Enhancing Monthly Lake Levels Forecasting Using Heuristic Regression Techniques with Periodicity Data Component: Application of Lake Michigan

Vahdettin DEMIR (✉ vahdettin.demir@karatay.edu.tr)

KTO Karatay Üniversitesi <https://orcid.org/0000-0002-6590-5658>

Research Article

Keywords: Forecast, Heuristic Regression Techniques, MARS, M5-Tree, LSSVR, Lake Michigan

Posted Date: August 17th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-726003/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Theoretical and Applied Climatology on February 21st, 2022. See the published version at <https://doi.org/10.1007/s00704-022-03982-0>.

Abstract

This paper investigates the accuracy of three different techniques with periodicity component for estimation of monthly lake levels. The compared methods are Least Square Support Vector Regression (LSSVR) Multivariate Adaptive Regression Splines (MARS) and M5 Model Tree (M5-Tree). Data from Lake Michigan, located in the USA, is used in the analysis. In the first stage of the study, three different techniques were applied to forecast monthly lake-levels variations up to 8-month ahead of time intervals. In the second stage, the influence of the periodicity component was applied (month number of the year, e.g., 1, 2, 3, ...12) as an external sub-set in modeling monthly lake levels. The Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2) were utilized are used for evaluating the accuracy of models. In both stages, the comparison results indicate that the MARS model generally performs superior to the LSSVR, and M5-Tree models. Furthermore, it has been discovered that including periodicity as an input to the models improves their accuracy in projecting monthly lake levels.

1 Introduction

Lake level fluctuations are significant for lakeshore structure planning, design, building, and operation, as well as for the management of fresh water lakes for water supply purposes. In order to regulate future lake level changes, models for modeling of high or abnormal level variations must be developed. The level measurements, or their future equally likely reproductions acquired by a simulation model, are a straightforward manner of getting lake management decision variables. Although comprehensive models incorporating hydrological and hydrometeorological variables such as precipitation, runoff, temperature, and evaporation can be found, it is more economically advantageous to use a model that simulates level variations based on past level records [1].

Lakes are used for a variety of domestic, industrial, and agricultural purposes [2, 3]. Forecasting lake water levels is important for water resource planning and management, lake navigation, tidal irrigation, and agricultural drainage canal management, among other things. The level of water in lakes is a complicated phenomenon that is primarily influenced by natural water exchange between the lake and its watershed, and consequently reflects hydrological changes in the watershed [4, 5]. For many practical applications, a model that predicts water-level changes based on previously measured levels is required [4].

Over the last few decades, hundreds of scholars have been interested in lake water level models. This is because global climate change has had an impact on the hydrological cycle, causing many lakes to dry up or flood unexpectedly. To model lake level fluctuations, several techniques have been devised. Sen et al. (2000) periodic and stochastic process [1]. Altunkaynak et al. (2003) used the diagram model and Markov process [6]. Altunkaynak (2007) used the artificial neural network [5]. Altunkaynak and Sen (2007) used the fuzzy logic [7]. Kisi (2009) used the wavelet conjunction model [8]. Karimi et al. (2012) used the gene expression programming and adaptive neuro-fuzzy inference system [4]. Sanikhani et al. (2015) used the adaptive-neuro-fuzzy inference system (ANFIS) and gene expression programming [9].

Young et al. (2015) used the Time Series Forecasting Model [10]. Shiri et al. (2016) used the extreme learning machine approach [3]. Shafaei and Kisi (2016), used the wavelet- Support Vector Regression (SVR), wavelet-ANFIS and Wavelet-ARMA conjunction models [11]. Liang et al. (2018) used the deep learning method [12]. Peprah and Larbi (2021) used the Integrated Moving Average and Kalman Filtering Techniques [13]. Luo et al. (2021) used the machine learning methods [14].

Most recently, three data-driven techniques, such as Least Square Support Vector Regression (LSSVR), Multivariate Adaptive Regression Splines (MARS), and M5 Model Tree, have achieved a remarkable emerging and promise in addressing difficult nonlinear situations. The methodologies mentioned above have been widely employed to solve hydrologic challenges [15–18]. LSSVR is a modified variant of support vector regression (SVR) that can solve problems involving quadratic programming [19]. It also avoids a number of flaws that other data-driven learning systems have (e.g., local minima, time consumption and over-fitting) [20]. In the field of engineering, LSSVR has had a successful application; for example, prediction of wastewater effluent parameters [21]. 2009), the expense of designing the structural components of a wing-box for an airplane [22], design of a superconducting magnetic energy storage controller with adaptive dampening [23], forecast of CO₂ in reservoir [24], economic analysis of oil recovery [25], forecast reservoir oil viscosity [26], In the hydrological study, there are a few studies have been conducted using LSSVR; for example, streamflow forecasting and estimation [15, 18, 27], estimation of daily water demand and daily inflow of dam [28], sediment transport modeling [29], modeling daily reference evapotranspiration [30], modeling of reservoir inflows [31], prediction of water pollution [32], forecast of air pollutants [17].

Multivariate adaptive regression splines are a newer artificial intelligence technique [33]. The ability to capture the natural difficulty of data mapping in high-dimensional data patterns, a rapid and adaptable model, and accurate forecasting of continuous and binary output variables are the key advantages of this method. Furthermore, this nonparametric statistical method provides a versatile procedure for organizing the relationship between input and output variables with fewer variable interactions [34]. Rainfall and temperature forecasting, streamflow forecasting, sediment concentration estimate, water pollution prediction, air pollutants prediction, freshwater distribution system modeling, and drought events river flow simulation are some of the previous studies using the MARS method in water resources applications [15, 18, 32, 34–38].

M5 model tree is a data mining methodology that uses the divide-and-conquer method to split the data time series into subspaces, allowing the multi-dimensional parameter space to be divided and the model to be generated automatically based on the overall quality requirement [39]. Scholars recently investigated the M5 model tree's utility in many hydrological applications, such as water level optimization [40], precipitation and river flow modeling [41], streamflow modeling [18], air pollutants modeling [17], evapotranspiration modeling [42], pan-evaporation modeling [37], flood events [43], and sediment yield modeling [44].

In this study, the major goals of the current research are (i) investigate three different novel heuristic regression techniques (M5 model-Tree, LSSVR and MARS) for modeling water levels forecasting, (ii) investigate influence of the periodicity component for water levels forecasting, (iii) in order to demonstrate the effectiveness, Lake Michigan in the USA have been used to perform the proposed models.

2. Case Study And And Data Preparation

The name Lake Michigan comes from the Ojibwa term Michi Gami, which means "large lake". Lake Michigan; In the United States (coordinates: 44°N 87°W), it is the third largest lake in the Lake District, consisting of five interconnected large lakes, and the sixth largest freshwater lake in the world (see Fig. 1). With a surface area of 58,016 square kilometers, drainage area 118,095 square kilometers, a width of 48–193 kilometers, a length of 494 kilometers, and a deepest point of 281 meters, the lake is the only lake in the middle northeast of the USA, among the Great Lakes, which remains entirely within the country's territory [45]. Lake Michigan is bordered by Wisconsin to the west, Illinois and Indiana to the south, and Michigan to the east. The surface of the lake, whose waters are fresh, is 177 meters above sea level. It is connected by the Strait of Mackinac to Lakes Superior, Huron, Erie, and Ontario from its northeast corner [46].

Forecasting lake level changes is critical for many operations in this region, including flood mitigation, reservoir management, drinking water distribution, water infrastructure management, trade, transportation, and beach erosion, among others. The observed data are 103 years (1236 months) long with an observation period between 1918 and 2020 for Lake Michigan station (IGLD 1985: Brochure on the International Great Lakes Datum 1985). The observed data were acquired from the report of the U.S. Army Corps of Engineers website "<https://www.lre.usace.army.mil/Missions/Great-Lakes-Information/Great-Lakes-Information-2/Water-Level-Data/>". The statistical parameters of the data used during the study period are shown in Table 1. The observed lake level fluctuation data for Lake Michigan, as well as the training and test datasets, are shown in Fig. 2.

Table 1
The monthly statistical informations of data sets for Lake Michigan

Data set	Period	X_{mean} (m)	X_{min} (m)	X_{max} (m)	Cs (m)	Sx (m)	Basik
All data	1918–2020	176.44	175.57	177.5	0.119	0.409	-0.763
Training	1918–1999	176.48	175.58	177.5	-0.075	0.389	-0.644
Test	1999–2020	176.28	175.57	177.46	0.949	0.445	-0.018

The partial auto-correlation and auto-correlation functions of the lake levels for Lake Michigan are also shown in Fig. 3. The figure shows that the lake level in Lake Michigan is highly correlated with past

month levels. The partial autocorrelation function indicates significant correlation up to lag 8 for the Lake Michigan and then stays within the confidence interval.

3 Methods

3.1. M5-Tree

M5 model tree algorithm is a new regression method developed by Quinlan in 1992. [39]. The two-component decision tree is the backbone of this model. The method defines the relationship between the independent and dependent variables with the linear regression function applied to the final leaf nodes. M5 model tree is better than other decision tree models used for categorical data [47].

The M5 model tree is a two-stage model. In the first step, the data is split into subsets to produce the decision schema (tree). The standard deviation of the class value reached at a node is used to categorize. The expected reduction is calculated based on the error that occurs as a result of testing the elements acting on this node. [42, 43]. The formulation of the standard deviation reduction (SDR) is as follows.

$$SDR = sd(T) - \sum \frac{|Ti|}{|T|} sd(Ti) \quad (1)$$

In this formula, sd is expressed as standard deviation. T represents a set of instances acting on the node. Subset samples with "i" results of potential data are represented by Ti [39].

3.2. MARS

Friedman proposed the MARS model, which is a nonparametric regression model [33]. MARS is a model used to predict non-linear continuous numerical results. It explains the complex nonlinear relationship between the model, estimation method and dependent variables. The MARS algorithm consists of two steps, forward and backward. It selects a set of suitable input variables with the forward step algorithm [48]. With the backward step algorithm, it eliminates unnecessary variables in the pre-selected set. A function is plotted from variable X to the new variable Y by two base functions or both variable values defined at the deviation point across the input range using the following fundamental equations. [49].

$$Y = \max(0, X - c) \quad (2)$$

$$Y = \max(0, c - x) \quad (3)$$

Here c represents the threshold (lower limit) value. MARS model is used especially in financial affairs management system, time series data and in many fields [15–18, 50, 51].

3.3. LSSVR

LSSVR is an extended version of the SVR model by Suykens and Vandewalle in 1999 [19]. It is used to estimate the current water levels statistically with the water levels in the historical time series and to

obtain the optimum function between the X input and Y output [18]. It performs this operation with a nonlinear relationship function with a multidimensional feature space. The regression function can be formulated as follows.

$$y(x) = w^1 \varphi(x) + b \quad (4)$$

Here, y is the value obtained in x, w is the coefficient vector, φ is the mapping function, b is the bias term obtained by minimizing the upper bound of the generalization error [19].

4. Application And Results

The three heuristic regression techniques evaluated (MARS, M5 Tree, and LSSVR) were created using MATLAB subroutines to estimate the compressive strength of foamed concrete. The data was divided into training and testing phase before you start modeling. The training data set accounts for 80% of all data (1236x0.8 = 989), while the testing data set accounts for 20% (247). Quantitative indicators are commonly used to evaluate hydrological applications. In their study, Legates and McCabe (1999) advised that predictive models in the field of hydrology be tested using goodness-of-fit for example Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and coefficient of determination (R^2) as shown in Eqs. (5–7) [52].

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (L_e - L_o)^2} \quad (5)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |L_e - L_o| \quad (6)$$

$$R^2 = \frac{\left[\sum_{i=1}^N (L_e - \bar{L}_e) (L_o - \bar{L}_o) \right]^2}{\sum_{i=1}^N (L_e - \bar{L}_e)^2 \sum_{i=1}^N (L_o - \bar{L}_o)^2} \quad (7)$$

In the Equations 5 to 7, L_e and L_o indicate the estimated and observed water levels values, and N indicates the raw water level amount of data. The aim of the current study is forecasting of lake level fluctuations by MARS, M5-Tree, and LSSVR. In this context, different input combinations were tried to predict the water levels. The inputs including the previous monthly lake levels (t-1, t-2, t-3, t-4, t-5, t-6, t-7, t-8) and the output corresponds to the lake level at time t.

The results of heuristic regression techniques in term of RMSE, MAE and R^2 are given in Table 2 with input combinations. RMSE of less than or equal to 60 cm indicates a very good and appropriate estimate [9, 53]. For Lake Michigan, an RMSE of less than or equal to 7.4 cm is a very good and satisfactory estimate.

Table 2
The results of heuristic regression techniques

Model	Inputs	Training			Test		
		RMSE	MAE	R ²	RMSE	MAE	R ²
MARS	I: t-1	0.068	0.055	0.970	0.069	0.055	0.971
	II: t-1, t-2	0.047	0.037	0.986	0.049	0.039	0.986
	III: t-1, t-2, t-3	0.045	0.035	0.987	0.046	0.037	0.987
	IV: t-1, t-2, t-3, t-4	0.044	0.034	0.987	0.045	0.036	0.988
	V: t-1, t-2, t-3, t-4, t-5	0.043	0.033	0.988	0.044	0.035	0.988
	VI: t-1, t-2, t-3, t-4, t-5, t-6	0.042	0.033	0.988	0.043	0.034	0.989
	VII: t-1, t-2, t-3, t-4, t-5, t-6, t-7	0.042	0.033	0.988	0.043	0.034	0.989
	VIII: t-1, t-2, t-3, t-4, t-5, t-6, t-7, t-8	0.042	0.033	0.988	0.0425	0.0332	0.9892
LSSVR	I: t-1	0.068	0.055	0.969	0.069	0.055	0.971
	II: t-1, t-2	0.047	0.037	0.986	0.048	0.038	0.986
	III: t-1, t-2, t-3	0.045	0.035	0.987	0.045	0.036	0.987
	IV: t-1, t-2, t-3, t-4	0.042	0.033	0.988	0.045	0.035	0.988
	V: t-1, t-2, t-3, t-4, t-5	0.042	0.032	0.989	0.043	0.034	0.989
	VI: t-1, t-2, t-3, t-4, t-5, t-6	0.041	0.032	0.989	0.043	0.033	0.989
	VII: t-1, t-2, t-3, t-4, t-5, t-6, t-7	0.040	0.031	0.989	0.043	0.033	0.989
	VIII: t-1, t-2, t-3, t-4, t-5, t-6, t-7, t-8	0.040	0.031	0.990	0.0427	0.0332	0.9890
M5-Tree	I: t-1	0.067	0.053	0.971	0.074	0.058	0.967
	II: t-1, t-2	0.045	0.035	0.986	0.050	0.040	0.985
	III: t-1, t-2, t-3	0.045	0.035	0.987	0.046	0.037	0.987
	IV: t-1, t-2, t-3, t-4	0.044	0.034	0.987	0.045	0.035	0.988
	V: t-1, t-2, t-3, t-4, t-5	0.043	0.034	0.988	0.0430	0.0340	0.9888
	VI: t-1, t-2, t-3, t-4, t-5, t-6	0.039	0.029	0.990	0.059	0.045	0.979
	VII: t-1, t-2, t-3, t-4, t-5, t-6, t-7	0.038	0.029	0.990	0.058	0.044	0.980
	VIII: t-1, t-2, t-3, t-4, t-5, t-6, t-7, t-8	0.038	0.028	0.990	0.057	0.044	0.980

According to the training results, the input combination (t-1 to t-8) has the most significant effect on forecasting lake levels of t. M5-Tree is the method that gives the least error in the training phase, followed by LSSVR and MARS methods with little difference. In the test phase, the input combinations are compatible with the autocorrelation and partial autocorrelation in Fig. 3 in the MARS and LSSVR methods. However, errors increase after the 5th combination (t-1 to t-5) in the M5-Tree method. LSSVR and M5-Tree methods are followed by the MARS method, which gives the least error and is closest to the best fit. The time series plot of the best results for each method and the scatter plot are given in Figs. 4 and 5.

In the second part of the modeling, the periodicity data component is examined and evaluated. In reality, the main purpose of integrating this periodic sub data, which is one year to forecast one month ahead, is to provide the modeling with an external flow pattern that can provide a more comprehensive understanding and higher outcomes accuracy [18]. The outcomes of the training and testing phase for periodic heuristic regression techniques were shown in Table 3. The addition of the periodicity component increased the average performance in all models. In particular, it has improved the models test performance accuracy in terms of RMSE and MAE by 15.53–13.25 %, 11.24–8.43 % and 4.98–11.08 for best MARS and best LSSVR respectively.

Table 3
The results of heuristic regression techniques with periodicity data component

Model	Inputs	Training			Test		
		RMSE	MAE	R ²	RMSE	MAE	R ²
P-MARS	I: t-1	0.040	0.031	0.989	0.040	0.032	0.991
	II: t-1, t-2	0.035	0.028	0.992	0.036	0.029	0.992
	III: t-1, t-2, t-3	0.035	0.028	0.992	0.036	0.029	0.992
	IV: t-1, t-2, t-3, t-4	0.035	0.028	0.992	0.036	0.029	0.992
	V: t-1, t-2, t-3, t-4, t-5	0.035	0.028	0.992	0.036	0.029	0.992
	VI: t-1, t-2, t-3, t-4, t-5, t-6	0.035	0.028	0.992	0.036	0.029	0.992
	VII: t-1, t-2, t-3, t-4, t-5, t-6, t-7	0.035	0.028	0.992	0.036	0.029	0.992
	VIII: t-1, t-2, t-3, t-4, t-5, t-6, t-7, t-8	0.035	0.028	0.992	0.0359	0.0288	0.9922
P-LSSVR	I: t-1	0.040	0.032	0.989	0.041	0.033	0.990
	II: t-1, t-2	0.035	0.027	0.992	0.036	0.029	0.992
	III: t-1, t-2, t-3	0.035	0.027	0.992	0.037	0.029	0.992
	IV: t-1, t-2, t-3, t-4	0.034	0.027	0.992	0.037	0.029	0.992
	V: t-1, t-2, t-3, t-4, t-5	0.034	0.027	0.992	0.037	0.030	0.992
	VI: t-1, t-2, t-3, t-4, t-5, t-6	0.034	0.027	0.992	0.038	0.030	0.991
	VII: t-1, t-2, t-3, t-4, t-5, t-6, t-7	0.034	0.027	0.992	0.038	0.030	0.991
	VIII: t-1, t-2, t-3, t-4, t-5, t-6, t-7, t-8	0.034	0.027	0.992	0.0379	0.0304	0.9912
P-M5-Tree	I: t-1	0.039	0.029	0.990	0.057	0.045	0.981
	II: t-1, t-2	0.033	0.025	0.993	0.053	0.042	0.983
	III: t-1, t-2, t-3	0.032	0.024	0.993	0.055	0.044	0.982
	IV: t-1, t-2, t-3, t-4	0.033	0.024	0.993	0.055	0.044	0.981
	V: t-1, t-2, t-3, t-4, t-5	0.032	0.024	0.993	0.055	0.044	0.981
	VI: t-1, t-2, t-3, t-4, t-5, t-6	0.032	0.023	0.993	0.055	0.044	0.982
	VII: t-1, t-2, t-3, t-4, t-5, t-6, t-7	0.032	0.023	0.993	0.054	0.043	0.982
	VIII: t-1, t-2, t-3, t-4, t-5, t-6, t-7, t-8	0.031	0.023	0.994	0.055	0.044	0.981

The time series plot of the best results for methods and the scatter plot are given in Figs. 6 and 7.

The graphs in Figs. 4 and 5 better represented the values observed in Figs. 6 and 7 with the effect of periodicity. The values observed during the training phase in Figs. 4 and 6 were generally captured by the models. In other words, it was generally forecasted correctly. However, while the long-term periodic fluctuations of the values observed in the test phase in Figs. 5 and 7 were well predicted, the short-time fluctuations were underestimated according to the three methods. In Table 2 better represented the values observed in Table 3 with the effect of periodicity. As can be seen from the table, the P-MARS model (VIII) in all data sets gave lower RMSE, MAE and higher R^2 values than the others (RMSE: 0.0359 MAE: 0.0288 R^2 : 0.9922). The worst results were obtained from the (I) 1 input model using the M5-Tree model (RMSE: 0.074 MAE: 0.058 R^2 : 0.967). In Figs. 7, it can be observed that the P-MARS (VIII inputs) model gives better estimation than the others, especially in the scatter diagrams (assuming the equation is $y = ax + b$), and the coefficients a and b (in the linear equation a and b are closer to 1 and 0, respectively). The reason behind this is that the MARS structure and the periodicity data component can accurately model the highly nonlinear lake level process [18].

5. Conclusion

In the current study, the applicability of MARS, M5-Tree, and LSSVR models in forecasting lake level fluctuations has been investigated. Lake-level observations from Lake Michigan in U.S. were used for training and testing of the applied models. In terms of performance indices, the obtained findings demonstrated the effectiveness of the employed models in reproducing the nonlinear behavior of lake level fluctuation.

- In both scenario forecasting, the MARS slightly better than LSSVR, and M5-Tree models.
- In general, P-MARS indicated better forecasted accuracies at input combinations VIII. Indeed, this is due to the capability of the application of multivariate adaptive regression which is to the skill to capture the complicated non-linear relationship.
- Modeling using a single input (I) gave the worst result in the estimation made with the M5-Tree.
- The periodic component feature was embedded and evaluated inside the modeling's input data sets, and the results revealed that integrating this component data was quite useful in offering a detailed intuition into the process of anticipated monthly lake levels.

Where resources are not available to operate complicated physically-based models, the proposed heuristic regression techniques may be a useful practical option for improved monthly lake level forecasts. In operational water level forecasting, the presented methods could be a valuable supplement to physical models. Understanding the causes of water level variations and the factors that influence them can help with lake conservation and management. It's critical to keep water levels in lake Michigan at a healthy level for the ecosystems and marshes that surround them. Effective strategies for

sustainable integrated water resources management should be implemented to preserve ecological integrity and assure the water release and storage capacity of Great Lakes under the pressure of unpredictable climate variables.

6. Declarations

Acknowledgments: There is not any conflict of interest.

Conflicts of Interest: The authors declare no conflicts of interest.

Funding: No funding or support has been received for research from funding institutions.

Author's Contribution: All chapters have been prepared by Vahdettin Demir.

Availability of data and material: Data are available U.S. Army Corps of Engineers.

Code availability: Not applicable.

Ethics approval: This study did not involve any protected area, private land, and endangered or protected species. And no specific permissions were required for this activity.

Consent to participate: Not applicable.

Consent for publication: Written informed consent for publication was obtained from all authors.

7. References

1. Şen, Z., Kadioğlu, M., Batur, E.: Stochastic modeling of the Van Lake monthly level fluctuations in Turkey. *Theor. Appl. Climatol.* 65, 99–110 (2000). <https://doi.org/10.1007/s007040050007>
2. Vuglinskiy, V.: Water Level: water level in lakes and reservoirs, water storage. , Assessment of the status of the development of the standards for the terrestrial essential climate variables, Global Terrestrial Observing System (GTOS), Rome, Italy (2009)
3. Shiri, J., Shamshirband, S., Kisi, O., Karimi, S., Bateni, S.M., Hosseini Nezhad, S.H., Hashemi, A.: Prediction of Water-Level in the Urmia Lake Using the Extreme Learning Machine Approach. *Water Resour. Manag.* 30, 5217–5229 (2016). <https://doi.org/10.1007/s11269-016-1480-x>
4. Karimi, S., Shiri, J., Kisi, O., Makarynsky, O.: Forecasting Water Level Fluctuations of Urmieh Lake Using Gene Expression Programming and Adaptive Neuro-Fuzzy Inference System. *Int. J. Ocean Clim. Syst.* 3, 109–125 (2012). <https://doi.org/10.1260/1759-3131.3.2.109>
5. Altunkaynak, A.: Forecasting surface water level fluctuations of lake van by artificial neural networks. *Water Resour. Manag.* 21, 399–408 (2007). <https://doi.org/10.1007/s11269-006-9022-6>
6. Altunkaynak, A., Özger, M., Sen, Z.: Triple diagram model of level fluctuations in Lake Van, Turkey. *Hydrol. Earth Syst. Sci.* 7, 235–244 (2003). <https://doi.org/10.5194/hess-7-235-2003>

7. Altunkaynak, A., Şen, Z.: Fuzzy logic model of lake water level fluctuations in Lake Van, Turkey. *Theor. Appl. Climatol.* 90, 227–233 (2007). <https://doi.org/10.1007/s00704-006-0267-z>
8. Kişi, Ö.: Neural network and wavelet conjunction model for modelling monthly level fluctuations in Turkey. *Hydrol. Process.* 23, 2081–2092 (2009). <https://doi.org/10.1002/hyp.7340>
9. Sanikhani, H., Kisi, O., Kiafar, H.: Comparison of Different Data-Driven Approaches for Modeling Lake Level Fluctuations: The Case of Manyas and Tuz Lakes (Turkey). *Water Resour Manag.* 29, 1557–1574 (2015). <https://doi.org/doi.org/10.1007/s11269-014-0894-6>
10. Young, C.C., Liu, W.C., Hsieh, W.L.: Predicting the Water Level Fluctuation in an Alpine Lake Using Physically Based, Artificial Neural Network, and Time Series Forecasting Models. *Math. Probl. Eng.* 2015, (2015). <https://doi.org/10.1155/2015/708204>
11. Shafaei, M., Kisi, O.: Lake Level Forecasting Using Wavelet-SVR, Wavelet-ANFIS and Wavelet-ARMA Conjunction Models. *Water Resour. Manag.* 30, 79–97 (2016). <https://doi.org/10.1007/s11269-015-1147-z>
12. Liang, C., Li, H., Lei, M., Du, Q.: Dongting Lake water level forecast and its relationship with the Three Gorges Dam based on a long short-term memory network. *Water (Switzerland)*. 10, (2018). <https://doi.org/10.3390/w10101389>
13. Peprah, M.S., Larbi, E.K., Peprah, M.S.: International Journal of Earth Sciences Knowledge and Applications Lake Water Level Prediction Model Based on Autocorrelation Regressive Integrated Moving Average and Kalman Filtering Techniques – An Empirical Study on Lake Volta Basin , Ghana. *Int. J. Earth Sci. Knowl. Appl.* 3, 1–11 (2021)
14. Luo, Y., Dong, Z., Liu, Y., Wang, X., Shi, Q., Han, Y.: Research on stage-divided water level prediction technology of rivers-connected lake based on machine learning: a case study of Hongze Lake, China. *Stoch. Environ. Res. Risk Assess.* 6, (2021). <https://doi.org/10.1007/s00477-021-01974-6>
15. Kisi, O., Shiri, J., Demir, V.: Hydrological Time Series Forecasting Using Three Different Heuristic Regression Techniques. In: *Handbook of Neural Computation*. pp. 45–65. Elsevier (2017)
16. Demir, V., Çubukçu, E.A.: Digital Elevation Modeling with Heuristic Regression Techniques Abstract. *Eur. J. Sci. Technol.* 484–488 (2021). <https://doi.org/10.31590/ejosat.916012>
17. Kisi, O., Parmar, K.S., Soni, K., Demir, V.: Modeling of air pollutants using least square support vector regression, multivariate adaptive regression spline, and M5 model tree models. *Air Qual. Atmos. Heal.* 10, (2017). <https://doi.org/10.1007/s11869-017-0477-9>
18. Yaseen, Z.M., Kisi, O., Demir, V.: Enhancing Long-Term Streamflow Forecasting and Predicting using Periodicity Data Component: Application of Artificial Intelligence. *Water Resour. Manag.* 30, (2016). <https://doi.org/10.1007/s11269-016-1408-5>
19. Suykens, J.A.K., Vandewalle, J.: No Title. *Neural Process. Lett.* 9, 293–300 (1999). <https://doi.org/10.1023/A:1018628609742>
20. Ji, Z., Wang, B., Deng, S., You, Z.: Predicting dynamic deformation of retaining structure by LSSVR-based time series method. *Neurocomputing.* 137, 165–172 (2014). <https://doi.org/10.1016/j.neucom.2013.03.073>

21. Huang, Z., Luo, J., Li, X., Zhou, Y.: Prediction of Effluent Parameters of Wastewater Treatment Plant Based on Improved Least Square Support Vector Machine with PSO. In: 2009 First International Conference on Information Science and Engineering. pp. 4058–4061. IEEE (2009)
22. Deng, S., Yeh, T.H.: Applying least squares support vector machines to the airframe wing-box structural design cost estimation. *Expert Syst. Appl.* 37, 8417–8423 (2010). <https://doi.org/10.1016/j.eswa.2010.05.038>
23. Pahasa, J., Ngamroo, I.: A heuristic training-based least squares support vector machines for power system stabilization by SMES. *Expert Syst. Appl.* 38, 13987–13993 (2011). <https://doi.org/10.1016/j.eswa.2011.04.206>
24. Shokrollahi, A., Arabloo, M., Gharagheizi, F., Mohammadi, A.H.: Intelligent model for prediction of CO₂ – Reservoir oil minimum miscibility pressure. *Fuel*. 112, 375–384 (2013). <https://doi.org/10.1016/j.fuel.2013.04.036>
25. Kamari, A., Nikookar, M., Sahranavard, L., Mohammadi, A.H.: Efficient screening of enhanced oil recovery methods and predictive economic analysis. *Neural Comput. Appl.* 25, 815–824 (2014). <https://doi.org/10.1007/s00521-014-1553-9>
26. Hemmati-Sarapardeh, A., Shokrollahi, A., Tatar, A., Gharagheizi, F., Mohammadi, A.H., Naseri, A.: Reservoir oil viscosity determination using a rigorous approach. *Fuel*. 116, 39–48 (2014). <https://doi.org/10.1016/j.fuel.2013.07.072>
27. Kisi, O.: Streamflow Forecasting and Estimation Using Least Square Support Vector Regression and Adaptive Neuro-Fuzzy Embedded Fuzzy c-means Clustering. *Water Resour. Manag.* 29, 5109–5127 (2015). <https://doi.org/10.1007/s11269-015-1107-7>
28. Hwang, S.H., Ham, D.H., Kim, J.H.: Forecasting performance of LS-SVM for nonlinear hydrological time series. *KSCE J. Civ. Eng.* 16, 870–882 (2012). <https://doi.org/10.1007/s12205-012-1519-3>
29. Kisi, O.: Modeling discharge-suspended sediment relationship using least square support vector machine. *J. Hydrol.* 456–457, 110–120 (2012). <https://doi.org/10.1016/j.jhydrol.2012.06.019>
30. Kisi, O.: Least squares support vector machine for modeling daily reference evapotranspiration. *Irrig. Sci.* 31, 611–619 (2013). <https://doi.org/10.1007/s00271-012-0336-2>
31. Okkan, U., Ali Serbes, Z.: The combined use of wavelet transform and black box models in reservoir inflow modeling. *J. Hydrol. Hydromechanics*. 61, 112–119 (2013). <https://doi.org/10.2478/johh-2013-0015>
32. Kisi, O., Parmar, K.S.: Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. *J. Hydrol.* 534, 104–112 (2016). <https://doi.org/10.1016/j.jhydrol.2015.12.014>
33. Friedman, J.H.: Multivariate Adaptive Regression Splines. *Ann. Stat.* 19, 590–606 (1991). <https://doi.org/10.1214/aos/1176347963>
34. Leathwick, J.R., Elith, J., Hastie, T.: Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecol. Modell.* 199, 188–196 (2006). <https://doi.org/10.1016/j.ecolmodel.2006.05.022>

35. Sotomayor, K.A.L.: Comparison of adaptive methods using multivariate regression splines (MARS) and artificial neural networks backpropagation (ANNB) for the forecast of rain and temperatures in the Mantaro river basin. *Hydrol. Days* 2010. 58–68 (2010)
36. Adamowski, J., Chan, H.F., Prasher, S.O., Sharda, V.N.: Comparison of multivariate adaptive regression splines with coupled wavelet transform artificial neural networks for runoff forecasting in Himalayan micro-watersheds with limited data. *J. Hydroinformatics*. 14, 731–744 (2012).
<https://doi.org/10.2166/hydro.2011.044>
37. Kisi, O.: Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *J. Hydrol.* 528, 312–320 (2015).
<https://doi.org/10.1016/j.jhydrol.2015.06.052>
38. Shortridge, J.E., Guikema, S.D., Zaitchik, B.F.: Empirical streamflow simulation for water resource management in data-scarce seasonal watersheds. *Hydrol. Earth Syst. Sci. Discuss.* 12, 11083–11127 (2015). <https://doi.org/10.5194/hessd-12-11083-2015>
39. Quinlan, J.R.: Learning With Continuous Classes. *World Sci.* 92, 343–348 (1992).
<https://doi.org/10.1.1.34.885>
40. Bhattacharya, B., Solomatine, D.P.: Neural networks and M5 model trees in modelling water level-discharge relationship. *Neurocomputing*. 63, 381–396 (2005).
<https://doi.org/10.1016/j.neucom.2004.04.016>
41. Solomatine, D.P., Dulal, K.N.: Model trees as an alternative to neural networks in rainfall-runoff modelling. *Hydrol. Sci. J.* 48, 399–411 (2003). <https://doi.org/10.1623/hysj.48.3.399.45291>
42. Pal, M., Deswal, S.: M5 model tree based modelling of reference evapotranspiration. *Hydrol. Process.* 23, 1437–1443 (2009). <https://doi.org/10.1002/hyp.7266>
43. Solomatine, D.P., Xue, Y.: M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China. *J. Hydrol. Eng.* 9, 491–501 (2004).
[https://doi.org/10.1061/\(ASCE\)1084-0699\(2004\)9:6\(491\)](https://doi.org/10.1061/(ASCE)1084-0699(2004)9:6(491))
44. Goyal, M.K.: Modeling of Sediment Yield Prediction Using M5 Model Tree Algorithm and Wavelet Regression. *Water Resour. Manag.* 28, 1991–2003 (2014). <https://doi.org/10.1007/s11269-014-0590-6>
45. Michigan Sea Grant, <https://www.michiganseagrant.org/topics/great-lakes-fast-facts/lake-michigan/>
46. Michigan, <https://www.goller.gen.tr/michigan-golu.html>
47. Mitchell, T.M.: Machine learning. The McGraw-Hill Companies, New York (1997)
48. De Andrés, J., Lorca, P., De Cos Juez, F.J., Sánchez-Lasheras, F.: Bankruptcy forecasting: A hybrid approach using fuzzy c-means clustering and multivariate adaptive regression splines (MARS). *Expert Syst. Appl.* 38, 1866–1875 (2011). <https://doi.org/10.1016/j.eswa.2010.07.117>
49. Sharda, V.N., Patel, R.M., Prasher, S.O., Ojasvi, P.R., Prakash, C.: Modeling runoff from middle Himalayan watersheds employing artificial intelligence techniques. *Agric. Water Manag.* 83, 233–242 (2006). <https://doi.org/10.1016/j.agwat.2006.01.003>

50. Sephton, P.: Forecasting recessions: can we do better on MARS? Review. 83, 39–49 (2001)
51. P. Bera, S. O. Prasher, R. M. Patel, A. Madani, R. Lacroix, J. D. Gaynor, C. S. Tan, S. H. Kim: Application of MARS in Simulating Pesticide Concentrations in Soil. Trans. ASABE. 49, 297–307 (2006).
<https://doi.org/10.13031/2013.20228>
52. Legates, D.R., McCabe, G.J.: Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35, 233–241 (1999).
<https://doi.org/10.1029/1998WR900018>
53. Coulibaly, P.: Reservoir Computing approach to Great Lakes water level forecasting. J. Hydrol. 381, 76–88 (2010). <https://doi.org/10.1016/j.jhydrol.2009.11.027>

Figures

Figure 1

Study area: Great Lakes (a) and Lake Michigan (b) [45].

Data set	Period	X_{mean} (m)	X_{min} (m)	X_{max} (m)	Cs (m)	Sx (m)	Basık
All data	1918-2020	176.44	175.57	177.5	0.119	0.409	-0.763
Training	1918-1999	176.48	175.58	177.5	-0.075	0.389	-0.644
Test	1999-2020	176.28	175.57	177.46	0.949	0.445	-0.018

Figure 2

Lake Michigan water level fluctuations and Training-test data sets

Figure 3

Monthly lake level autocorrelation and partial autocorrelation coefficients for Lake Michigan

Figure 4

The observed and forecasted lake level time series and scatter plot for training phase

Figure 5

The observed and forecasted lake level time series and scatter plot for test phase

Figure 6

The observed and forecasted lake level time series and scatter plot for periodic training phase.

Figure 7

The observed and forecasted lake level time series and scatter plot for periodic test phase