0.024	0.030	0.957
M5 Tree	Q(t), P(t) ve GSS(t-1)	0.010	0.018	0.965
Anfis	Q(t), P(t) ve GSS(t-1)	0.009	0.016	0.971

MLR Results

In the MLR model, stream flow rate (Q(t), m^3/s), precipitation height in the basin (P(t), cm) and offset lake water level (LWL(t-1), m) parameters were used in LWL estimation. The results of the test phase of the MLR method are given as distribution and scatter graphs in Figures 5-6, respectively.

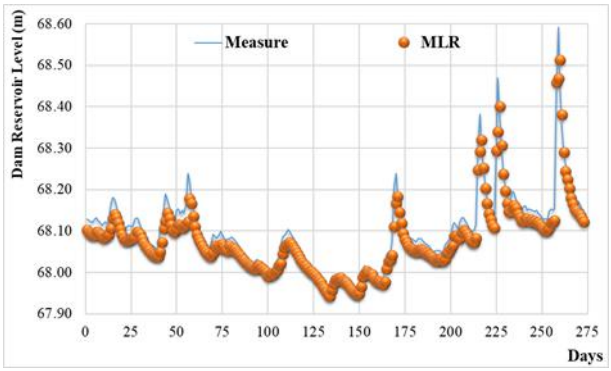

Fig.5: Scatter graph of MLR model

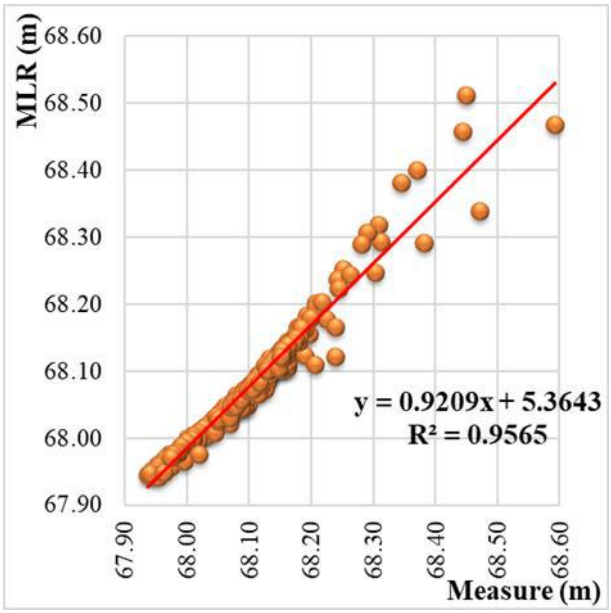

Fig.6: Distribution graph of MLR model

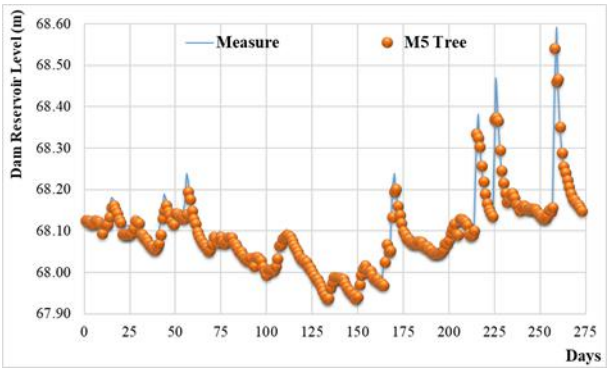

Fig.7: Scatter graph of M5 Tree model

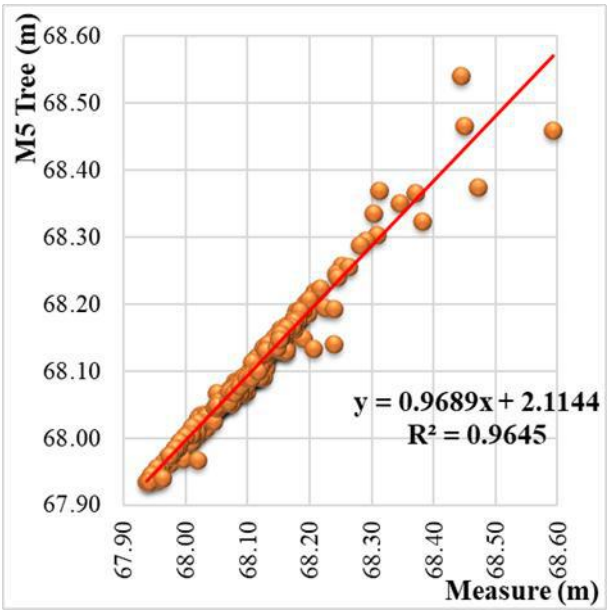

Fig.8: Distribution graph of MLR model

According to the scatter graph (Figure 6) and Table 1, it was seen that the coefficient of determination obtained was R^2 = 0.957. When the MLR model in the test phase was examined, it was determined that it had the lowest determination value. It was determined that some peak LWL amounts gave lower estimates than the actual LWL values in the MLR model. Therefore, it is seen that there is a decrease in the determination values.

M5 Tree Results

In the M5 Tree model (as in the MLR method), LWL was estimated using the stream flow rate (Q(t), m^3/s), precipitation height in the basin (P(t), cm) and offset Lake Water Level (LWL(t-1), m) parameters. The distribution and scatter graphs in the test phase results of the M5 Tree model are given in Figures 7-8, respectively.

According to the distribution and scatter graphs given in Figures 7 and 8, it was obtained that there was a good agreement between the real LWL and M5 Tree estimation results. It was seen from Table 1 and Figure 8 that the coefficient of determination R^2 = 0.965. The M5 Tree method performed better than the MLR method in LWL estimation.

Anfis Results

In the Anfis model (as in the MLR and M5 Tree models), the LWL was estimated using the stream flow rate (Q(t), m3/s), precipitation height in the basin (P(t), cm) and offset Lake Water Level (LWL(t-1), m) parameters. The distribution and scatter graphs in the test phase results of the Anfis model are given in Figures 9-10, respectively

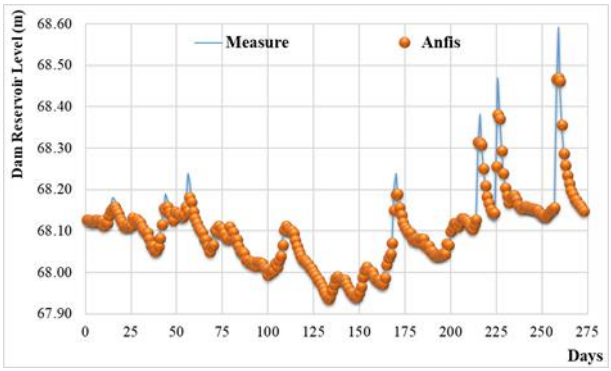

Fig.9: Scatter graph of Anfis model

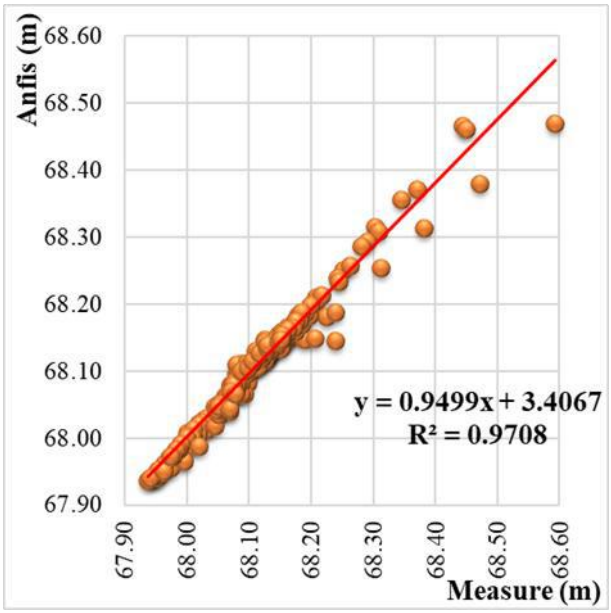

Fig.10: Distribution graph of Anfis model

According to the distribution graph, it was obtained that there was a harmony between the real results and the Anfis estimation results. When Figure 10 and Table 1 were examined, it was seen that the determination coefficient obtained was R^2 = 0.971. It was determined that the Anfis model generally gave estimates closer to the LWL peak values. Therefore, it was seen that there was an increase in the determination values compared to other methods. It was determined that the Anfis results had the best estimation performance for LWL estimations. The results of the Anfis estimation values of the real-time LWL showed better performance than the other model estimates and good estimation results were observed according to the real values. When we look at the MAE, RMSE and R^2 shown in Table 1, the Anfis (0.009; 0.016; 0.971) model showed the best performance compared to the other models.

IV. CONCLUSION

For the purpose of designing and building lakeshore constructions, other industrial operations, and integrated water resources management, it is critical to predict fluctuations in dam reservoir levels. The current study used MLR, M5 Tree, and Anfis models to anticipate dam reservoir Tuscaloosa lake level in the United States. For the performance evaluation of multiple linear regression (MLR), M5 decision tree (M5 Tree) and adaptive network-based fuzzy inference system (ANFIS) models, coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE) were calculated. From the study, the following conclusions can be made.

As a result of the created models' performance evaluation, all models successfully estimated the reservoir lake level. The results of the MLR method and the M5Tree method showed similar results.

It was seen that the adaptive network-based fuzzy inference system (ANFIS) model was more successful than the other three models due to its lower error values and high coefficient of determination. Compared to the traditional models, the proposed Anfis model yields more accurate estimations of the fluctuations in the reservoir level.

ACKNOWLEDGEMENTS

The data used in this study were downloaded from the web server of the USGS. The authors wish to thank the staff of the USGS who are associated with data observation, processing, and management of USGS Web sites.

REFERENCES

[1] Ripple, W. (1883). The Capacity of Storage for water supply, Proc., Institution of Civil Engineers, 71, 270.

[2] Sudler, C. E., (1927), Storage Required for Regulation of Streamflow, Trans., ASCE, 91, No. 622.

[3] Sudheer, K.P., Jain A., (2004). Explaining the internal behaviour of artificial neural network river flow models. Hydrological Processes 18 (4): 833-844.

[4] Sudheer K.P., (2005). Knowledge extraction from trained neural network river flow models, Journal of Hydrologic Engineering 10 (4): 264-269.

[5] Üneş, F.(2010a), Dam reservoir level modelıng by neural network approach. A case study, Neural Network World 4, 461- 474.

[6] Üneş, F., Demirci, M., Kişi, Ö., (2015a), Prediction of millers ferry dam reservoir level in USA using artificial neural network, Periodica Polytechnica Civil Engineering 59, 309–318.

[7] Hong, J.; Lee, S.; Bae, J.H.; Lee, J.; Park, W.J.; Lee, D.; Kim, J.; Lim, K.J. Development and Evaluation of the Combined Machine Learning Models for the Prediction of Dam Inflow. Water 2020, 12, 2927. [CrossRef].

[8] Choi, C.; Kim, J.; Han, H.; Han, D.; Kim, H.S. Development of Water Level Prediction Models Using Machine Learning in. Wetlands: A Case Study of Upo Wetland in South Korea. Water 2020, 12, 93. [CrossRef].
[9] Wang, Q.; Wang, S. Machine Learning-Based Water Level Prediction in Lake Erie. Water 2020, 12, 2654. [CrossRef].
[10] Gronewold, A.D.; Clites, A.H.; Hunter, T.S.; Stow, C.A. An Appraisal of the Great Lakes Advanced Hydrologic Prediction System. J. Great Lakes Res. 2011, 37, 577–583. [CrossRef].
[11] Quinlan, J. R. (1992, November). Learning with continuous classes. In 5th Australian joint conference on artificial intelligence (Vol. 92, pp. 343-348).
[12] Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & Data, M. (2005, June). Practical machine learning tools and techniques. In Data mining (Vol. 2, No. 4, pp. 403-413). Amsterdam, The Netherlands: Elsevier.
[13] Jang JSR. Fuzzy modeling using generalized neural networks and kalman filter algorithm. Proceedings of the 9th National Conference on Artificial Intelligence, Anaheim, CA, USA, 1991 July 14–19. p.762–767.
[14] Jang JSR. ANFIS: Adaptive-Network-Based Fuzzy Inference System. IEEE Transactions on Systems, Man and Cybernetics 1993;23(3):665-685.
[15] Cansız ÖF, Erginer İ, ve Genç GG. Ulaştırma sektöründe karayollarının payına düşen enerji tüketiminin yapay sinir ağları ve çok değişkenli lineer regresyon yöntemleri ile tahmini. International Eurasian Conference on Science, Engineering and Technology 2018:627-633; Ankara.
[16] Üneş F, Kaya YZ, Mamak M. Daily reference evapotranspiration prediction based on climatic conditions applying different data mining techniques and empirical equations. Theoretical and Applied Climatology 2020;141(1-2):763-773.
[17] Demirci M, Unes F, Kaya YZ, Tasar B, Varçin H. Modeling of dam reservoir volume using adaptive neuro-fuzzy method, in: 10th international Water and Air Components Conference, 2018 March 145–152; Sovata, ROMANIA.