



Enhancing monthly lake levels forecasting using heuristic regression techniques with periodicity data component: application of Lake Michigan

Vahdettin Demir¹

Received: 19 July 2021 / Accepted: 14 February 2022 / Published online: 21 February 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2022

Abstract

This study investigates the accuracy of three different techniques with the periodicity component for estimating monthly lake levels. The three techniques are multivariate adaptive regression splines (MARS), least-square support vector regression (LSSVR), and M5 model tree (M5-tree). Data from Lake Michigan, located in the USA, is used in the analysis. In the first stage of modeling, three techniques were applied to forecast monthly lake level fluctuations up to 8 months 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 subset in modeling monthly lake levels. The root-mean-square error, mean absolute error, and coefficient of determination were used for evaluating the accuracy of the models. In both stages, the comparison results indicate that MARS generally outperforms LSSVR and M5-tree. Further, 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, designing, building, and operation, as well as for managing freshwater lakes for water supply purposes. To regulate future lake level fluctuations, methods for modeling high or abnormal level fluctuations should be devised. The level measurements, or their future equally likely reproductions acquired using a simulation model, are straightforward manners of obtaining 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 models that simulate lake level fluctuations based on past level records (Şen et al. 2000).

Lakes are used for various domestic, industrial, and agricultural purposes (Vuglinskiy 2009; Shiri et al. 2016). Forecasting lake water levels is crucial for water resource planning and management, lake navigation, tidal irrigation,

and agricultural drainage canal management, etc. Lake water level is a complicated phenomenon; it is primarily influenced by natural water exchange between the lake and its watershed, and consequently reflects hydrological changes in the watershed (Altunkaynak 2007; Karimi et al. 2012). For many practical applications, a model that forecasts lake level fluctuations based on previously measured levels is required (Karimi et al. 2012).

Over the last few decades, numerous researchers have studied lake water level models because global climate change impacts the hydrological cycle, causing many lakes to dry up or flood unexpectedly. To model lake level fluctuations, several techniques have been devised. Şen et al. (2000) employed periodic and stochastic processes. Altunkaynak et al. (2003) used the diagram model and Markov process. Altunkaynak (2007) employed an artificial neural network. Altunkaynak and Şen (2007) used fuzzy logic. Kişi (2009) used a wavelet conjunction model. Karimi et al. (2012) employed gene expression programming and an adaptive neuro-fuzzy inference system (ANFIS). Sanikhani et al. (2015) also used ANFIS and gene expression programming. Young et al. (2015) used a timeseries forecasting model. Shiri et al. (2016) employed an extreme learning machine approach. Shafaei and Kisi (2016) employed the wavelet-support vector regression (SVR), wavelet-ANFIS, and wavelet-autoregressive

✉ Vahdettin Demir
vahdettin.demir@karatay.edu.tr

¹ Faculty of Engineering and Natural Sciences, KTO Karatay University, Konya 42020, Turkey

moving average model conjunction models. Liang et al. (2018) used a deep learning method. Peprah et al. (2021) employed integrated moving average and kalman filtering techniques. Luo et al. (2021) used machine learning methods.

Recently, three machine learning techniques, multivariate adaptive regression splines (MARS), least-square SVR (LSSVR), and M5 model tree (M5-tree), have achieved a remarkable emerging and promise in addressing difficult nonlinear situations. These techniques have been widely employed in solving hydrologic challenges (Yaseen et al. 2016; Kisi et al. 2017a, b; Demir and Çubukçu 2021). MARS is a newer artificial intelligence technique (Friedman 1991). 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 its main advantages. Further, this nonparametric statistical method provides a versatile procedure for organizing the relationship between input and output variables with fewer variable interactions (Leathwick et al. 2006). Rainfall and temperature forecasting, streamflow forecasting, sediment concentration estimate, water pollution forecast, air pollutant forecast, freshwater distribution system modeling, and drought events river flow simulation are previous water resources applications of MARS (Leathwick et al. 2006; Sotomayor 2010; Adamowski et al. 2012; Kisi 2015a; Shortridge et al. 2015; Kisi and Parmar 2016; Yaseen et al. 2016; Kisi et al. 2017b).

LSSVR is a modified variant of SVR that can solve problems involving quadratic programming (Suykens and Vandewalle 1999). It also avoids some flaws that other data-driven learning systems have (e.g., local minima, time consumption, and overfitting) (Ji et al. 2014). In the engineering field, LSSVR has been successfully applied; e.g., to predict wastewater effluent parameters (Huang et al. 2009), design the structural components of a wing-box for an airplane (Deng and Yeh 2010), design a superconducting magnetic energy storage controller with adaptive dampening (Pahasa and Ngamroo 2011), forecast CO₂ emission in reservoir (Shokrollahi et al. 2013), analyze oil recovery (Kamari et al. 2014), forecast reservoir oil viscosity (Hemmati-Sarapardeh et al. 2014). In the hydrological study, few studies have been conducted using LSSVR; for example, streamflow forecasting and estimation (Kisi 2015b; Yaseen et al. 2016; Kisi et al. 2017b), daily water demand and dam daily inflow estimation of (Hwang et al. 2012), sediment transport modeling (Kisi 2012), daily reference evapotranspiration modeling (Kisi 2013), reservoir inflow modeling (Okkan and Ali Serbes 2013), water pollution forecast (Kisi and Parmar 2016), air pollutant forecast (Kisi et al. 2017a).

M5-tree is a data mining technique that uses the divide-and-conquer method to split data timeseries into subspaces, allowing a multidimensional parameter space to be divided and the model to be automatically generated on the basis of the overall quality requirement (Quinlan 1992). Scholars recently investigated the M5-tree's utility in many hydrological applications, such as water level optimization (Bhattacharya and Solomatine 2005), precipitation and river flow modeling (Solomatine and Dulal 2003), streamflow modeling (Yaseen et al. 2016), wind speed modeling (Başakın et al. 2022), air pollutant modeling (Kisi et al. 2017a), evapotranspiration modeling (Pal and Deswal 2009), pan-evaporation modeling (Kisi 2015a), flood events (Solomatine and Xue 2004), and sediment yield modeling (Goyal 2014).

Based on the reported database of Scopus for “machine learning” and “lake level” over 161 document results were appeared. A set of significant keywords for this study domain has been created using the VOSviewer algorithm (Fig. 1a). In addition, when the adopted research is analyzed based on the time scale (Fig. 1a), it is seen that many studies have been published in 2016 and beyond. These studies seem to have more research interest on data science, prediction, time series, water quality, classification ect., and new machine learning models such as deep learning, support vector machine, random forest, regression tree, extreme learning machine etc. Figure 1b shows the main regions where machine learning and lake level have been investigated. It is the region of China with the most research (47), followed by USA (40), Canada (18), Iran (17).

In this study, the major goals are to (i) investigate three different novel heuristic regression techniques (MARS, LSSVR, and M5-tree) for modeling water level forecasting, (ii) investigate the influence of the periodicity component (months of the observed data) for water level forecasting, (iii) demonstrate the effectiveness of the proposed models; Lake Michigan in the USA was employed.

2 Case study and data preparation

The name, Lake Michigan comes from the Ojibwa term, Michi Gami, which means “large lake”. Lake Michigan is in the USA (coordinates: 44°N 87°W); it is the third-largest lake in the Lake District, comprising five interconnected large lakes, and the sixth-largest freshwater lake globally (see Fig. 2). With a surface area of 58,016 km², drainage area of 118,095 km², a width of 48–193 km, a length of 494 km, and deepest point of 0.281 km, 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

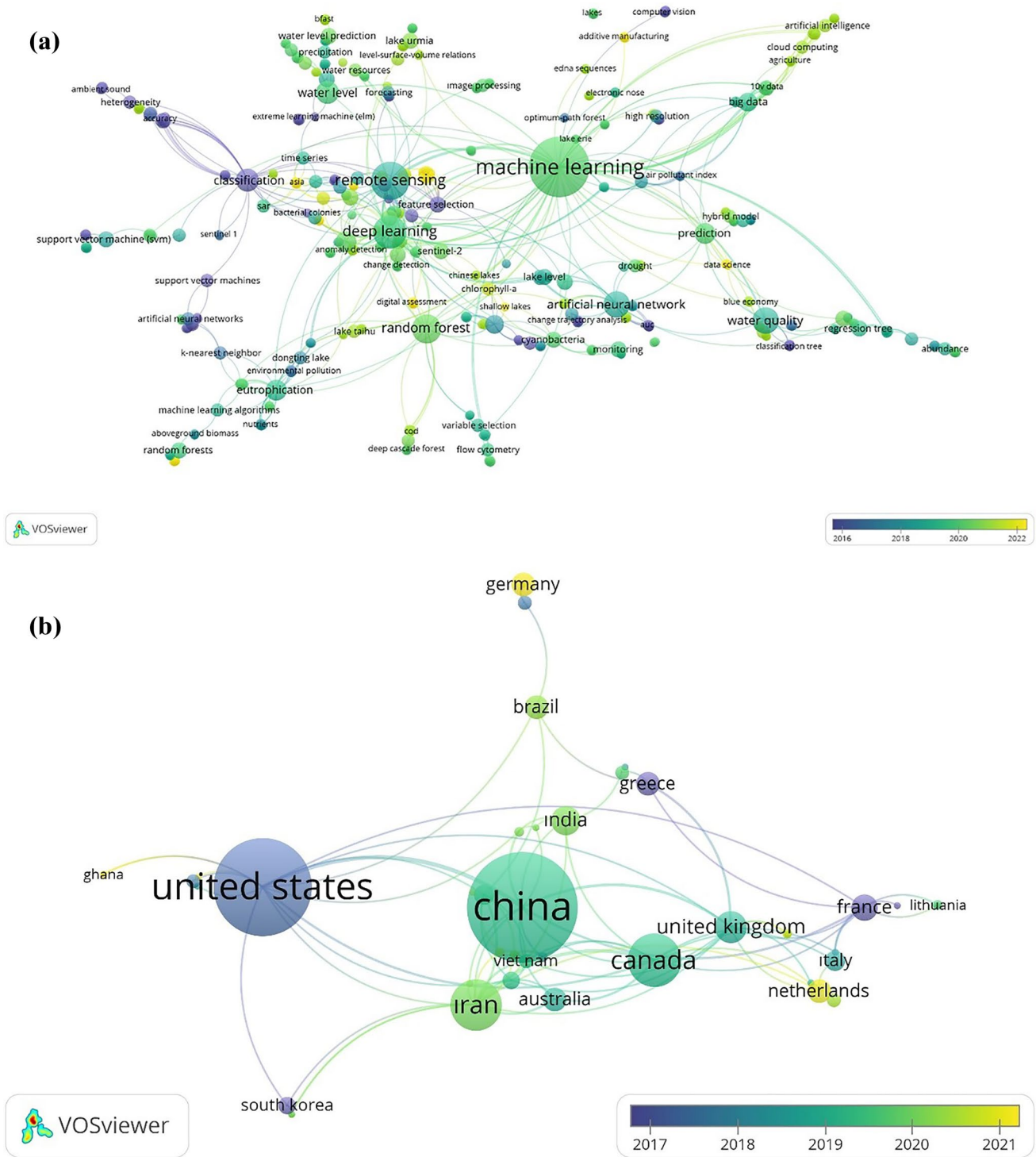


Fig. 1 The literature review keywords (a) and research regions (b)

(Michigan 2021a). 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 0.177 km above sea level. It is connected by the Strait of Mackinac to Lakes Superior, Huron, Erie, and Ontario from its northeast corner (Michigan 2021b).

Forecasting lake level fluctuations is critical for many operations in Lake Michigan region, including flood mitigation, reservoir management, drinking water distribution, water infrastructure management, trade, transportation, and beach erosion. The observed data are 103 years (1236 months) long with an observation

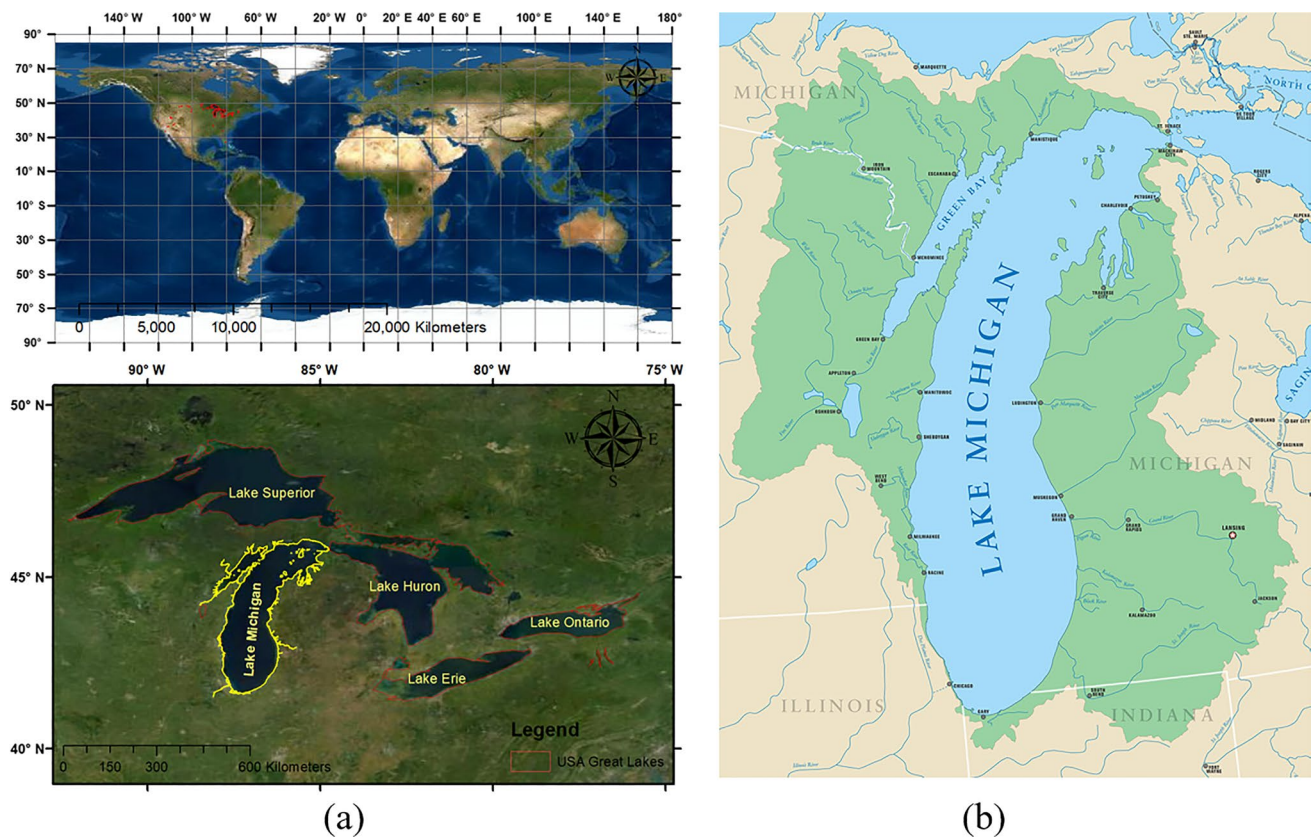


Fig. 2 Study area: (a) Great Lakes, and (b) Lake Michigan (Michigan 2021a) (b)

period between 1918 and 2020 for the 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. 3.

The partial autocorrelation and autocorrelation functions of the lake levels for Lake Michigan are also shown in Fig. 4. The figure shows that the lake level in Lake Michigan highly correlates with past month levels. The partial autocorrelation function indicates a significant correlation up to lag 8 for Lake Michigan and then stays within the confidence interval.

Table 1 Monthly statistical information of datasets for Lake Michigan

| Data set | Period | X_{mean} (m) | X_{min} (m) | X_{max} (m) | Cs_x | S_x | Ck_x |
|----------|-----------|-----------------------|----------------------|----------------------|--------|-------|--------|
| 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 |

3 Methods

3.1 MARS

Friedman proposed the MARS model, which is a nonparametric regression model (Friedman 1991). MARS is a model used to forecast nonlinear continuous numerical results. It explains the complex nonlinear relationship between a model, estimation method, and dependent variables. The MARS algorithm comprises two steps: forward and backward steps. It selects a set of suitable input variables with the forward step algorithm (De Andrés et al. 2011). With the backward step algorithm, it eliminates unnecessary variables in the preselected set. A graph 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 (Sharda et al. 2006):

Fig. 3 Lake Michigan water level fluctuations and training–test datasets

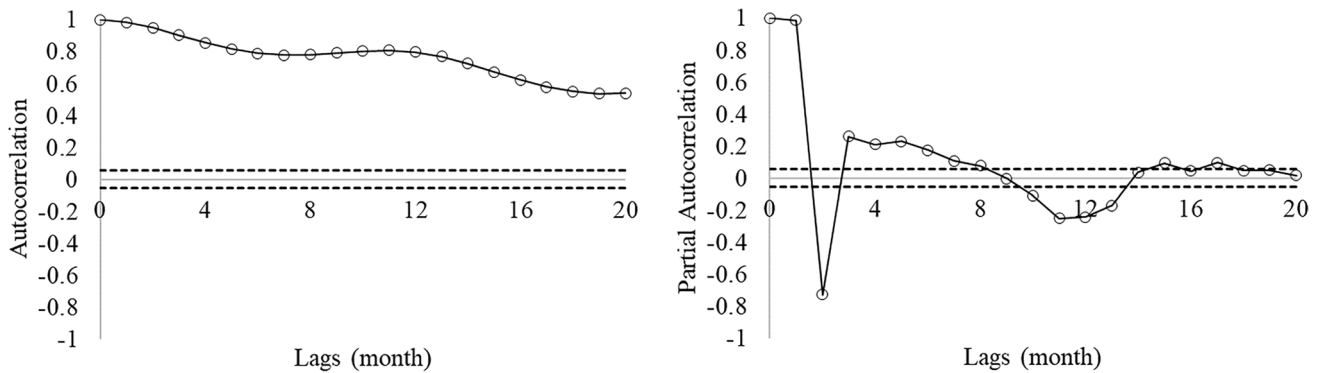
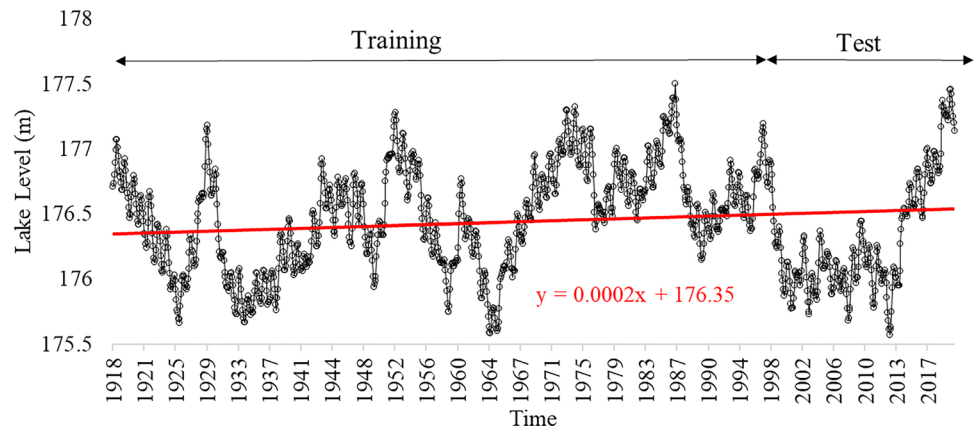


Fig. 4 Monthly lake level autocorrelation and partial autocorrelation coefficients for Lake Michigan

$$Y = \max(0, X - c) \tag{1}$$

$$Y = \max(0, c - x) \tag{2}$$

where c represents the threshold (lower limit) value. The MARS model is used especially in financial affairs management systems and timeseries data (Sephton 2001; Bera et al. 2006; Yaseen et al. 2016; Kisi et al. 2017a, b; Demir and Çubukçu 2021).

3.2 LSSVR

LSSVR is an extension of SVR, proposed by Suykens and Vandewalle in 1999 (Suykens and Vandewalle 1999). It is employed to statistically estimate water levels with the water levels in the historical timeseries and obtain the optimum function between the X input and Y output (Yaseen et al. 2016). It performs this operation with a nonlinear relationship function in a multidimensional feature space. The regression function can be expressed as follows:

$$y(x) = w^T \varphi(x) + b \tag{3}$$

Where y is the value obtained in x , w is the coefficient vector, φ is the mapping function, b is the bias term obtained from minimizing the generalization error's upper bound (Suykens and Vandewalle 1999).

3.3 M5-tree

M5-tree algorithm is a new regression method developed by Quinlan in 1992 (Quinlan 1992). Its backbone is a two-component decision tree. The method defines the relationship between the independent and dependent variables with a linear regression function applied to the final leaf nodes. M5-tree is better than other decision tree models used for categorical data (Mitchell 1997).

M5-tree is a two-stage model. In the first stage, data are split into subsets to produce the decision schema (tree). The standard deviation of the class value reached at a node is used to classify. The expected reduction is calculated on the basis of the error that occurs due to testing the elements acting on this node. (Solomatine and Xue 2004; Pal and Deswal 2009). The expression of the standard deviation reduction (SDR) is as follows.

$$SDR = sd(T) - \sum \frac{|Ti|}{|T|} sd(Ti) \tag{4}$$

In this formula, sd is the standard deviation, and T represents a set of instances acting on the node. Subset samples with “i” results of potential data are represented by Ti (Quinlan 1992).

4 Results and discussion

The three heuristic regression techniques evaluated (MARS, LSSVR and M5-Tree) were created using MATLAB subroutines to estimate the lake levels forecasting. The data were divided into the training and test datasets before modeling. The training dataset accounts for 80% (1236 × 0.8 = 989), whereas the test dataset accounts for 20% (247). Quantitative indicators are commonly used to evaluate hydrological applications. Legates and McCabe (1999) suggested that predictive models in the hydrology field be tested using goodness-of-fit methods, e.g., root-mean-square error (RMSE), mean absolute error (MAE), and coefficient of

determination (R²), as shown in Eqs. (5)–(7), respectively (Legates and McCabe 1999).

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

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

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

In Eqs. (5)–(7), L_e and L_o indicate the estimated and observed water levels values, respectively, and N indicates the raw water level amount of data. This study aims to forecast lake level fluctuations by MARS, M5-Tree, and LSSVR. In this context, different input combinations were explored to forecast the water levels. The inputs include the previous

Table 2 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 |

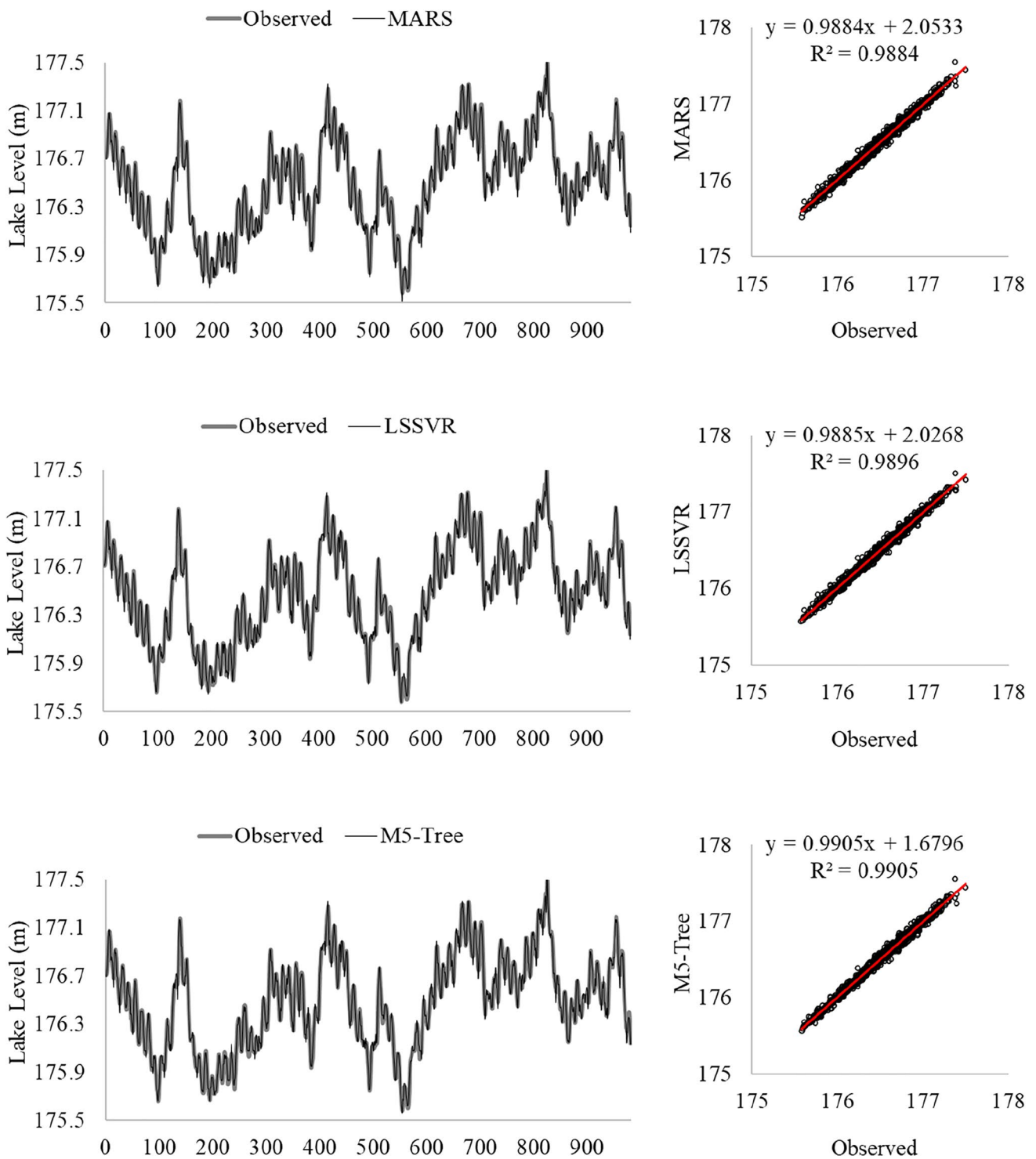


Fig. 5 Observed and forecasted lake level timeseries and scatter plots for the training phase

monthly lake levels ($t - 1, t - 2, t - 3, t - 4, t - 5, t - 6, t - 7, t - 8$), and the outputs correspond to the lake level at time t .

The results of heuristic regression techniques in terms of RMSE, MAE, and R^2 are summarized in Table 2 with

input combinations. $RMSE \leq 60$ cm indicates an excellent and appropriate estimate (Coulibaly 2010; Sanikhani et al. 2015). For Lake Michigan, $RMSE \leq 7.4$ cm is an excellent and satisfactory estimate.

According to the training results, the input combination ($t-1$ to $t-8$) had the most significant effect on forecasting lake levels of t . M5-tree yielded the least error in the training phase, followed by LSSVR and MARS with little difference. In the test phase, the input combinations were compatible

with the autocorrelation and partial autocorrelation in Fig. 4 in MARS and LSSVR. However, errors increased after the fifth combination ($t-1$ to $t-5$) in M5-tree. MARS were followed by LSSVR and M5-tree, which yielded the least error and was closest to the best fit. The timeseries plot of the best

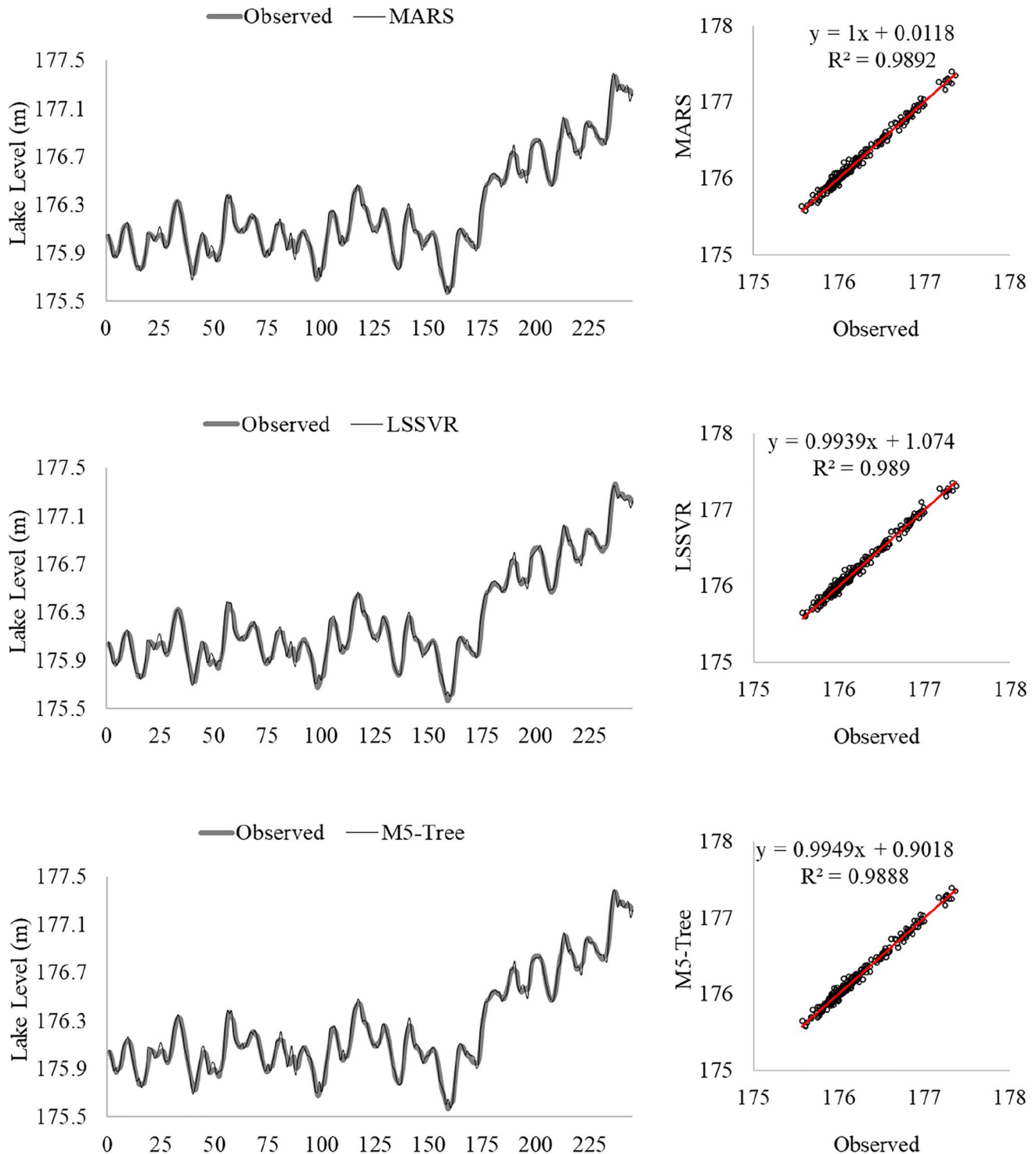


Fig. 6 Observed and forecasted lake level timeseries and scatter plots for the test phase

results for each method and the scatter plot are depicted in Figs. 5 and 6.

In the second modeling part, the periodicity data component was examined and evaluated. In reality, the main purpose of integrating this periodic subdata, which is one year to forecast one month ahead, was to provide the modeling with an external flow pattern that could yield a more comprehensive understanding and higher outcome accuracy (Yaseen et al. 2016). The outcomes of the training and test phase for periodic heuristic regression techniques are summarized in Table 3. The addition of the periodicity component increased the average performance in all models. In particular, it improved the model test performance accuracy in terms of RMSE and MAE by 15.53–13.25% (For example RMSE for best MARS: $0.0425 - 0.0425 \times 0.1553 = 0.0359$), 11.24–8.43%, and 4.98–11.08% for best MARS and best LSSVR, respectively.

The timeseries plot of the best results for the methods and the scatter plots are depicted in Figs. 7 and 8.

Figures 5 and 6 better represented the values observed in Figs. 7 and 8 with the effect of periodicity. The values observed during the training phase in Figs. 5 and 6 were generally captured by the models. In other words, it was

generally forecasted correctly. However, although the long-term periodic fluctuations of the values observed in the test phase in Figs. 7 and 8 were well predicted, the short-time fluctuations were under estimated according to the three techniques. Table 2 better represented the values observed in Table 3 with the effect of periodicity. From Table 2, P-MARS (VIII inputs) in all datasets yielded lower RMSE and 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 M5-tree (RMSE = 0.074; MAE = 0.058; $R^2 = 0.967$). In Fig. 8, P-MARS (VIII inputs) yielded better estimates than 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 was that the MARS structure and the periodicity data component could accurately model the highly nonlinear lake level process (Yaseen et al. 2016).

Although MAE, RMSE, and R^2 error criteria demonstrated the accuracy of the estimated variables, these error statistics do not reveal information about the distribution of the models (Citakoglu 2021). Therefore, the Taylor diagram and Violin plot containing statistical analysis was used while

Table 3 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 |

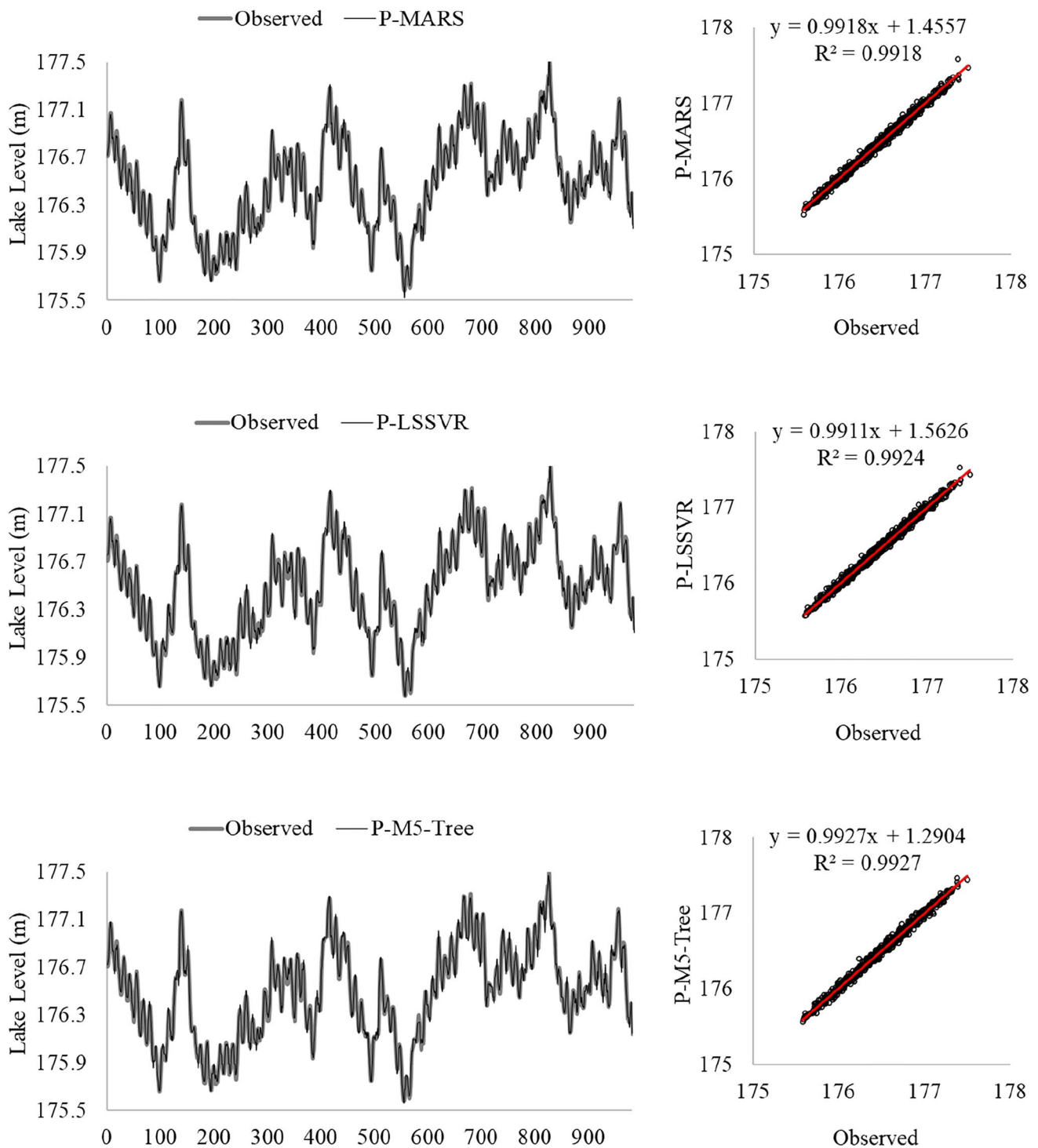


Fig. 7 Observed and forecasted lake level timeseries and scatter plots for the periodic training phase

comparing methods (Taylor 2001; Legouhy 2021; Başakın et al. 2022). Taylor diagram assessed compliance of estimation data with observed data. With the use of the Taylor diagram, further comparisons of the models were achieved.

Taylor diagrams for the best result of MARS, LSSVR, M5-tree models and P-MARS, P-LSSVR, P-M5-tree models are presented in Fig. 9. As can be inferred from Fig. 9, the models were relatively similar to each other. However,

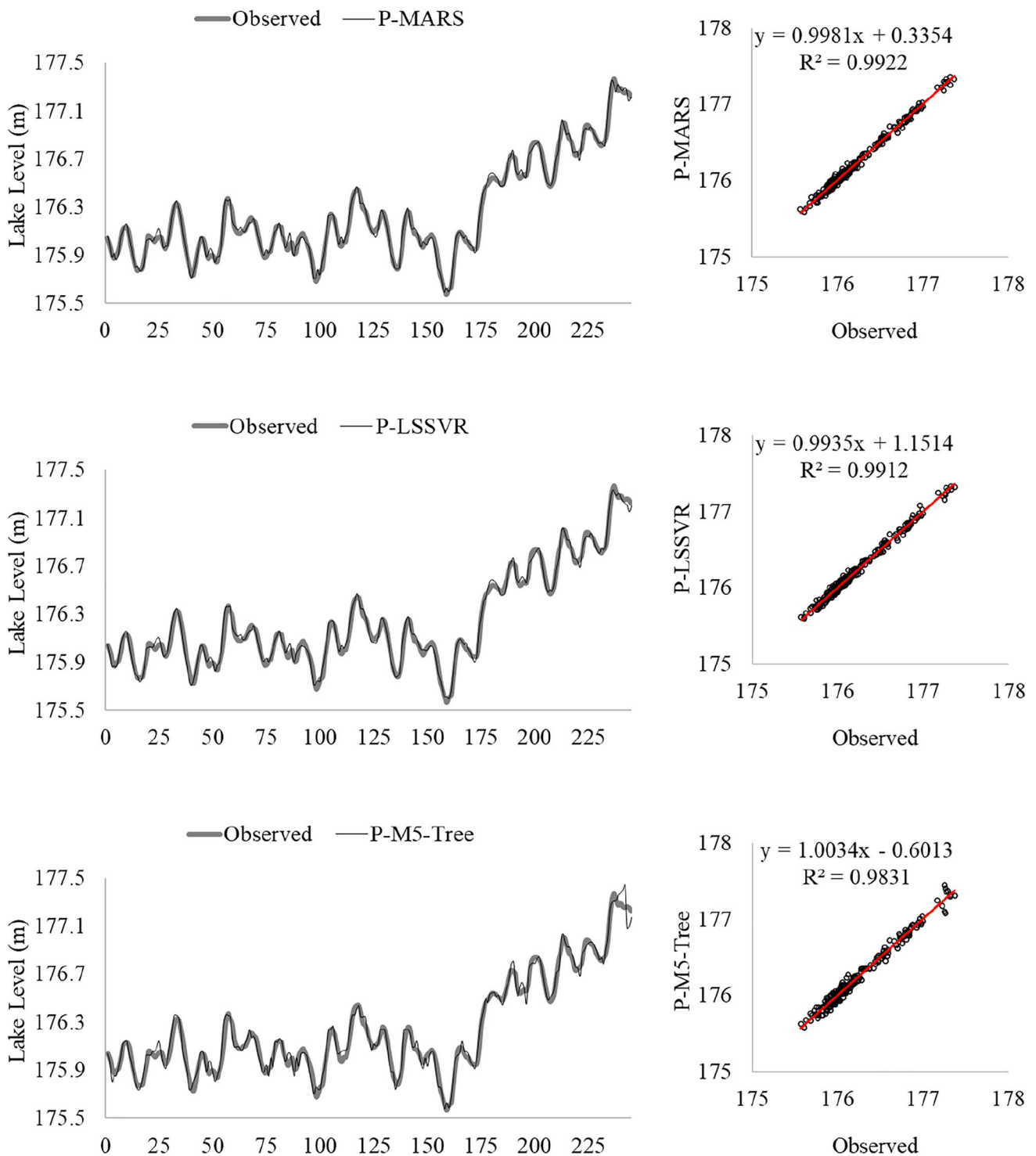
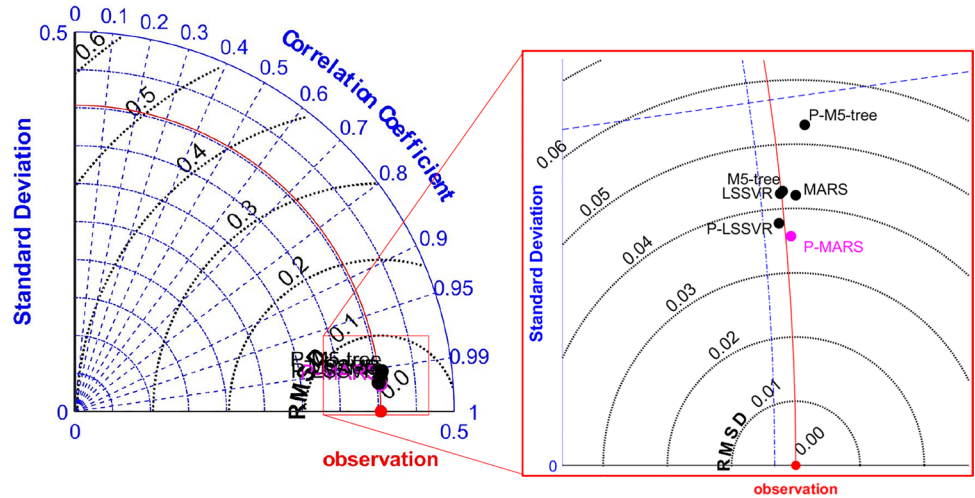


Fig. 8 Observed and forecasted lake level timeseries and scatter plots for the periodic test phase

P-MARS was separable from the other approaches at the Taylor diagram. P-MARS model yielded quite close to observed data. Therefore, the Taylor diagram also revealed that the P-MARS approach was more successful than the other models.

Violin plot shows the compatibility of the forecast data with the observed data with the help of statistical parameters. An additional comparison was made using the Violin plot for the models (Fig. 10). The Violin plot used in this study was modified from Legouhy (2021). The model

Fig. 9 Taylor diagrams of MARS, LSSVR and M5-tree models for testing phase



results are expressed in the original part (before) at the first stage, and then the results using the periodic component are given in the after part. With the addition of the periodic component, the performance of the MARS and LSSVR methods increased, but the performance of the M5-tree method decreased. This situation is understood by the similarity of the observed violin graph and the graphs of the other methods.

Finally, in this study, a statistical significance test was performed between the results of the best methods and the observed data. The Kruskal–Wallis (KW) test was used to determine whether the estimated and measured data had similar distributions (Citakoglu 2021; Görkemli et al. 2022). As seen in Table 4, the H_0 hypothesis is rejected in

the estimations of the methods of lake level fluctuations. In other words, it shows that there is no significant difference between the means of the forecasted and observed data. The KW test was performed at 95% of the confidence interval, and the critical value was $p_{cri} = 0.05$.

5 Conclusion

In this study, the applicability of MARS, M5-tree, and LSSVR in forecasting lake level fluctuations was investigated. Lake level observations from Lake Michigan in the USA were used for training and testing the three models. In terms of performance indices, the results demonstrated

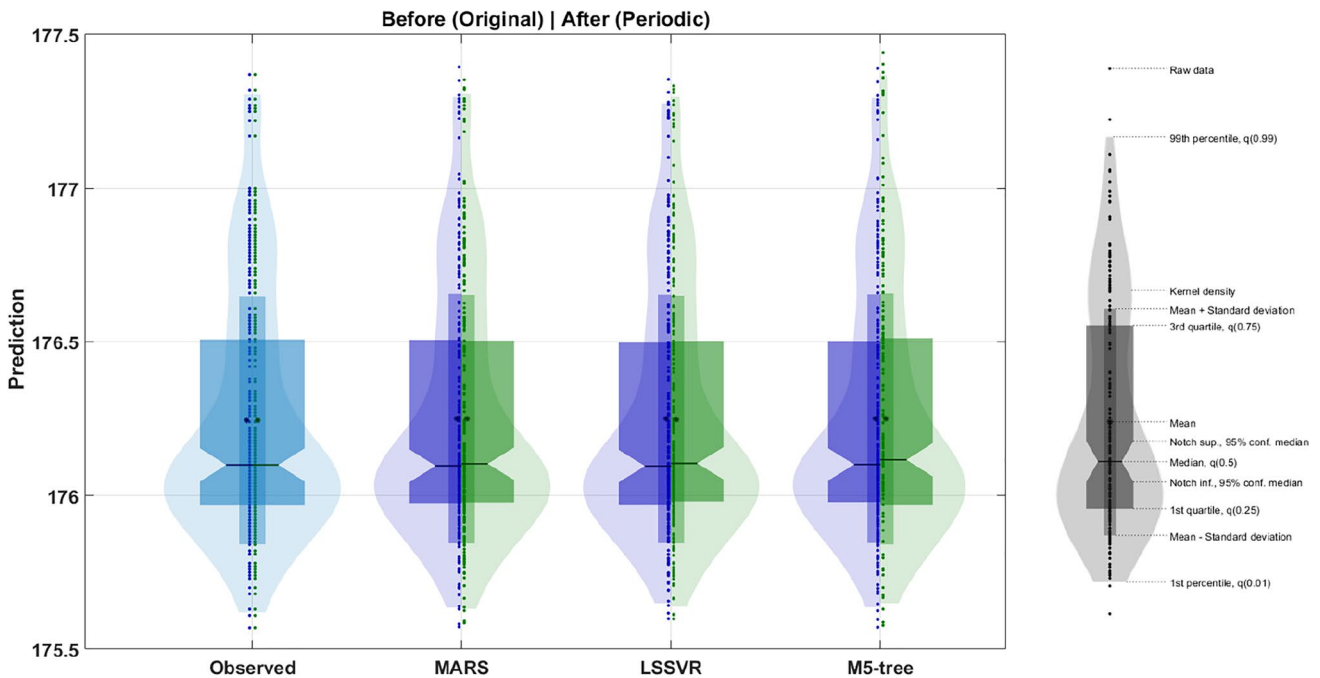


Fig. 10 Violin plot of MARS, LSSVR and M5-tree models for testing phase

Table 4 Kruskal–Wallis test results

| Models | Lake levels (m) | |
|-----------|-----------------|--------|
| | P value | Ho* |
| MARS | 0.8173 | Reject |
| LSSVR | 0.8579 | Reject |
| M5-tree | 0.8003 | Reject |
| P-MARS | 0.8937 | Reject |
| P-LSSVR | 0.9002 | Reject |
| P-M5-tree | 0.9164 | Reject |

*Ho means that there are differences between the mean estimate and measurement values

the effectiveness of the three models in reproducing the nonlinear behavior of lake level fluctuation.

- In both models with and without the periodic component introduced as input data, MARS performed slightly better than the LSSVR and M5-Tree model tree.
- In general, P-MARS indicated better forecast accuracies at input combinations VIII, mainly due to the capability of the application of multivariate adaptive regression, which could capture the complicated nonlinear relationship.
- Modeling using a single input (I) yielded the worst result in the estimation with M5-tree.
- The periodic component feature was embedded and evaluated inside the modeling's input datasets, and the results revealed that integrating this component data was 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 useful practical options for improved monthly lake level forecasts. In operational water level forecasting, the proposed methods could be valuable supplements to physically-based models. Understanding the causes of water level fluctuations and the factors influencing them can help in lake conservation and management. It is 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 the Great Lakes under the pressure of unpredictable climate variables.

Acknowledgements The author would like to thank KTO Karatay University.

Author contribution All chapters have been prepared by Vahdettin Demir.

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

Code availability Not applicable.

Declarations

Ethics approval and consent to participate 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 for publication Written informed consent for publication was obtained from all authors.

Conflict of interest The authors declare no competing interests.

References

- Adamowski J, Chan HF, Prasher SO, Sharda VN (2012) 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. <https://doi.org/10.2166/hydro.2011.044>
- Altunkaynak A (2007) Forecasting surface water level fluctuations of lake van by artificial neural networks. *Water Resour Manag* 21:399–408. <https://doi.org/10.1007/s11269-006-9022-6>
- Altunkaynak A, Özger M, Sen Z (2003) Triple diagram model of level fluctuations in Lake Van, Turkey. *Hydrol Earth Syst Sci* 7:235–244. <https://doi.org/10.5194/hess-7-235-2003>
- Altunkaynak A, Şen Z (2007) Fuzzy logic model of lake water level fluctuations in Lake Van, Turkey. *Theor Appl Climatol* 90:227–233. <https://doi.org/10.1007/s00704-006-0267-z>
- Başakın EE, Ekmekcioğlu Ö, Çitakoğlu H, Özger M (2022) A new insight to the wind speed forecasting: robust multi-stage ensemble soft computing approach based on pre-processing uncertainty assessment. *Neural Comput Appl* 34:783–812. <https://doi.org/10.1007/s00521-021-06424-6>
- Bera, S. O. Prasher, R. M. Patel, et al (2006) Application of MARS in simulating pesticide concentrations in soil. *Trans ASABE* 49:297–307. <https://doi.org/10.13031/2013.20228>
- Bhattacharya B, Solomatine DP (2005) Neural networks and M5 model trees in modelling water level-discharge relationship. *Neurocomputing* 63:381–396. <https://doi.org/10.1016/j.neucom.2004.04.016>
- Citakoglu H (2021) Comparison of multiple learning artificial intelligence models for estimation of long-term monthly temperatures in Turkey. *Arab J Geosci* 14:2131. <https://doi.org/10.1007/s12517-021-08484-3>
- Coulibaly P (2010) Reservoir computing approach to Great Lakes water level forecasting. *J Hydrol* 381:76–88. <https://doi.org/10.1016/j.jhydrol.2009.11.027>
- De Andrés J, Lorca P, De Cos Juez FJ, Sánchez-Lasheras F (2011) Bankruptcy forecasting: a hybrid approach using fuzzy c-means clustering and multivariate adaptive regression splines (MARS). *Expert Syst Appl* 38:1866–1875. <https://doi.org/10.1016/j.eswa.2010.07.117>
- Demir V, Çubukçu EA (2021) Digital elevation modeling with heuristic regression techniques. *Eur J Sci Technol* 484–488. <https://doi.org/10.31590/ejosat.916012>

- Deng S, Yeh TH (2010) Applying least squares support vector machines to the airframe wing-box structural design cost estimation. *Expert Syst Appl* 37:8417–8423. <https://doi.org/10.1016/j.eswa.2010.05.038>
- Friedman JH (1991) Multivariate adaptive regression splines. *Ann Stat* 19:590–606. <https://doi.org/10.1214/aos/1176347963>
- Görkemli B, Citakoglu H, Haktanir T, Karaboga D (2022) A new method based on artificial bee colony programming for the regional standardized intensity–duration–frequency relationship. *Arab J Geosci* 15. <https://doi.org/10.1007/s12517-021-09377-1>
- Goyal MK (2014) Modeling of sediment yield prediction using M5 model tree algorithm and wavelet regression. *Water Resour Manag* 28:1991–2003. <https://doi.org/10.1007/s11269-014-0590-6>
- Hemmati-Sarapardeh A, Shokrollahi A, Tatar A et al (2014) Reservoir oil viscosity determination using a rigorous approach. *Fuel* 116:39–48. <https://doi.org/10.1016/j.fuel.2013.07.072>
- Huang Z, Luo J, Li X, Zhou Y (2009) 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. IEEE, pp 4058–4061
- Hwang SH, Ham DH, Kim JH (2012) Forecasting performance of LS-SVM for nonlinear hydrological time series. *KSCE J Civ Eng* 16:870–882. <https://doi.org/10.1007/s12205-012-1519-3>
- Ji Z, Wang B, Deng S, You Z (2014) Predicting dynamic deformation of retaining structure by LSSVR-based time series method. *Neurocomputing* 137:165–172. <https://doi.org/10.1016/j.neucom.2013.03.073>
- Kamari A, Nikookar M, Sahranavard L, Mohammadi AH (2014) Efficient screening of enhanced oil recovery methods and predictive economic analysis. *Neural Comput Appl* 25:815–824. <https://doi.org/10.1007/s00521-014-1553-9>
- Karimi S, Shiri J, Kisi O, Makarynsky O (2012) 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. <https://doi.org/10.1260/1759-3131.3.2.109>
- Kisi O (2015) Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *J Hydrol* 528:312–320. <https://doi.org/10.1016/j.jhydrol.2015.06.052>
- Kisi O (2015) 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. <https://doi.org/10.1007/s11269-015-1107-7>
- Kisi O (2012) Modeling discharge–suspended sediment relationship using least square support vector machine. *J Hydrol* 456–457:110–120. <https://doi.org/10.1016/j.jhydrol.2012.06.019>
- Kisi O (2013) Least squares support vector machine for modeling daily reference evapotranspiration. *Irrig Sci* 31:611–619. <https://doi.org/10.1007/s00271-012-0336-2>
- Kişî Ö (2009) Neural network and wavelet conjunction model for modelling monthly level fluctuations in Turkey. *Hydrol Process* 23:2081–2092. <https://doi.org/10.1002/hyp.7340>
- Kisi O, Parmar KS (2016) 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. <https://doi.org/10.1016/j.jhydrol.2015.12.014>
- Kisi O, Parmar KS, Soni K, Demir V (2017a) Modeling of air pollutants using least square support vector regression, multivariate adaptive regression spline, and M5 model tree models. *Air Qual Atmos Heal* 10. <https://doi.org/10.1007/s11869-017-0477-9>
- Kisi O, Shiri J, Demir V (2017b) Hydrological Time Series Forecasting Using Three Different Heuristic Regression Techniques. In: *Handbook of Neural Computation*. Elsevier, pp 45–65
- Leathwick JR, Elith J, Hastie T (2006) Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecol Modell* 199:188–196. <https://doi.org/10.1016/j.ecolmodel.2006.05.022>
- Legates DR, McCabe GJ (1999) Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. *Water Resour Res* 35:233–241. <https://doi.org/10.1029/1998WR900018>
- Legouhy A (2021) al_goodplot - boxplot & violin plot. In: *MATLAB Cent. mathworks*. https://www.mathworks.com/matlabcentral/fileexchange/91790-al_goodplot-boxplot-violin-plot
- Liang C, Li H, Lei M, Du Q (2018) Dongting Lake water level forecast and its relationship with the Three Gorges Dam based on a long short-term memory network. *Water (switzerland)* 10. <https://doi.org/10.3390/w10101389>
- Luo Y, Dong Z, Liu Y (2021) 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. <https://doi.org/10.1007/s00477-021-01974-6>
- Mitchell TM (1997) *Machine learning*. The McGraw-Hill Companies, New York
- Michigan (2021a) Sea Grant. <https://www.michiganseagrant.org/topics/great-lakes-fast-facts/lake-michigan/>. Accessed 15 Jul 2021
- Michigan (2021b). <https://www.goller.gen.tr/michigan-golu.html>. Accessed 15 Jul 2021b
- Okkan U, Ali Serbes Z (2013) The combined use of wavelet transform and black box models in reservoir inflow modeling. *J Hydrol Hydromechanics* 61:112–119. <https://doi.org/10.2478/johh-2013-0015>
- Pahasa J, Ngamroo I (2011) A heuristic training-based least squares support vector machines for power system stabilization by SMES. *Expert Syst Appl* 38:13987–13993. <https://doi.org/10.1016/j.eswa.2011.04.206>
- Pal M, Deswal S (2009) M5 model tree based modelling of reference evapotranspiration. *Hydrol Process* 23:1437–1443. <https://doi.org/10.1002/hyp.7266>
- Peprah MS, Larbi EK, Peprah MS (2021) 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
- Quinlan JR (1992) Learning with continuous classes. *World Sci* 92:343–348. <https://doi.org/10.1.1.34.885>
- Sanikhani H, Kisi O, Kiafar H (2015) 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. <https://doi.org/10.1007/s11269-014-0894-6>
- Şen Z, Kadioğlu M, Batur E (2000) Stochastic modeling of the Van Lake monthly level fluctuations in Turkey. *Theor Appl Climatol* 65:99–110. <https://doi.org/10.1007/s007040050007>
- Sephton P (2001) Forecasting recessions: can we do better on MARS? *Review* 83:39–49
- Shafaei M, Kisi O (2016) Lake level forecasting using wavelet-SVR, wavelet-ANFIS and wavelet-ARMA conjunction models. *Water Resour Manag* 30:79–97. <https://doi.org/10.1007/s11269-015-1147-z>
- Sharda VN, Patel RM, Prasher SO et al (2006) Modeling runoff from middle Himalayan watersheds employing artificial intelligence techniques. *Agric Water Manag* 83:233–242. <https://doi.org/10.1016/j.agwat.2006.01.003>
- Shiri J, Shamshirband S, Kisi O et al (2016) Prediction of water-level in the Urmia Lake using the extreme learning machine approach. *Water Resour Manag* 30:5217–5229. <https://doi.org/10.1007/s11269-016-1480-x>

- Shokrollahi A, Arabloo M, Gharagheizi F, Mohammadi AH (2013) Intelligent model for prediction of CO₂ — reservoir oil minimum miscibility pressure. *Fuel* 112:375–384. <https://doi.org/10.1016/j.fuel.2013.04.036>
- Shortridge JE, Guikema SD, Zaitchik BF (2015) Empirical streamflow simulation for water resource management in data-scarce seasonal watersheds. *Hydrol Earth Syst Sci Discuss* 12:11083–11127. <https://doi.org/10.5194/hessd-12-11083-2015>
- Solomatine DP, Dulal KN (2003) Model trees as an alternative to neural networks in rainfall-runoff modelling. *Hydrol Sci J* 48:399–411. <https://doi.org/10.1623/hysj.48.3.399.45291>
- Solomatine DP, Xue Y (2004) 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. [https://doi.org/10.1061/\(ASCE\)1084-0699\(2004\)9:6\(491\)](https://doi.org/10.1061/(ASCE)1084-0699(2004)9:6(491))
- Sotomayor KAL (2010) 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
- Suykens JAK, Vandewalle J (1999) No Title Neural Process Lett 9:293–300. <https://doi.org/10.1023/A:1018628609742>
- Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. *J Geophys Res Atmos* 106:7183–7192. <https://doi.org/10.1029/2000JD900719>
- Vuglinskiy V (2009) 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
- Yaseen ZM, Kisi O, Demir V (2016) Enhancing long-term streamflow forecasting and predicting using periodicity data component: application of artificial intelligence *Water Resour Manag* 30 <https://doi.org/10.1007/s11269-016-1408-5>
- Young CC, Liu WC, Hsieh WL (2015) Predicting the water level fluctuation in an Alpine Lake using physically based, artificial neural network, and time series forecasting models *Math Probl Eng* 2015 <https://doi.org/10.1155/2015/708204>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.