Water Resources Management (2026) 40:444
https://doi.org/10.1007/s11269-026-04806-w

Robust Monthly Lake Water Level Forecasting Using an Ensemble of Singular Spectrum Analysis Based Deep Learning and Categorical Boosting Models

Okan Mert Katipoğlu[1] · Veysi Kartal[2] · Türker Tuğrul[3] · Veysel Süleyman Yavuz[2] · Sema Ariman[4]

Received: 7 April 2025 / Accepted: 5 June 2026
© The Author(s) 2026

Abstract
This study developed an innovative hybrid modelling approach to predict one-month-ahead water levels (LWL (T+1)) for two major lakes in Türkiye (Eber and Eğirdir), aiming to model complex dynamics of hydrological systems and optimize sustainable water resources management. In the proposed methodological framework, Singular Spectrum Analysis (SSA) was employed to decompose hydrometeorological time series into subcomponents, while Mutual Information (MI) and Partial Autocorrelation Function (PACF) techniques were utilized to determine optimal input variables. The hybrid model architecture, incorporating Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and CatBoost algorithms, was optimized using 70% of the data for training and 30% for testing, and its performance was evaluated through deterministic and probabilistic metrics. The analyses revealed that SSA-based data preprocessing significantly improved the performance of prediction models, and depending on lake characteristics, CNN-based models demonstrated superior performance for Lake Eber, while DNN architecture excelled for Lake Eğirdir. The developed hybrid approach presents an original methodology that can be integrated into decision support systems for water resources management, contributing to the development of sustainable water management strategies and adaptability to similar hydrological systems.

Keywords Hybrid Deep Learning · Water Level Prediction · Data Decomposition · Hydrological Forecasting · Feature Selection · Time Series Analysis

1 Introduction

Accurate lake water level forecasting is essential for the sustainable management of water resources, especially in light of the increasing threats posed by climate change, extreme weather events, and hydrological variability (Fijani and Khosravi 2023; Hai et al. 2019). Various techniques, including wavelet-based models, season-neuro approaches, and hybrid

Extended author information available on the last page of the article

Published online: 26 June 2026

machine learning methods, have been employed to forecast lake water levels (Altunkaynak 2019). However, the performance of these models can be sensitive to the specific characteristics of the lake and the available data. (Yadav and Eliza 2017). Predictive models for lake water levels provide crucial insights that can mitigate the risks associated with flooding, drought, and ecosystem disturbances (Chen et al. 2023). Traditional statistical methods such as autoregressive integrated moving average (ARIMA) and linear regression have been used in hydrological forecasting as well as lake water level modelling (Aksoy et al. 2013). Still, they are often limited by the inability to capture the nonlinear, dynamic nature of hydrological data, particularly in large, complex water systems (Zhang et al. 2011; Vionnet et al. 2020). This has driven a shift toward advanced machine learning (ML) and deep learning (DL) techniques, which offer substantial improvements in prediction accuracy by leveraging large datasets, diverse input features, and flexible model architectures (Ruma et al. 2023; Bonakdari et al. 2019; Ahmed et al. 2022).

A prominent technique in this area is Singular Spectrum Analysis (SSA), which decomposes time series data into trends, oscillatory components, and noise, making it ideal for extracting key patterns and seasonal fluctuations from complex datasets (Nguyen et al. 2022; Ghafarian Malamiri et al. 2018). SSA has been applied successfully in water level forecasting models, allowing for improved performance by isolating the most relevant features from raw data and reducing noise, thus enhancing subsequent model predictions (Mohammed et al. 2023; Liao et al. 2014). By breaking down intricate hydrological data into managable components, SSA has demonstrated significant benefits when integrated with other models like artificial neural networks (ANNs) and support vector regression (SVR) (Polomcic et al. 2017; Jamei et al. 2022). This combination has proven effective in capturing seasonal variations and trends within water data, ultimately contributing to more robust and adaptable forecasting models (Reddy et al. 2022).

Both machine learning methods and traditional statistical forecasting techniques have demonstrated successful performance in the forward-looking prediction of numerous parameters. Although these are statistically highly effective methods, the increasing use of advanced deep learning techniques and hybrid models has led to more satisfactory model performance results in the literature, successfully capturing the attention of researchers (Ergüven et al. 2025).

Deep learning approaches, particularly Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Convolutional Neural Networks (CNN), have made transformative contributions to water level forecasting due to their ability to capture complex temporal dependencies and handle nonlinearities within sequential data (Atashi et al. 2022; Lin et al. 2022; Cho et al. 2022). For example, LSTM and GRU models, well-suited for long-term dependencies, and they have demonstrated remarkable accuracy in flood prediction, river discharge, and groundwater levels (Kogekar et al. 2021). Furthermore, hybrid models that combine CNN with LSTM or GRU have shown improved performance by learning both spatial and temporal features, making them particularly advantageous in complex water systems where data exhibit spatial dependencies (Barzegar et al. 2021). Recently, the LSTM algorithm has been utilized by researchers in various structures to predict future meteorological and hydrological parameters. These include hybridizations with CNNs and other machine learning methods. In fact, the most critical criterion is the hybrid model structure developed specifically for the data being predicted and the relevant region, as the performance of these models can vary accordingly. In this context, the hybridization of CNN with LSTM has

emerged as one of the most focused and effective hybrid models in recent days (Natarajan et al. 2026). Irfan et al. (2026), who advocate that leaving a better future for coming generations is possible through the conservation of water resources, proposed several models to facilitate future predictions of Total Water Consumption (TWC) and Agricultural Water Use (AWU). In their study, they utilized popular GRU-based hybrid models, similar to LSTM. They also stated that they achieved highly effective performance with their models. Although the performance of hybrid models depends on hyperparameters, the model input structure also significantly affects performance. It has been noted in the literature that the performance of new deep learning algorithms may vary accordingly, and it is emphasized that feature selection can be performed using tree-based algorithms (Tandiverdi and Batmaz 2025). These models also benefit from SSA preprocessing, which refines data and allows deep-learning models to focus on significant patterns, thereby minimizing forecast errors (Ahmed et al. 2022).

Categorical boosting algorithms, particularly CatBoost, represent a significant advancement in ensemble learning by efficiently handling categorical features and delivering robust predictions in diverse applications, including water forecasting (Qian et al. 2023; Dasi et al. 2023; Wei et al. 2023). CatBoost's capability to manage high-dimensional data, coupled with its inherent ability to handle multicollinearity issues, is especially advantageous in hydrological applications where data is often complex and subject to numerous environmental influences (Kim et al. 2023; Sapitang et al. 2020). The integration of CatBoost with other machine learning models, such as XGBoost and LightGBM, has demonstrated enhanced predictive accuracy in various hydrological and environmental contexts, showing promise in capturing the intricate relationships between different variables (Kim et al. 2023; Fan et al. 2023). The CatBoost algorithm can be utilized as an ensemble method as well as directly for parameter estimation. The application of these algorithms in hydrology, hybrid methods, not only enhances model performance but also lays the groundwork for the development of various hybrid methods. It is particularly known for demonstrating effective performance in drought prediction within semi-arid regions (Tuğrul et al. 2025). In this scope, studies applying CatBoost with neural networks in hybrid models have shown improvements in predictive reliability, especially under dynamic environmental conditions where traditional models fail (Louis et al. 2019; Herath et al. 2023).

Ensemble learning, a broader category of machine learning techniques, combines multiple models to achieve greater accuracy and robustness, a method increasingly applied to complex environmental datasets (Demir and Şahin 2023; Mahajan et al. 2023). Ensemble frameworks, especially those that incorporate decision trees, gradient boosting, and categorical boosting, have been shown to outperform single models by capturing a diverse range of patterns in data (Qian et al. 2023; Ahn et al. 2023). CatBoost, as part of such an ensemble framework, complements deep learning and SSA, making it a highly effective approach for forecasting water levels in complex lake systems (Kaur and Sood 2020; Barzegar et al. 2021; Nguyen et al. 2022).

This study was built based upon these advancements by proposing a novel forecasting framework that integrates SSA, deep learning architectures (LSTM, GRU, CNN), and CatBoost for monthly lake water level forecasting. The proposed model leverages SSA to decompose time series data, enhancing the accuracy of subsequent deep learning models in capturing complex temporal dynamics, while CatBoost provides resilience by managing categorical and high-dimensional environmental data (Polomcic et al. 2017; Prokhorenkova et al. 2018). By integrating these techniques, this study aims to create a robust and adaptable

model that can be applied across varying environmental conditions, providing a powerful tool for water resource management and environmental monitoring in an era marked by rapid climate variability.

2 Material and Method

2.1 Study Area and Dateset

Türkiye is characterized by four distinct seasons and is primarily situated within a semi-arid climatic zone. The inland regions, in particular, face significant challenges due to high temperatures and limited precipitation. Consequently, the strategic planning and management of water resources are of paramount importance for the country. In this context, this study focuses on Lake Eber and Lake Eğirdir as the primary research areas. Both regions exhibit a rigorous climate, characterized by an average annual temperature fluctuation of approximately 20 °C and annual precipitation levels aobut 650 mm. The combination of low precipitation and highly variable temperatures reflects the harsh climatic conditions of these basins.

Egirdir Lake, located in the Isparta province, is Türkiye's second-largest freshwater lake. Situated in the Lakes Region in the western Mediterranean, it is a vital freshwater source with a complex karstic and tectonic structure. The lake's formation resulted from the merging of depression cavities within primary limestone, and its basin was further shaped by tectonic processes. Classified as a pluvial lake in the context of Anatolia's ancient climate, Eğirdir is replenished by subterranean springs, nearby streams, and direct precipitation (Atılgan et al. 2020; Yücel et al. 2022).

Lake Eber, located in the Bolvadin district of Afyon, is a closed basin within the Akarçay Basin. Unlike Eğirdir, Lake Eber is primarily replenished by direct precipitation and surface runoff, and its surface is notably covered with dense aquatic vegetation (Sener et al. 2010).

For this study, a comprehensive dataset consisting of meteorological and hydrological parameters from both lakes was utilized. The variables include maximum temperature (Tmax), minimum temperature (Tmin), average temperature (Tave), rainfall, relative humidity (RH), sunshine duration (SH), wind speed (WS), and lake water levels. The statistical characteristics of these datasets are summarized in Table 1. In this study, monthly average lake water level (LWL, m) data obtained from the General Directorate of State Water Works (DSI) were used. The dataset covers the period between 1990 and 2022 and represents a 33-year monthly time series, as shown in Fig. 1. The data were collected through the regular hydrological observation system of DSI.

The monthly LWL data were used as the main input dataset in the applied analysis and modeling procedures. First, the time-series data were organized and checked before the analysis process. Then, these data were utilized in the selected techniques to evaluate temporal variations and obtain the desired prediction and performance results.

Candidate input variables for one-month-ahead lake water level forecasting [LWL(T + 1)] included the contemporaneous and lagged lake water level values and monthly meteorological variables. In this study, LWL(T + 1) denotes the target lake water level for the next month; LWL(T), LWL(T − 1), and LWL(T − 2) denote the current, one-month-lagged, and two-month-lagged lake water levels, respectively. The meteorological variables considered as candidate predictors were RH, Tmin, Tmax, Tave, WS, rainfall, and SH.

Table 1 Details of hydrometerorological variables

Eber						
Location	31°10'55" E-38°37'7" N			Drainage area (km^2)		5288
Variable	initial	end	mean	std	min	max
Tmax ºC	1990	2022	24.97	8.26	4.40	39.00
Tmin ºC			-0.74	8.67	-23.70	14.40
Tave ºC			11.64	8.15	-5.64	26.13
RH %			62.16	11.67	30.82	89.43
WS (m/s)			2.03	0.47	0.45	3.06
Rainfall (mm)			32.31	25.18	0.00	136.60
SH (hours)			192.98	88.08	20.20	409.60
Lake level (m)			965.92	0.56	964.55	967.89
Egirdir						
Location	30°51'16" E-37°52'33" N			Drainage area (km^2)		3321
Variable	initial	end	mean	std	min	max
Tmax ºC	1990	2022	23.99	8.31	4.00	38.60
Tmin ºC			1.55	7.48	-14.90	14.80
Tave ºC			12.42	7.69	-3.71	25.63
RH %			66.96	9.91	41.30	91.30
WS (m/s)			3.03	0.56	0.17	4.94
Rainfall (mm)			66.68	66.44	0.00	354.80
SH (hours)			216.88	84.16	6.90	396.20
Lake level (m)			917.10	0.58	915.82	918.48

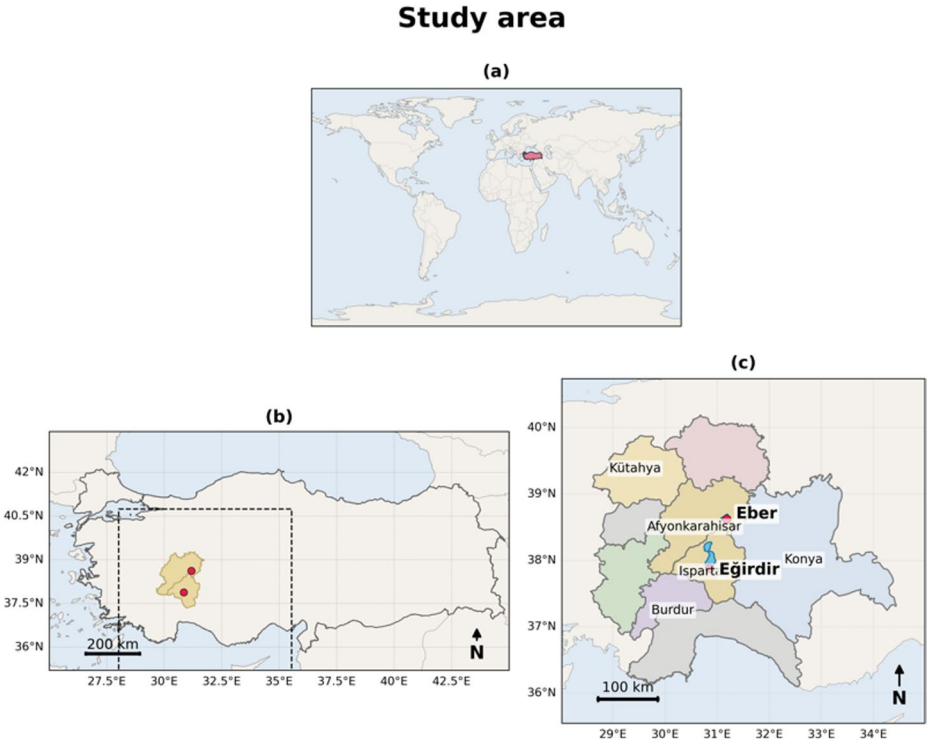


Fig. 1 Study area

3 Methods

The methodological framework was developed specifically for hydrological forecasting of lake water levels. In this context, the applied techniques, including Singular Spectrum Analysis (SSA), Mutual Information (MI), Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), CatBoost, and Deep Neural Networks (DNN), were integrated into the prediction process of lake water level variations. Initially, SSA was employed to decompose the original lake water level time series into meaningful temporal components and reduce noise effects. Subsequently, MI analysis was used to determine the most informative lagged lake water level inputs associated with future water level changes. The processed hydrological inputs were then utilized in RNN, CNN, CatBoost, and DNN models to estimate and predict future lake water level fluctuations. Thus, all techniques presented in Sect. 3.1–3.6 were directly implemented for hydrological analysis and forecasting purposes.

3.1 Singular Spectrum Analysis (SSA)

One of the main factors that makes forecasting difficult in time series is the variable statistical structure. Due to structure difficulties, non-parametric methods are generally preferred to facilitate analyzes in time series and increase the accuracy of forecasts (Pham et al. 2021). One of them is Singular Spectrum Analysis (SSA), which is a spectral estimation technique. SSA extracts the time series interpretable components. A primary characteristic of SSA is its ability to decompose the time series into more comprehensible components, making it more accurate for prediction. Each component may possess a distinct structure based on its features(Sun et al. 2019; Zhang et al. 2011).

The SSA technique takes place in two stages: (1) embedding and singular value decomposition (SVD); (2) grouping and cross averaging in the reconstruction.

3.1.1 Stage decomposition

Embedding Embedding can be called mapping that converts a one-dimensional time series $(Y_T = y_1, \ldots, y_T)$ into a multi-dimensional matrix $X_i, \ldots, X_K$ with $(X_i = y_i, \ldots, y_{i+L-1})$, that is a vector (where $K = T - L + 1$). X_i vectors are also called L lag vectors. The first parameter here is the embedded window length L $(2 \leq L < T)$. The trajectory obtained here creates the matrix. The trajectory matrix $X = [X_1, \ldots, X_k] = (X_{ij})_{I,j=1}^{L,K}$. Here, X is the diagonal matrix and $i + j$ are constant.

Singular Value Decomposition (SVD) At this stage, SVD decomposes the trajectory matrix into a singular value. Here denote the eigenvalues of XX' by $\lambda_1, \ldots, \lambda_L$ and $U_1, \ldots, U_L$ are the corresponding eigen values of the eigenvectors of the orthogonal system XX' $((U_i, U_j) = 0\ i \neq j)$.

$$d = \max(i,\ such\ that\ \lambda_i > 0) = rank\ X$$

Let $V_i = X' U_i / \sqrt{\lambda_i}$, then the SVD can be addressing as:

$$X = X_1 + \ldots + X_d \quad (1)$$

where V_i is the principal components, X_1 with rank 1, U_i is factor empirical orthogonal functions. Here $\sqrt{\lambda_i}$, U_i, V_i is the matrix i_{th} are their eigenvalues and also $\sqrt{\lambda_i}$ is named as the spectrum of the matrix X. If $\sqrt{\lambda_i}$, U_i, V_i here are equal to 1, the equation $\sum_{i=1}^{r} \lambda_i / \sum_{i=1}^{d} \lambda_i$ is accepted (r *and* d are the rank values based on X_1.

3.1.2 Stage Reconstruction

Grouping The first step in this stage is to divide the matrix X_1 into several the basic matrices and summing the matrices within each group. Let $I = \{i_1, \ldots .i_p\}$ be a group of indices. Next, the matrix X_I corresponding to I takes the form $X_I = X_{i1} + \ldots . + X_{p1}$. The separation of index sets $J = 1, \ldots, d$ corresponds to the subset $I_1, \ldots ., I_m$ and finally represent as $X = X_{I1} + \ldots . + X_{Im}$. In group I, the contribution of component X_I to the expansion is quantified by the proportion of the relevant eigenvalues: $\sum_{i \in I} \lambda_i / \sum_{i=1}^{d} \lambda_i$.

Diagonal Averaging The Diagonal averaging process, called the diagonal averaging, is transferred to an I series with the initial series Y_T. If z_{ij} is an element of a Z matrix, the k_{th} one term of the time series is continued by averaging the z_{ij}. The Z matrix obtained is the Hankel matrix which is the trajectory matrix and correspond to the times series obtained from the diagonal matrix. Here, we obtain the matrix $\mathcal{H}Z$ which is the nearest to Z matrix and another expansion: $X = \widetilde{X}_{I1} + \ldots . + \widetilde{X}_{Im}$. Here $\mathcal{H}Z = \widetilde{X}_{I1}$. This corresponds to the decomposition of $Y_T = (y_1, \ldots, y_T)$ into a summation of m series: $y_t = \sum_{k=1}^{m} \widetilde{y}_t^{(k)}$ 00 which corresponds to X_{Ik}. More information for calculation see: Sun et al. (2019) and Hassani (2007).

3.2 Recurrent Neural Networks (RNN)

Artificial Neural Network (ANN) work biologically by imitating human neural networks and is widely used in hydrological studies (Katipoğlu et al. 2023; Köyceğiz and Büyükyıldız 2022; Tuğrul and Hinis 2024). According to many scientists, it was said that ANN and various modifications were pretty good for forecasting hydrologic parameters such as precipitation and streamflow. There are multiple modifications of ANN such as feedforward, feedback, recurrent and multilayer to be effective in many different situations and parameters. Recurrent Neural Network (RNN) is an ANN-based machine learning method. Although it has many similarities with ANN, it differs fundamentally on one point While the previous situation does not affect the result output in the known ANN, it affects the result output in the RNN, leading to a more effective outcome (Le et al. 2017). In time series predictions in hydrology, dynamic input data can be defined as a function of time. When using this function, there may be difficulties in solving or predicting the problem due to the need for many input parameters based on the function. To overcome the difficulties in such cases, RNN has been introduced (Elman 1993; Hopfield 1982). Compared to traditional ANN, RNN has a dynamic structure for solving problems, making it more suitable for estimating target parameters (Hammer 2000). The hidden state consists of units known as input, output,

and forget gates, which enable evaluation and control. Detailed formulas related to RNN's structure are shown in the supplementary material. (the supplementary material: https://doi.org/10.1016/j.jhydrol.2023.129924)

3.3 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN), one of the deep learning methods, is used in many studies to estimate hydrological parameters, and researchers state that it provides effective results (Danandeh Mehr et al. 2023). CNN yields superior outcomes both alone and when integrated with various machine learning or optimization techniques. The structure of CNN is very similar to that of traditional ANN. It may be seen as almost equivalent to Multi-Layer Perception (MLP), which is a variant of ANN. Just like in MLP, the network structure consists of multiple layers. The main advantage of this method is that the features in the input data are used in these networks, making it easier to calculate and reliable (Fig. 2). A CNN fundamentally comprises a convolutional layer, a pooling layer, and a fully linked layer. The main module of the CNN is comprised of the convolutional and pooling layers, which are coupled to one another. The task of the convolution layer is to distinguish and process features. A convolution kernel does a dot product operation inside a localized region of the input data or the output from the preceding layer to extract sophisticated features. The pool layer is located after the convolution layer and divides the samples into smaller groups. This plays an effective role in achieving better results. The last layer, the fully connected layer, performs the same tasks as in an MLP. The preceding layers can be regarded as pre-processing, which enhances the estimation of the resultant parameters. The completely linked layer consolidates all activations from preceding levels and produces the final output.

According to the Eq. 2, the CNN is computed.

$$y^n\,(i,w) = \left(c_w^n\; x\; d\,\right)(i) = \sum {}_{j=-\infty}^{\infty}\, c_w^n\; x\,(i-j) \tag{2}$$

where c_w^1 (w=l,…, M_l) shows filter, first layer indicates $f^1 = w\left(z^1\right) \in R$ for activation function, n represents hidden layer numbers, $d\,and\,y$ are input and output valuable, respectively. More details can see: Hancock and Khoshgoftaar (2020), Danandeh Mehr et al. (2023).

3.4 Categorical Boosting (CatBoost)

CatBoost, designed to enhance the efficacy of decision trees, is available as open-source code (Prokhorenkova et al. 2018). CatBoost, characterized by several qualities, is utilized

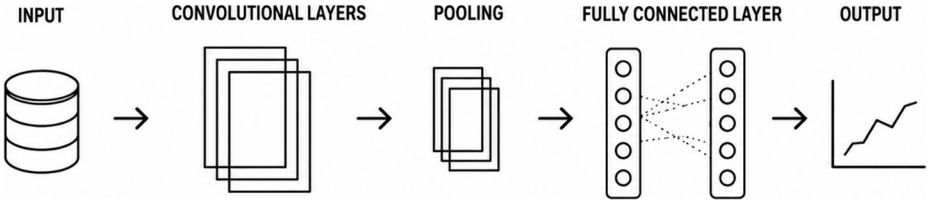


Fig. 2 Structure of the CNN

by Yandex researchers and engineers and is favoured by many hydrological researchers (Najm et al. 2023). Numerous experts assert that one of its paramount attributes is its excellent utilization of data characteristics without necessitating any pre-processing (Dorogush et al. 2018). Significant results may be achieved by either deleting missing data or using the data with CasBoost without requiring any correction (Ibragimov and Gusev 2019). The fact that this technique, in contrast to other ways, discourages overlearning is yet another significant benefit that this method has over traditional methods (Hancock and Khoshgoftaar 2020). CatBoost comprises entirely symmetrical trees that have two leaves and six levels of depth. Here, if the learning rate is to be determined automatically, it is recommended that this process be calculated according to the iteration and data properties in the data set. To enhance learning efficiency, one may decrease the number of iterations while simultaneously increasing the learning rate (Najm et al. 2023). Overall, these make it especially suitable and low cost for calculation in prediction parameters.

3.5 Mutual Information

Mutual information, shown as a stochastic measuring technique by several studies, is also used efficiently in hydrology (Steuer et al. 2002). This approach quantitatively indicates the relationship between two independent variables. For instance, a drought value that is desired to be calculated based on parameters such as precipitation, flow, and temperature is mathematically expressed as the value that has the greatest effect on drought. Here, a relative value is calculated by taking the difference between the degree of degradation of the system and the degree of degradation of the features. The feature with the highest entropy is the one most associated with the feature to be predicted. This approach is calculated as following:

$$H\left(X\right) = -\sum_{i=1}^{M_x} p\left(x_i\right) \log p\left(x_i\right) \tag{3}$$

$$H\left(X,Y\right) = -\sum_{i=1}^{M_x} \sum_{j=1}^{M_y} p\left(x_i, y_j\right) \log p\left(x_i, y_j\right) \tag{4}$$

where $p\left(x_i, y_j\right)$ shows the joint possibility which X is in state x_i and Y is in state y_j. M_x and M_y represent the number of potential variations. The mutual information $I\left(X, Y\right)$ between systems X and Y is described mathematically by Eq. 5. More details, see: Rossi et al. (2006).

$$I\left(X,Y\right) = H\left(X\right) + H\left(Y\right) - H\left(XY\right), \geq 0 \tag{5}$$

3.6 Deep Neural Network (DDN)

Deep Neural Network (DDN) is one of the most complex versions of artificial neural networks, and due to its complex structure, in fact, it is one of the successful algorithms according to many researchers because it is applied to the architecture of deep learning in the process of learning (Abiodun et al. 2018). One of the biggest reasons why this algorithm is successful is that it has two or more hidden layers, as in Fig. 3, and this structure is associated with deep learning (Sahiner et al. 2019; Wardah et al. 2019). The vertical three-dot symbols shown below the hidden layers are schematic ellipsis marks indicating that addi-

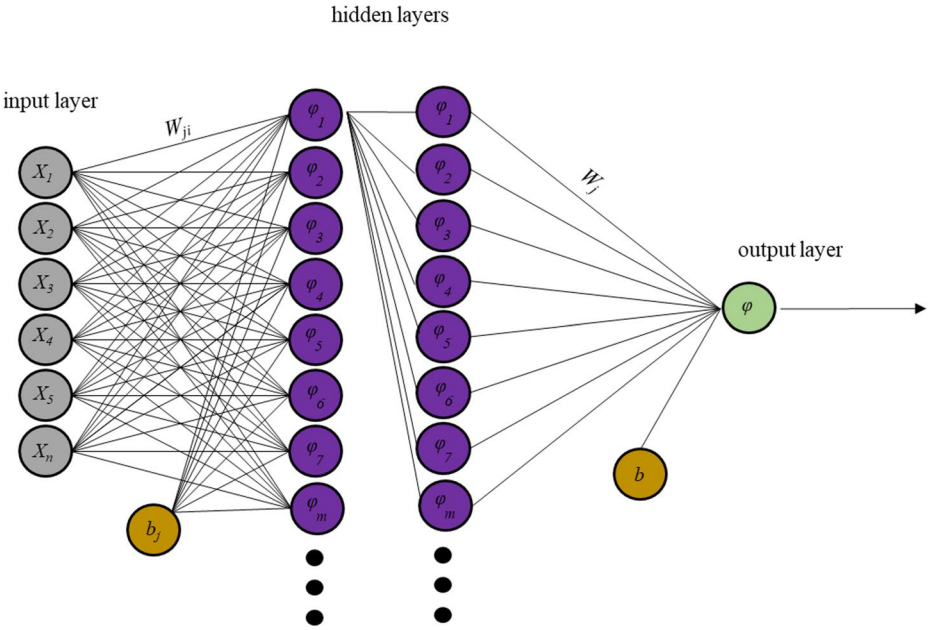


Fig. 3 Structure of the DNN (The vertical three-dot symbols shown below the hidden layers are schematic ellipsis marks indicating that additional hidden neurons are present but omitted from the illustration for simplicity)

tional hidden neurons are present but omitted from the illustration for simplicity. Despite its architectural similarities to MLP, the primary distinction is in the number of hidden layers; MLP has a single hidden layer, while DNN may have two, three, or more hidden layers. Thanks to this feature, it provides learning in the input structure consisting of many features and analyzes the contribution of each feature to the result in the most efficient way and good results are obtained. By analyzing the input and output structure efficiently, one may get highly effective outcomes by elucidating their links. Compared to ANN and MLP, learning can be achieved with a larger network due to the larger scale of the input data (Awad and Khanna 2015). Regardless of the extended duration of the training procedure relative to other approaches, the classic backpropagation strategy yields effective results based on the dataset (Masood and Ahmad 2021). The disadvantage here is that due to the backpropagation policy, it can be overfitting in the learning process. The same calculation methods used in MLP are applied here. Due to its favourable outcomes, the linear activation function was used in DNN for hydrological forecasting, among several options, including linear, logistic, and sigmoid functions. This approach is calculated in Eq. 6 for each neuron:

$$y = \left[\sum\nolimits_{j=1}^{m} w_j \left(\sum\nolimits_{i=1}^{N} w_{ji} + b_j\right) + b\right] \tag{6}$$

where m is number of hidden neurons, N is number of samples, w_{ji} is weight which connects the ith and jth neurons in the input layer and the hidden layer, respectively; b_j is bias for the jth hidden neuron. For more details, see Kim et al. (2018) and Zhang et al. (2020).

3.7 Performance Metrics

To measure the performance of the applied models several performance indicators were commonly used in the field of hydrology such as Root Mean Square Error (RMSE), the Mean Absolute Error (MAE), the Akaike information criterion (AIC), the Nash–Sutcliffe model efficiency (NSE), the Mean Bias Error (MBE) and the Percent Bias (Pbias).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - y_i)^2} \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|o_i - y_i| \tag{8}$$

$$NSE = 1 - \left[\frac{\sum_{i=1}^{n}(o_i - y_i)^2}{\sum_{i=1}^{n}\left(o_i - \bar{o}\right)^2}\right] \tag{9}$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(y_i - o_i) \tag{10}$$

$$Pbias = \left[\frac{\sum_{i=1}^{n}(y_i - o_i)}{\sum_{i=1}^{n}o_i}\right] x100 \tag{11}$$

where o_i: observed value, y_i: the predicted value, $\bar{o}$ is the mean observedc value; n: the number of sample.

The proposed framework was developed as a sequential integration of PACF, MI, SSA, and machine/deep learning algorithms for one-month-ahead lake water level forecasting in Fig. 4. PACF was first employed to determine the statistically significant lagged lake water level terms, and MI was then used to identify the most informative lagged and meteorological predictors for LWL(T+1). Based on these selected inputs, two modelling branches were established: MI-based baseline models using the raw predictors directly, and SSA-MI-based models using SSA-decomposed subcomponents of the same predictors. This framework enabled a clear assessment of the added value of SSA preprocessing in improving the predictive performance of CatBoost, CNN, DNN, and RNN models.

4 Results

In this study, a hybrid prediction model was proposed by integrating machine learning techniques such as Catboost, DNN, CNN, RNN with MI and SSA techniques for the prediction of the next month's water level values (LWL (T+1)) of 2 major lakes in Türkiye. The SSA approach was used for effective modeling by separating the data into sub-bands and the mutual information (MI) approach was used for feature selection. The importance of features according to the MI approach is shown in Fig. 5. Accordingly, the variables that have the highest MI scores on LWL (T+1) values in Eber Lake are LWL (T), LWL (T-1),

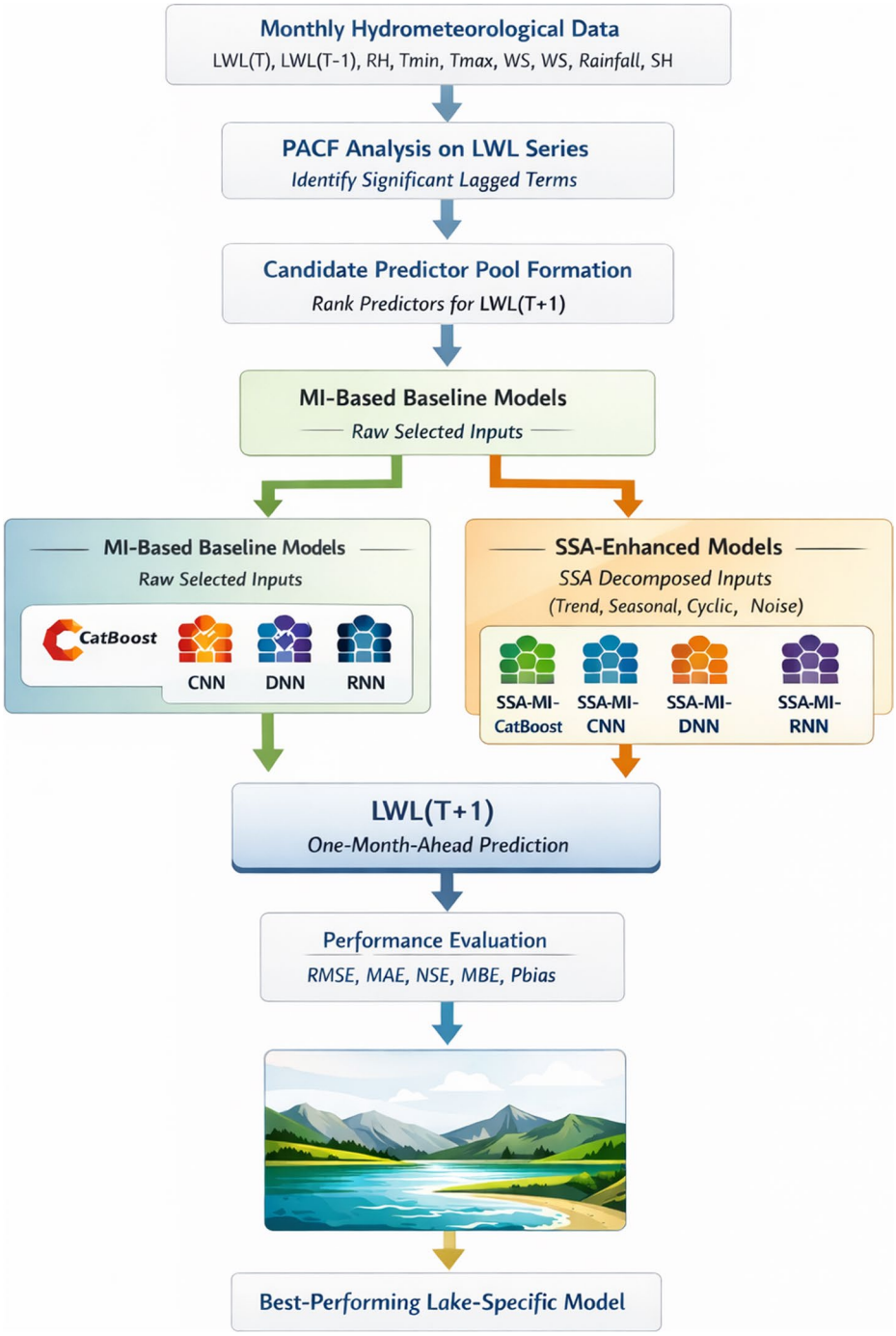


Fig. 4 Flowchart of the present study

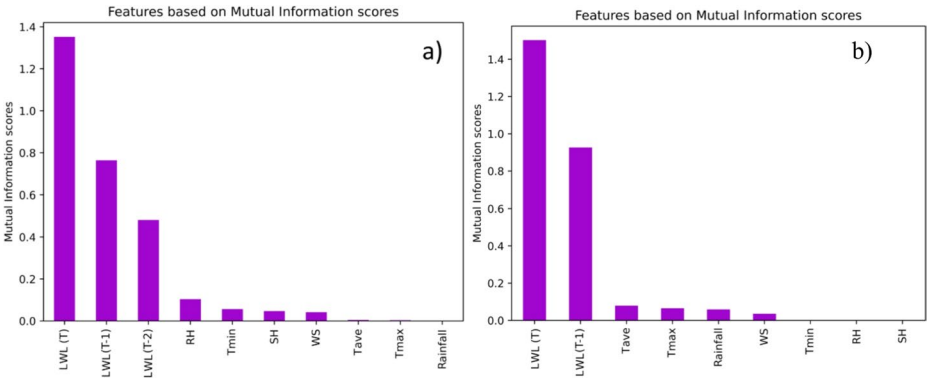


Fig. 5 The importance of features according to mutual information approach: (a) Eber lake, (b) Eğirdir lake

LWL (T-2), RH, Tmin, SH and WS. In Eğirdir Lake, the effective parameters are LWL (T), LWL (T-1), Tave, Tmax, Rainfall, and WS. The features based on MI scores indicating the improtant factors in the prediction of LWL are indicated in Table 2. Accordingly, the highest score was obtained with LWL (T) value in two lakes.

PACF analysis was used to identify statistically meaningful lagged lake water level terms. For Eber Lake, the 1-, 2-, and 3-month lagged LWL values exceeded the 95% confidence limits, whereas for Eğirdir Lake, the 1- and 2-month lagged values exceeded these limits. These lagged terms were then evaluated together with meteorological variables using MI-based feature ranking. For this reason, these values and meteorological parameters showing high internal dependence were subjected to the mutual information test and the most important parameters were revealed (Fig. 6).

Figure 7 presents temperature data (Tmin) belonging to Lake Eber separated into its subcomponents by the SSA method. The raw form of the analyzed minimum temperature data (Tmin) is given. This series, which has a wavy structure, includes the general trends, cyclic structure and noise components of temperature changes. The original series is the combination of all subcomponents. The trend component represents long-term changes. Tmin_Seasonal (Seasonal component) shows that temperatures have a distinct annual cycle. It is seen

Table 2 Features based on MI scores:

Eber Lake		Eğirdir Lake	
Feature	MI score	Feature	MI score
LWL (T)	**1.353**	**LWL(T)**	**1.502**
LWL (T-1)	**0.764**	**LWL(T-1)**	**0.926**
LWL (T-2)	**0.481**	**Tave**	**0.0792**
RH	**0.104**	**Tmax**	**0.0657**
Tmin	**0.057**	**Rainfall**	**0.0592**
SH	**0.047**	**WS**	**0.0360**
WS	**0.042**	Tmin	0.000
Tave	0.005	RH	0.000
Tmax	0.003	SH	0.000
Rainfall	0.000		

Note: Bold characters were selected as input

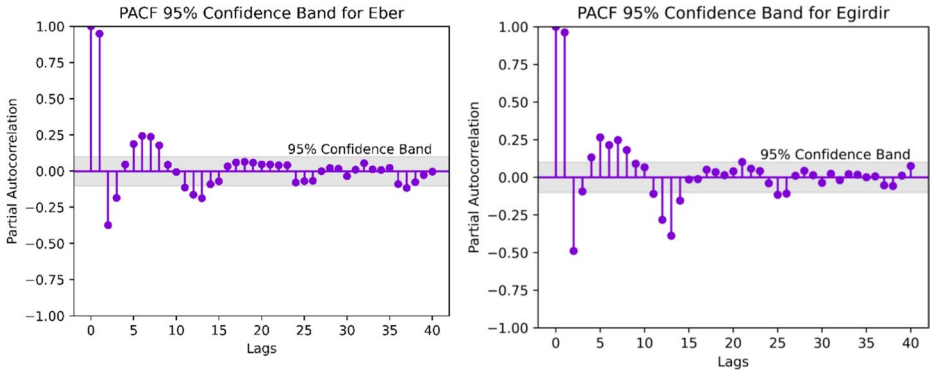


Fig. 6 PACF Plots of LWL times series

that minimum temperature values increase and decrease regularly every year around Lake Eber, therefore seasons have a strong effect on temperature. This reflects the region's characteristic seasonal climate pattern. Tmin_Cyclic (Cyclic Component) represents longer-term and irregular cycles, unlike trend and seasonality. Here, it is seen that there are long-term fluctuations in the Tmin series, and these fluctuations are not regular but tend to increase over time. This may be due to large-scale climate systems (e.g. AMO, PDO or Nino 3.4) or local environmental changes. Tmin_Noise (Noise Component) represents the random fluctuations in the series. This component contains the unexplained variations remaining after the other subcomponents are removed. The noise in the graph consists largely of low-amplitude random fluctuations, but sudden jumps are occasionally observed. These jumps may be due to sudden atmospheric events or measurement errors.

Figure 8 indicates that a RH belonging to Lake Eber is separated into its subcomponents by the SSA. RH (Original Series) has a mixed structure, including seasonal cycles, trend and noise components. The wavy movement indicates that the relative humidity is strongly affected by the atmospheric conditions in the region. RH_Trend (Trend component) represents the long-term changes in relative humidity. The graph shows a wavy long-term trend. This indicates that the relative humidity increases and decreases in certain periods but does not contain a general trend. RH_Seasonal (Seasonal Component) indicates the regular annual cycles in relative humidity. This indicates that there are pronounced seasonal effects in the Eber Lake region and that the humidity level increases or decreases in certain periods of the year. Such a pattern reflects the local climate characteristics and seasonal precipitation regime. RH_Cyclic (Cyclic Component) indicates long-term fluctuations that are not seasonal but occur at regular intervals. The amplitude and frequency of the cycles indicate the recurrence of events in the relative humidity at certain intervals. RH_Noise (Noise Component) represents the random fluctuations in relative humidity.

Figure 9 visualizes LWL values for Lake Eber separated into its subcomponents by the SSA. Original LWL includes seasonal cycles, trends and noise components. The water level exhibits significant fluctuations and trends, indicating that it is affected by long-term atmospheric and hydrological effects. LWL(T)_Trend (Trend component) reflects long-term changes in water level. Here, large fluctuations, rising and falling, are noticeable. LWL(T)_Seasonal (Seasonal Component) represents regular annual cycles. This pattern, which has a distinct repeating structure, indicates that annual cycles such as seasonal precipitation, snow

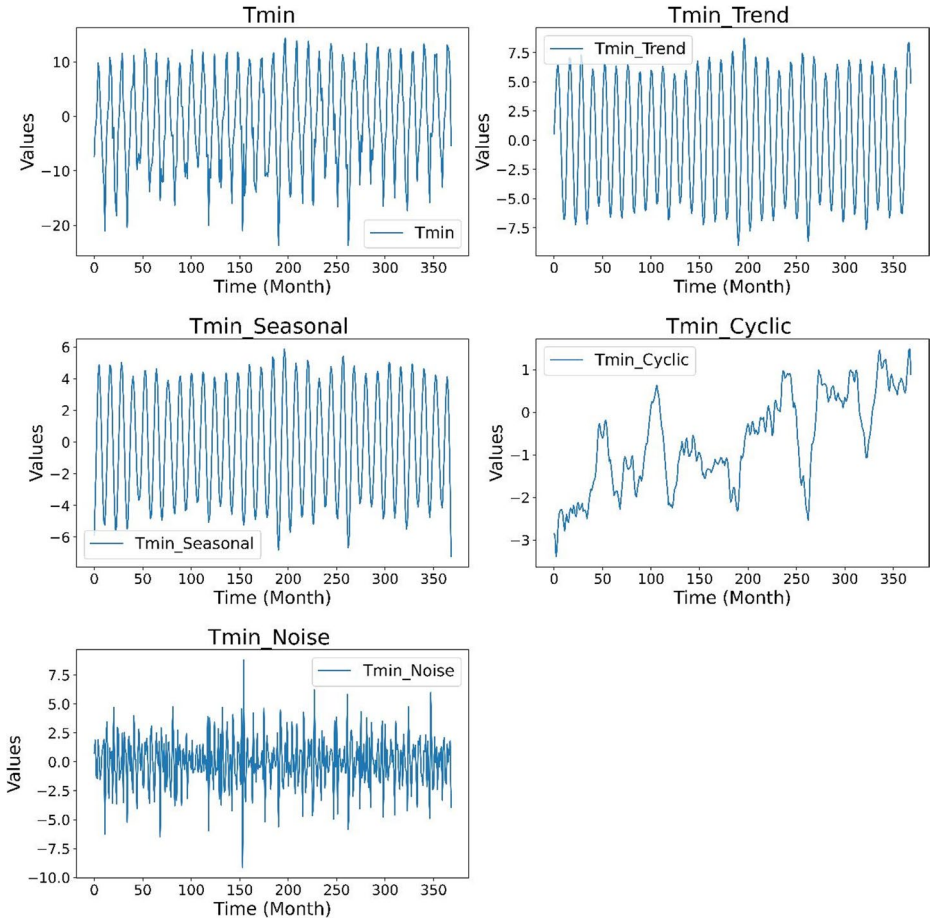


Fig. 7 Subcomponents obtained by SSA of Tmin in Lake Eber

melt or temperature changes affect the water level regularly. LWL(T)_Cyclic (Cyclic Component) represents non-seasonal, medium and long-term cyclic changes. LWL(T)_Noise (Noise Component) represents random and short-term fluctuations in water level. The amplitudes of the noise increased significantly in some periods. This may be due to measurement errors, sudden hydrological events (e.g. floods) or small-scale local variability.

For methodological transparency, the forecasting framework was implemented in two parallel forms. The first form consisted of MI-based baseline models, where the predictors selected by mutual information from the original time series were directly supplied to CatBoost, CNN, DNN, and RNN. The second form consisted of SSA-enhanced models, where those same selected predictors were first decomposed into interpretable SSA components (e.g., trend, seasonal, cyclic, and residual/noise-related components), and the resulting subseries were used as inputs to the same learning algorithms. Thus, the comparison in Table 3 was designed to quantify the incremental benefit of SSA preprocessing beyond MI-based feature selection alone. Table 3 shows the performance evaluation of the lake water level prediction models for the test phase after 1 month. When the prediction results

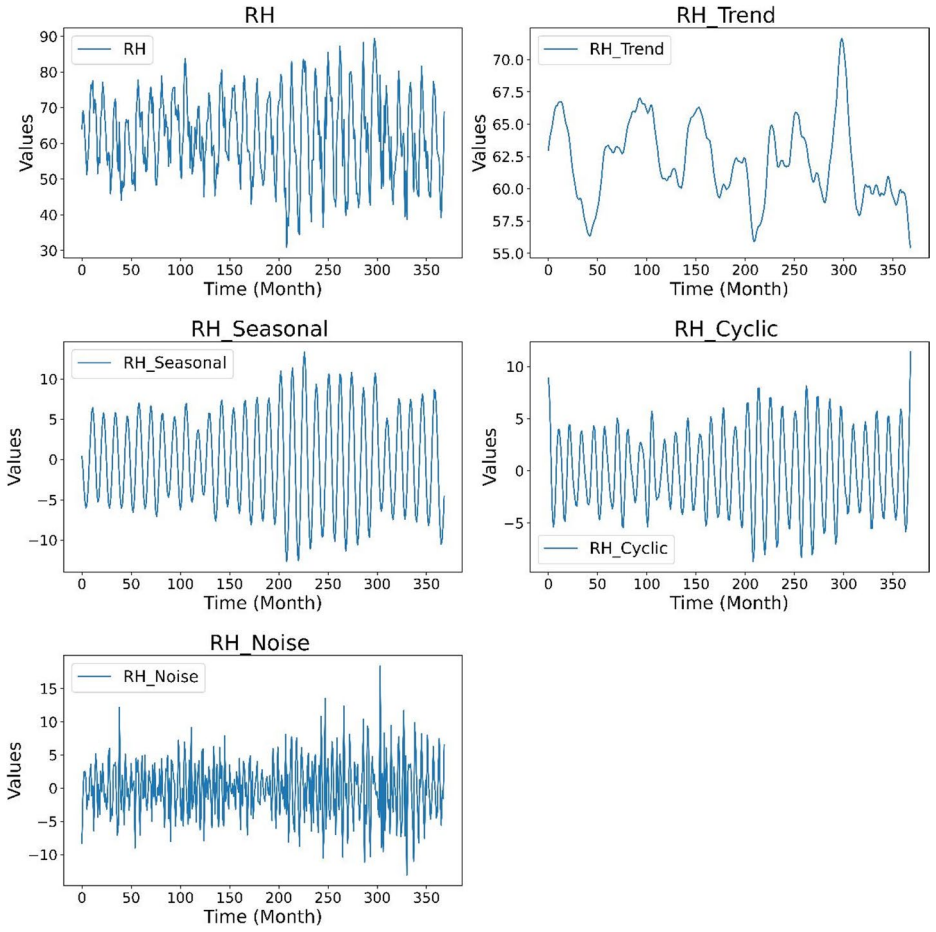


Fig. 8 Subcomponents obtained by SSA of RH in Lake Eber

are compared, it is determined that the SSA technique significantly increases the prediction accuracy of artificial intelligence techniques. According to findings, as seen in Table 3, the most accurate prediction results in Lake Elber were obtained with SSA-MI-CNN with the values of RMSE: 0.107, MAE: 0.069, NSE: 0.937 and MBE: 0.034 Pbias: 0.004. SSA-MI-RNN model showed the most accurate predictions in Ereğli Lake with the values of RMSE: 0.098, MAE: 0.060, NSE: 0.961, MBE: -0.001 and Pbias: 0.000.

Figure 10 (a) shows scatter plot of AI prediction models and observed values in Eber lake. MI-Catboost, MI-CNN and MI-DNN show a very good fit with $R^2 = 0.86$, $R^2 = 0.82$, and $R^2 = 0.87$, respectively. It is seen that the model's predictions have a high correlation with the observed values. The closeness of the points to the fit line indicates good performance. MI-RNN, on the other hand, does not show a linear scatter as the other models and shows a relatively scattered structure, so it can be inferred that weaker predictions are made. Figure 10 (b) visualizes scatter plot of hybrid AI prediction models in Eber Lake. When the scatter plots are examined, the SSA-MI-CNN model algorithm shows the most superior

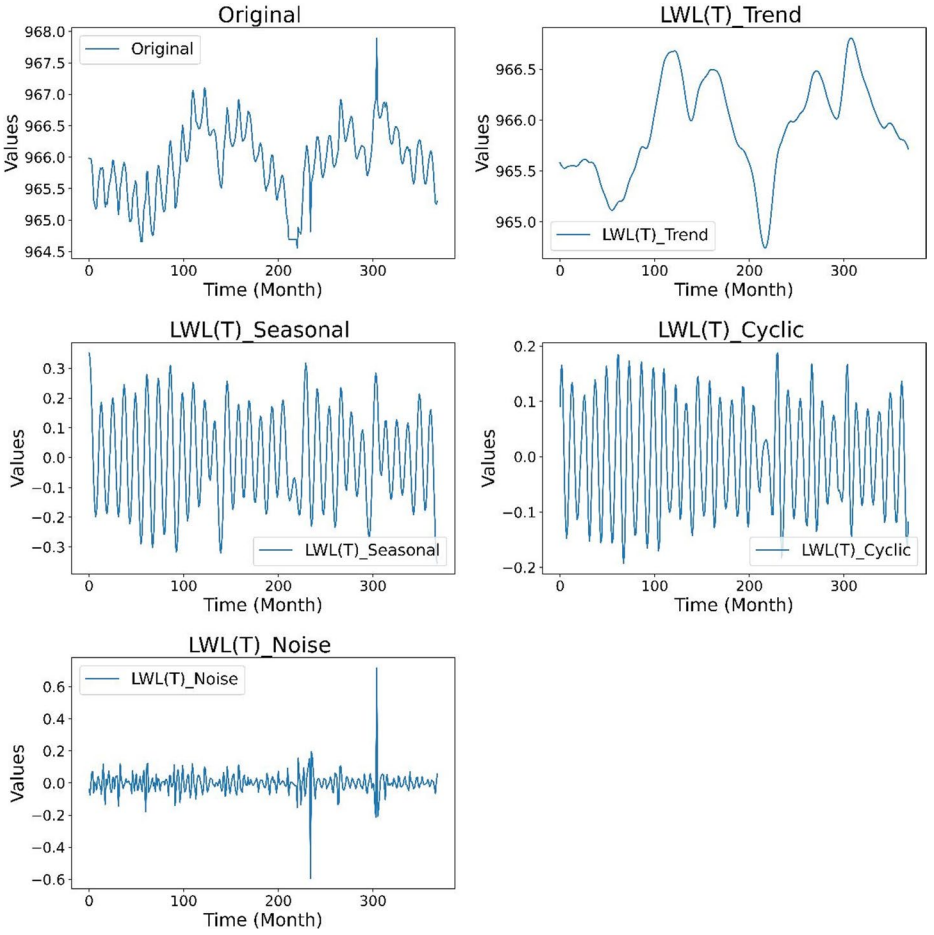


Fig. 9 Subcomponents obtained by SSA of mean temperature in Lake Eber

distribution and relationship with the highest R^2: 0.95 values. SSA-MI-Catboost shows the weakest prediction results and relationship with the lowest R^2: 0.95 values.

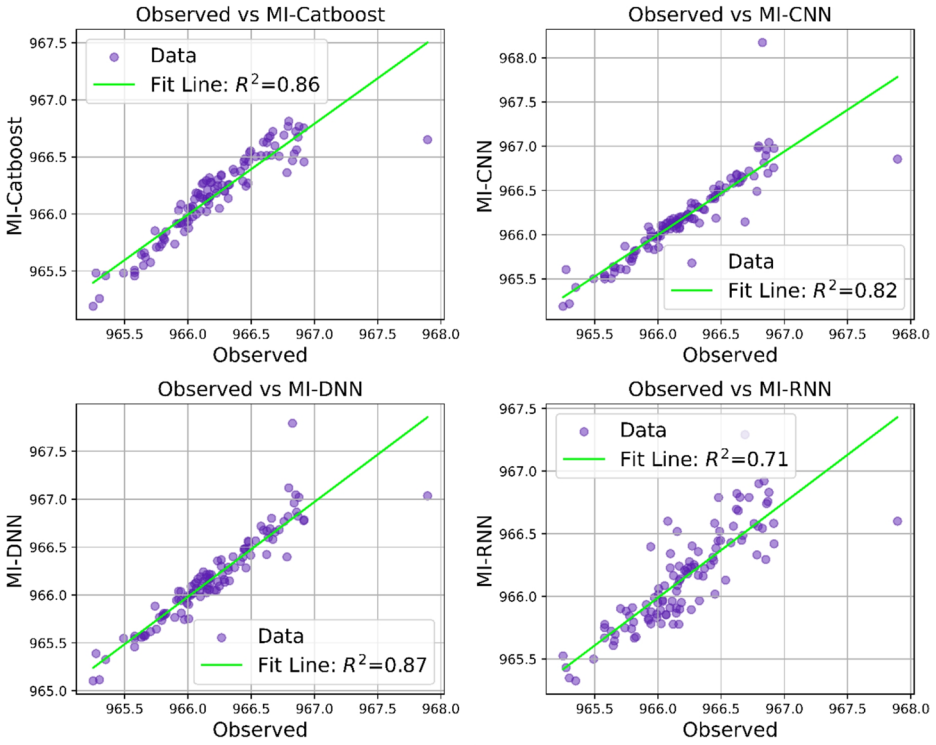
Figure 10 (c) visualizes scatter plot of hybrid AI prediction models in Eğirdir Lake. When the scatter plots are examined, the SSA-MI-RNN model algorithm shows the most superior distribution and relationship with the highest R2: 0.96 values. SSA-MI-DNN represents the weakest estimation results and relationship with the lowest R^2: 0.93 values. Figure 10 (d) shows the scatter plot of AI prediction models and observed values in Eğirdir Lake. MI-Catboost, MI-CNN and MI-DNN are satisfactory results in terms of exhibiting a very close linear fit line with $R^2 = 0.93$, $R^2 = 0.90$ and $R^2 = 0.92$, respectively. It is also seen that the predictions in these three models have a high correlation with the observed values. Also, As seen in Fig. 10, the first three models have 70% accuracy and produce satisfactory predictions, while the MI-RNN model produces relatively weaker predictions than the other models with $R^2 = 0.84$.

Table 3 Comparison of the forecast results of the lake water level test phase

Eber Lake								
	MI-based ML models				MI and SSA-based ML models			
Statist.	MI-Catboost	**MI-CNN**	MI_DNN	MI-RNN	SSA-MI-Catboost	**SSA-MI-CNN**	SSA-MI-DNN	SSA-MI-RNN
RMSE	0.170	0.193	**0.163**	0.238	0.229	**0.107**	0.123	0.144
MAE	0.098	**0.088**	0.097	0.163	0.154	**0.069**	0.086	0.098
NSE	0.841	0.795	**0.854**	0.687	0.711	**0.937**	0.916	0.885
MBE	-0.047	**-0.013**	-0.021	-0.060	-0.103	**0.034**	-0.005	0.043
Pbias	-0.005	**-0.001**	-0.002	-0.006	-0.011	0.004	**0.000**	0.004
Eğirdir Lake								
Statist.	MI-Catboost	MI-CNN	**MI_DNN**	MI-RNN	SSA-MI-Catboost	SSA-MI-CNN	SSA-MI-DNN	**SSA-MI-RNN**
RMSE	**0.136**	0.158	0.139	0.205	0.134	0.117	0.133	**0.098**
MAE	0.081	0.073	**0.068**	0.134	0.094	0.083	0.081	**0.060**
NSE	**0.924**	0.898	0.921	0.827	0.927	0.944	0.928	**0.961**
MBE	0.005	0.013	**-0.003**	0.033	**0.000**	-0.047	0.016	-0.001
Pbias	0.001	0.001	**0.000**	0.004	0.000	-0.005	0.002	**0.000**

Note: Bold characters show the best model

Eber Lake AI prediction data visualization, based on Bland-Altan graphs during the test phase, is shown in Fig. 11 (a). A Bland-Altman plot evaluates the fit of two measurement methods by plotting the difference between their means. In this case, the two measurement methods are the model's predicted and the actual value. The x-axis represents the mean of the actual and predicted values. The y-axis represents the difference between the actual and predicted values (prediction minus the true). The dashed line in the middle shows the mean of the differences (bias). MI-Catboost estimates are distributed between −0.4 and 0.2, and the estimate differences are generally consistent. MI-CNN estimates are generally concentrated between −0.4 and 0.4 and are symmetric. MI-DNN estimates are between −0.3 and 0.3 and are symmetric. MI-RNN estimates show estimates ranging between −0.5 and 0.4. In general, all models perform similarly, but it can be said that Catboost and DNN models make slightly more consistent estimates. The presence of a few outliers in each model indicates that the models have difficulty in some special cases. Eber Lake SSA-MI-based hybrid prediction framework, prediction data visualization, based on Bland-Altan graphs during the test phase, is shown in Fig. 11 (b). SSA-MI-Catboost results are distributed between −0.5 and 0.25. A negative trend is observed as the mean values increase. SSA-MI-CNN and SSA-MI-DNN estimates are concentrated between −0.2 and 0.2, and it can be inferred that they produce more consistent and superior outputs than the other models. SSA-MI-RNN estimates range from −0.2 to 0.4 and have a highly symmetric structure. Figure 11 (c) presents Eğirdir AI prediction data based on Bland-Altan graphs. MI-DNN and MI-RNN seem to perform best with low bias and better fit. MI-Catboost shows reasonable performance, but there are some outliers.

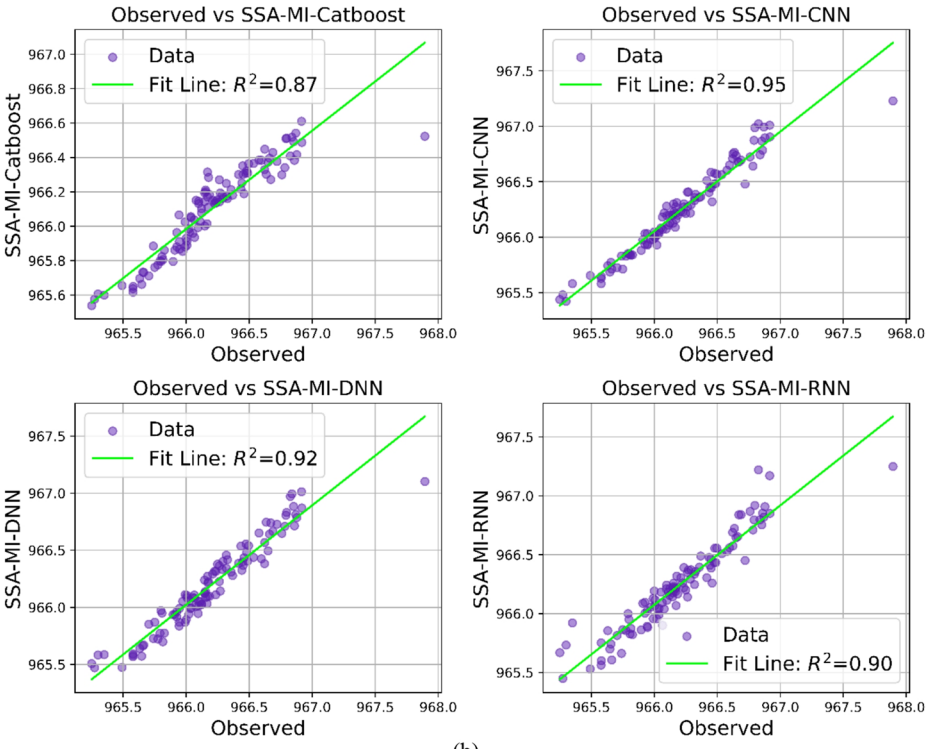


(a)

Fig. 10 Scatter plots of AI and hybrid models in Lakes Eber and Eğirdir: (**a**) AI predicition models in Eber, (**b**) SSA-MI-based hybrid prediction framework in Eber, (**c**) SSA-MI-based hybrid prediction framework in Eğirdir, (**d**) AI predicition models in Eğirdir

Figure 11 (d) indicates Eğirdir hybrid AI prediction data based on Bland-Altan graphs. All four models perform relatively well, with low bias and most points within the limits of fit. SSA-MI-RNN appears to perform slightly better with the tightest data distribution. SSA-MI-Catboost, SSA-MI-CNN, and SSA-MI-DNN have some outliers, especially for the lower mean values, suggesting that these models may be less accurate in this range.

Figure 12 presents the test phase box plots of different models (MI-Catboost, MI-CNN, MI-DNN, MI-RNN) predicting the water level of Lake Eber and the combination of these models with SSA. The box plots are used to compare the central tendency and spread of the predictions of each model and observation. When the median values of the horizontal lines passing through the middle of the boxes are examined, the median of the observed value and the median of the predicted values largely overlap. This indicates that the predictions of the models are quite consistent. However, the models that are most compatible with the observed values were determined to be CNN-based models. The points seen above and below the box represent outliers. The outliers are similar in the observed values and in the individual models except for MI-Catboost. The length of the box and the upper and lower whiskers show the distribution of the model's predictions. In terms of spread, the MI-CNN



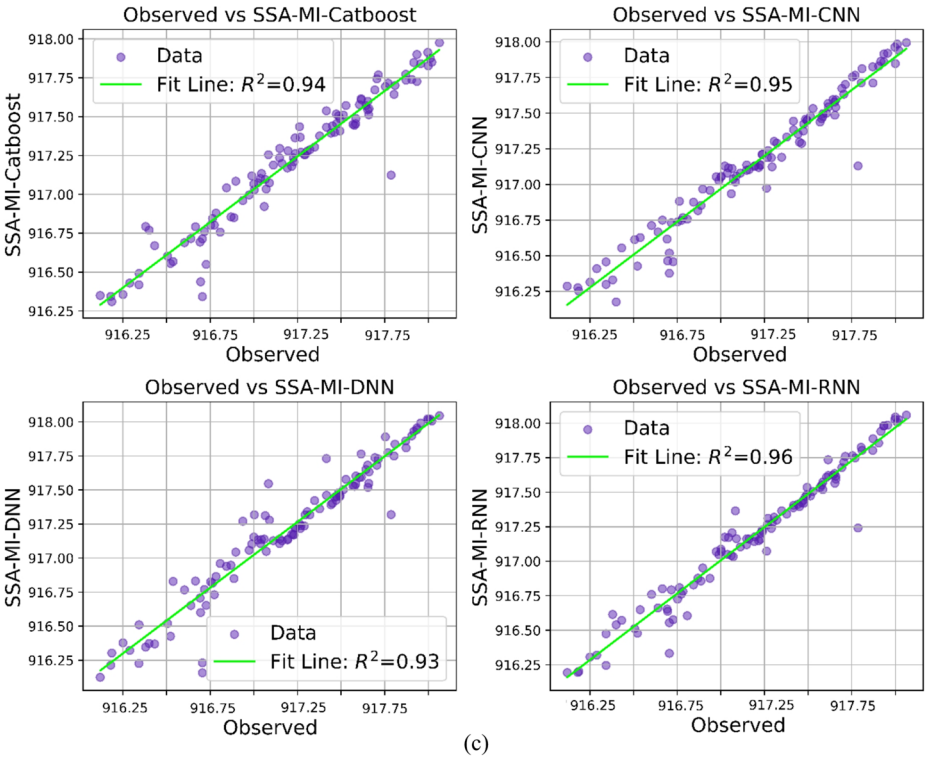
(b)

Figure 10 (continued)

and SSA-MI-CNN models with the closest median to the observed values and a reasonable spread indicate that they perform best. Figure 13 shows the box diagram of the estimation results of the test phase in Eğirdir Lake. Accordingly, according to the data distribution and outliers, it can be concluded that all models except Catboost produce promising results in terms of the data distribution and the similarity of the upper-lower whiskers.

5 Discussion

The present study revealed that integrating SSA with advanced machine learning techniques significantly enhances the accuracy of lake water level forecasting. The specific focus on Eber Lake and Eğirdir Lake demonstrated the effectiveness of using a hybrid model that combines SSA-based deep learning with categorical boosting to manage complex, nonlinear, and seasonal hydrological patterns. In particular, SSA provided an efficient decomposition of time series data into trends, seasonal, and residual components, enabling the models to capture lake-specific patterns with higher accuracy. For instance, the decomposition of water levels for Eber Lake into T + 1 components showed that including lagged data (LWL(T-1) and LWL(T-2) as inputs increased predictive accuracy, corroborating findings

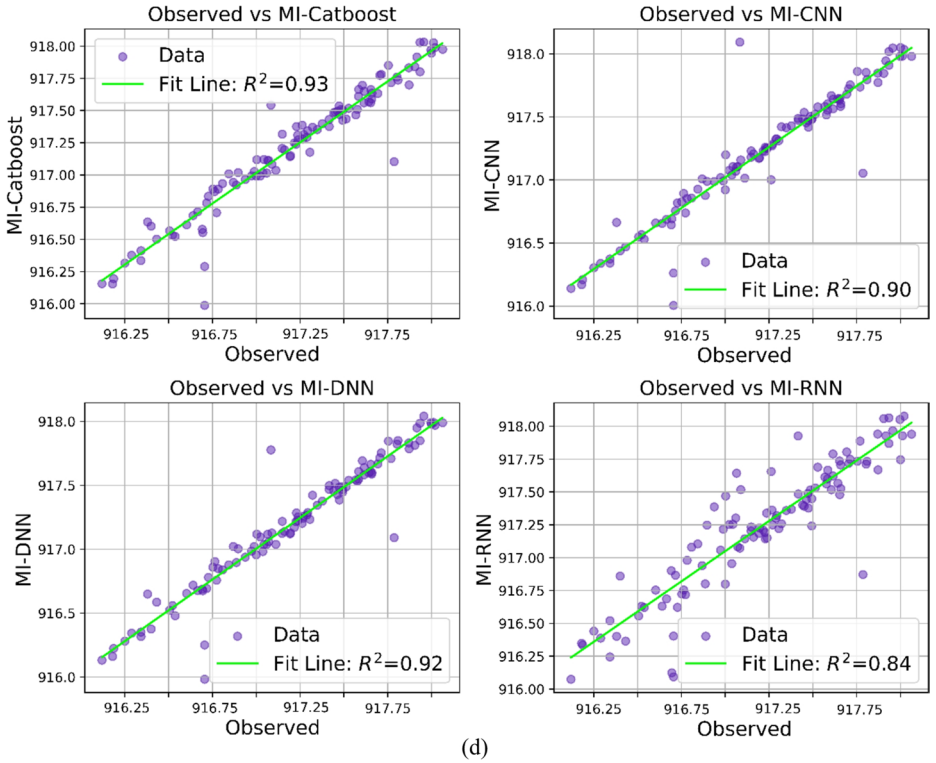


(c)

Figure 10 (continued)

from previous studies that highlight the value of lagged water level inputs in hydrological forecasting (Nguyen et al. 2022).

According to the mutual information (MI) ranking, the most informative predictors for LWL(T + 1) in Eber Lake were LWL(T), LWL(T − 1), LWL(T − 2), RH, Tmin, SH, and WS. For Eğirdir Lake, the highest-ranked predictors were LWL(T), LWL(T − 1), Tave, Tmax, Rainfall, and WS. These variables were selected from the candidate predictor set because they yielded the highest MI scores with the target series, indicating the strongest statistical dependence and information content for one-month-ahead forecasting, rather than a direct causal effect. The models incorporating mutual information feature selection further demonstrated that selecting features with the highest scores, such as LWL(T) and LWL(T-1), for Eber Lake provided superior performance. This process, applied to Eğirdir Lake as well, allowed the model to prioritize essential predictors while reducing noise, thus improving accuracy. Mutual information-based models, such as MI-CNN and MI-DNN, showed consistently better performance in capturing lake water level dynamics compared to traditional methods, aligning with studies by Zhang et al. (2011) and Reddy et al. (2022) on the importance of selecting features that maximize relevant information for forecasting.



(d)

Figure 10 (continued)

Deep learning techniques (LSTM, CNN, and RNN) combined with SSA increased prediction performance and yielded high predictive performance in different lakes. For instance, the SSA-MI-CNN and SSA-MI-LSTM models yielded notable accuracy for Eber Lake by capturing temporal and spatial dependencies, which has been a challenging aspect of hydrological modelling in complex lake systems (Barzegar et al. 2021; Kogekar et al. 2021). The effectiveness of hybrid SSA-MI-CatBoost illustrated the robustness of categorical boosting in handling multicollinear and high-dimensional environmental data, making it well-suited for dynamic environments such as Eğirdir Lake, where lake water levels are influenced by various meteorological and geographical factors (Kim et al. 2023; Qian et al. 2023).

It was determined that although the hybrid models yielded high prediction accuracy, there was a noticeable variance in performance in different lakes. For example, Eber Lake showed high accuracy at T + 1 monthly lag, likely due to strong intrinsic dependencies within its water levels, whereas Eğirdir required additional preprocessing to manage the seasonal variability observed in their lake water levels (Altunkaynak 2019; Yadav and Eliza 2017). This sensitivity to lake-specific characteristics suggests that incorporating local hydrological attributes may improve model generalizability, especially for lakes with unique seasonal fluctuations.

Although the proposed SSA-based hybrid models achieved high forecasting accuracy, several limitations should be noted. The study was conducted using data from only two lakes, which may limit the generalizability of the findings to other hydrological systems.

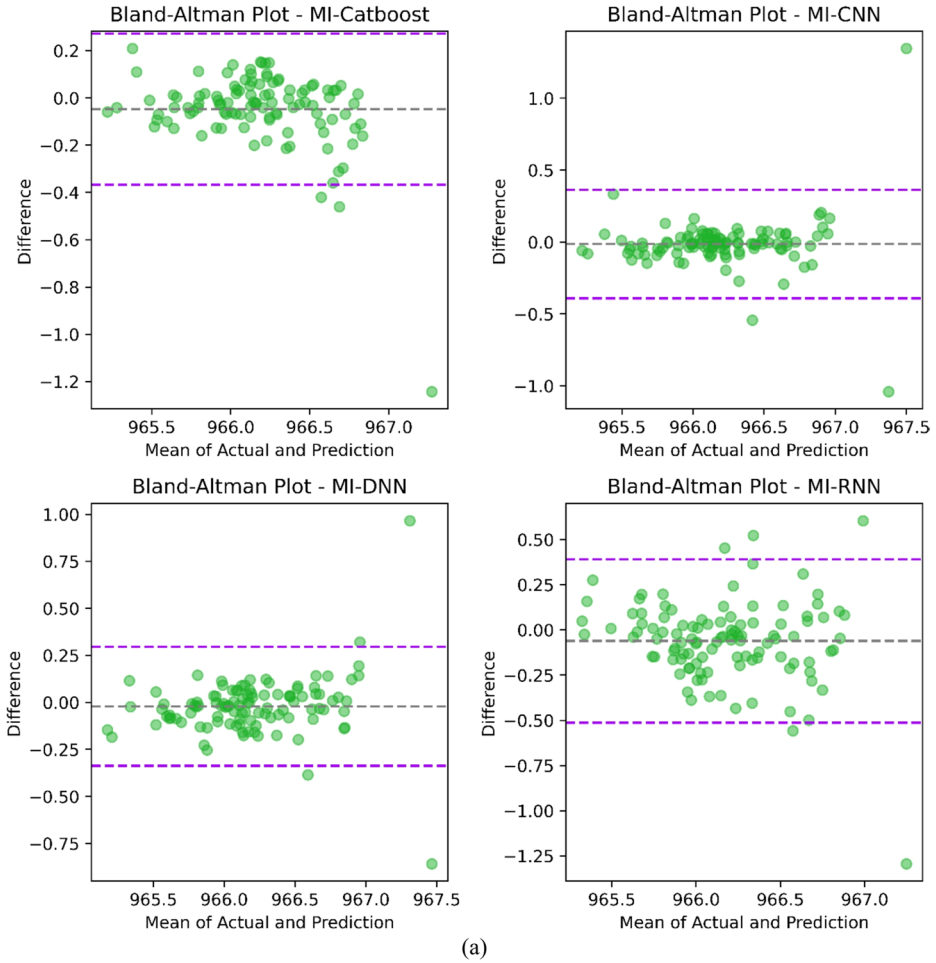


Fig. 11 Visualization of AI and SSA-enhanced hybrid models in Lakes Eber and Eğirdir based on Bland-Altan graphs: (**a**) AI predicition models in Eber, (**b**) SSA-MI-based hybrid prediction framework in Eber(**c**) AI predicition models in Eğirdir, (**d**) SSA-MI-based hybrid prediction framework in Eğirdir

Model performance also depends on the quality and availability of historical lake water level and meteorological data. Finally, external factors such as water management practices, land-use changes, and anthropogenic interventions were not explicitly considered, which may influence lake water level dynamics and forecasting performance.

6 Conclusion

This study proposes a robust and adaptive framework for monthly lake water level forecasting by integrating Singular Spectrum Analysis (SSA), mutual information-based feature selection, advanced deep learning architectures (DNN, CNN, RNN), and the categorical boosting model CatBoost. The model's unique combination of SSA for data decomposition

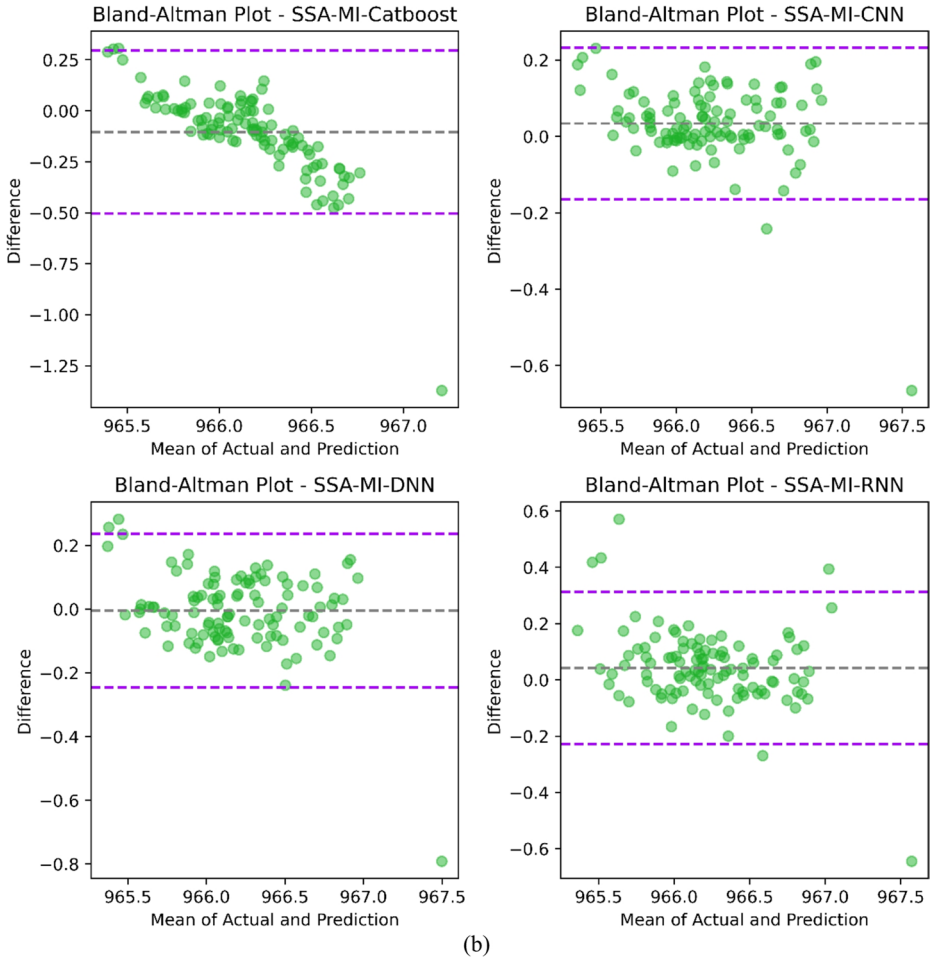


Figure 11 (continued)

and mutual information for optimized feature selection demonstrates its potential to capture the complex, nonlinear patterns inherent in lake water levels across varied hydrological conditions. Applied to Eber Lake and Eğirdir Lake, this approach achieved notable improvements in prediction accuracy, highlighting the value of hybrid models in handling intricate temporal dependencies and spatial variances in hydrological data.

The model's performance underscores the efficacy of SSA in enhancing the predictive power of deep learning by isolating meaningful components from noise and residuals, which, when combined with CNN and LSTM architectures, allow for more precise forecasting even under challenging conditions. The inclusion of mutual information-based selection enables the model to prioritize the most informative variables, contributing to reduced dimensionality and noise, which are critical in maintaining model accuracy across

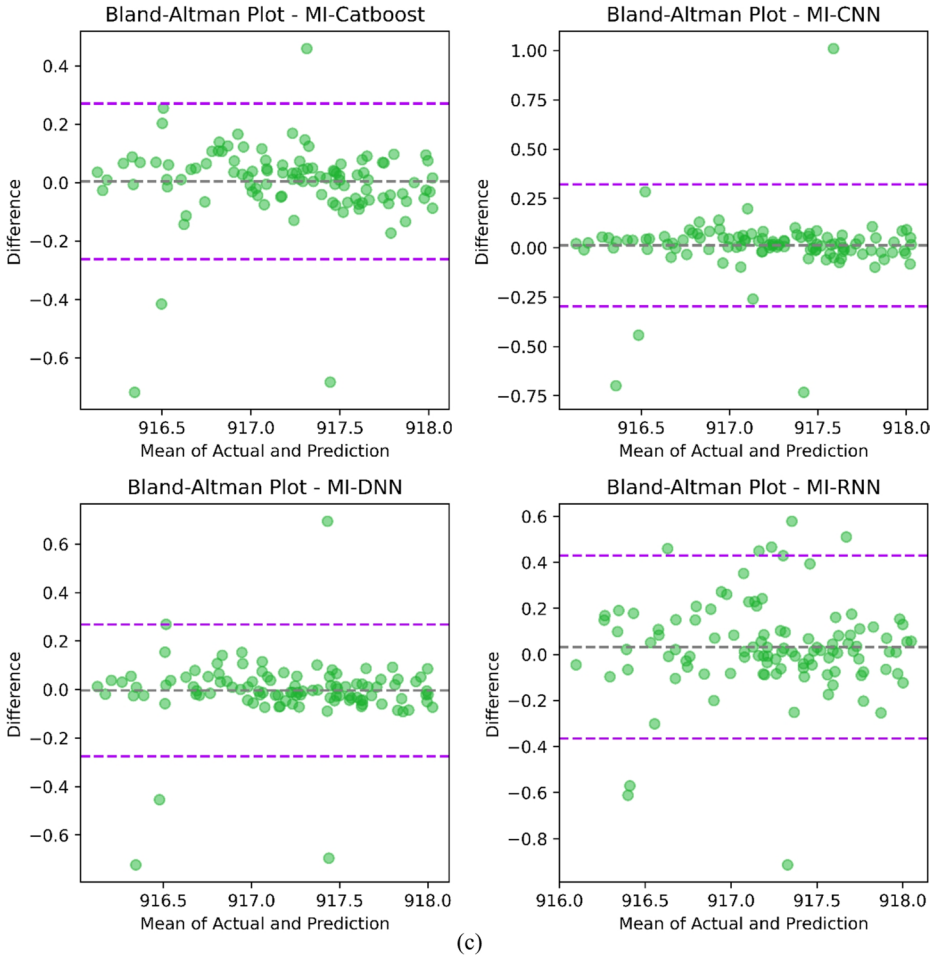
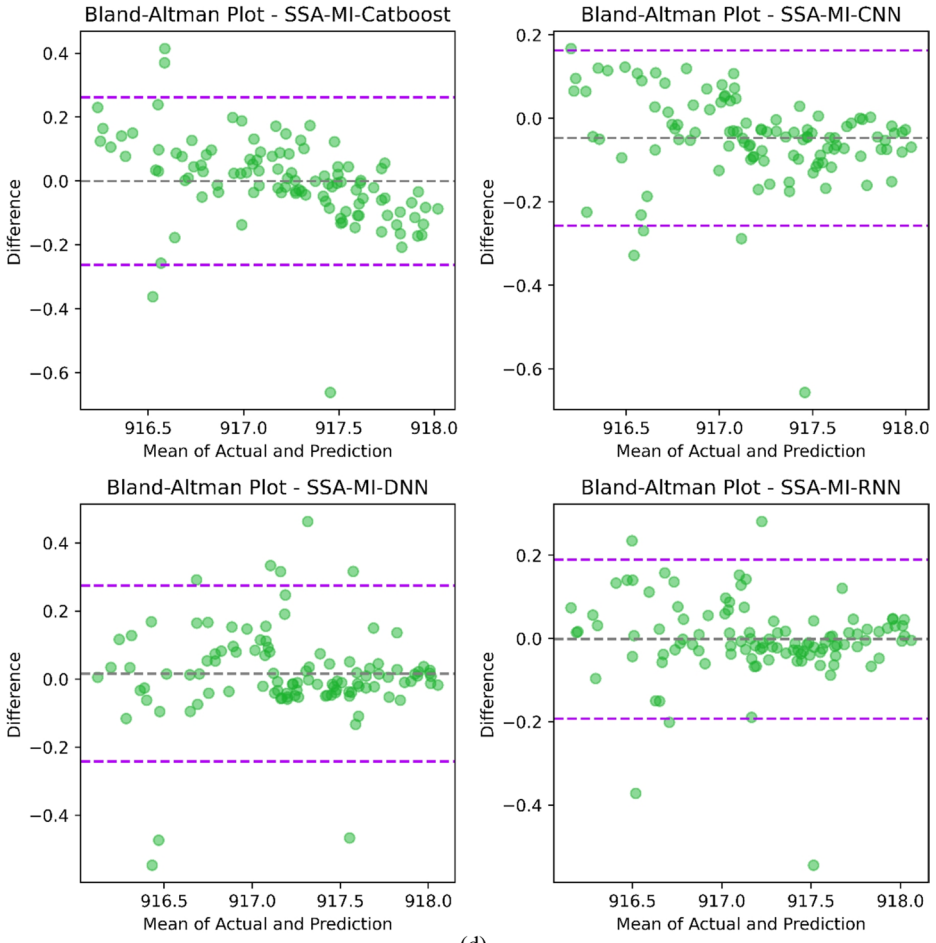


Figure 11 (continued)

different time lag, such as T + 1. Furthermore, CatBoost's ability to handle categorical and high-dimensional data positions it as a valuable asset in managing the diverse environmental and meteorological variables impacting lake water levels.

This research contributes to the existing literature by providing empirical evidence for the effectiveness of SSA-enhanced deep learning models and categorical boosting in water level forecasting. However, it also reveals that model performance can vary depending on lake-specific attributes, such as hydrological regime and seasonal variability, suggesting that future research should focus on adaptive frameworks that dynamically adjust based on the unique characteristics of each lake. Such refinements could further enhance the model's generalizability and utility for broader applications, including flood forecasting and reservoir management.



(d)

Figure 11 (continued)

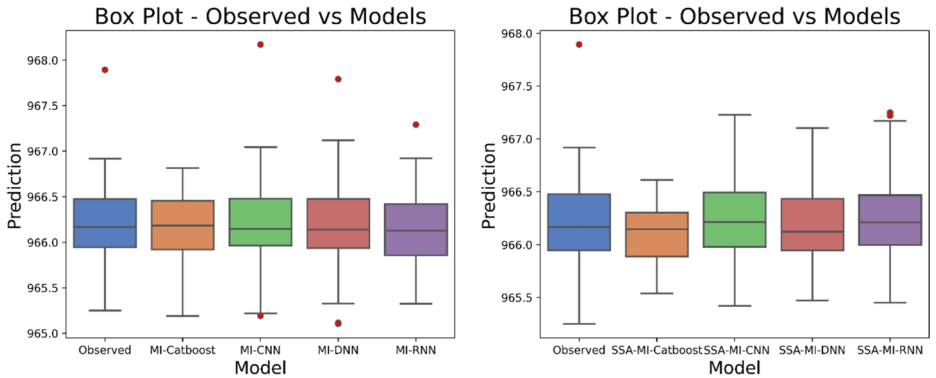


Fig. 12 Model evaluation with box diagonals of the test phase of Lake Eber

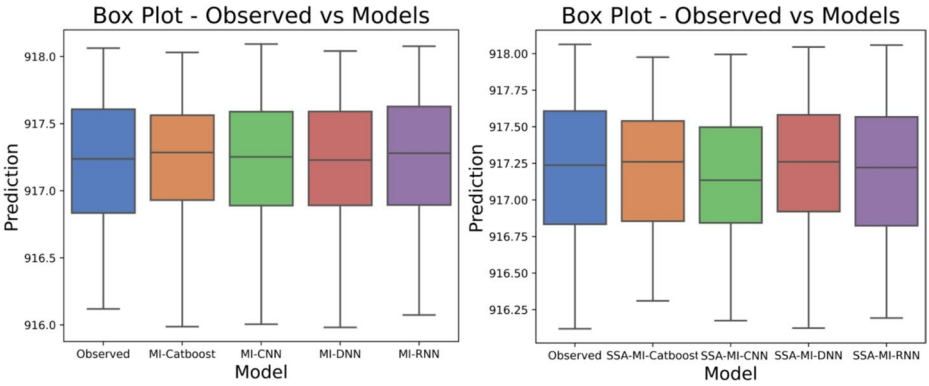


Fig. 13 Model evaluation with box diagonals of the test phase of Eğirdir

In conclusion, the proposed hybrid model not only offers a powerful tool for lake water level forecasting under diverse conditions but also sets a foundation for further exploration into SSA-based deep learning ensembles in hydrological and environmental modelling. The findings underscore the potential for integrated machine learning frameworks to transform predictive accuracy in water resource management, addressing a critical need for resilience in the face of climate variability and environmental change.

Author Contributions Conceptualization, VK and OMK; Investigation, VK and OMK; Methodology, TT; Visualization, OMK and TT; Writing – original draft, VK, VSY and TT; Writing – review & editing, VK, OMK and VSY.; Supervision, Editing SA. All authors have read and agreed to the submitted version of the manuscript.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). This study has not received any funding from any institution/organization.

Data Availability Thanks to General Directorate of State Water Works (DSI) and Meteorology (MGM) for providing data. Data obtained from the General Directorate of State Hydraulic Works (DSI) is not available to the public. Thanks to Grammarly Premium and ChatGPT for helping correct grammatical and spelling errors. Thanks to the free version of the Deepl translator it was used to translate the article into English.

Declarations

Ethics approval Not applicable.

Consent for Publication Not applicable.

Consent to participate Not applicable.

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Clinical trial number Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H (2018) State-of-the-art in artificial neural network applications: A survey. Heliyon 4:e00938

Ahmed AM, Deo RC, Ghahramani A, Feng Q, Raj N, Yin Z, Yang L (2022) New double decomposition deep learning methods for river water level forecasting. Sci Total Environ 831:154722

Ahn JM, Kim J, Kim K (2023) Ensemble machine learning of gradient boosting (XGBoost, LightGBM, CatBoost) and attention-based CNN-LSTM for harmful algal blooms forecasting. Toxins 15(10):608

Aksoy H, Unal NE, Eris E, Yuce MI (2013) Stochastic modeling of Lake Van water level time series with jumps and multiple trends. Hydrol Earth Syst Sci 17:2297–2303

Altunkaynak A (2019) Predicting water level fluctuations in Lake Van using hybrid season-neuro approach. J Hydrol Eng, *24*(8)

Atashi V, Gorji HT, Shahabi SM, Kardan R, Lim YH (2022) Water level forecasting using deep learning time-series analysis: A case study of Red River of the North. *Water, 14*(12), 1971

Atilgan A, Yucel A, Markovic M (2020) Determination of relationship between water level, volume and meteorological variables: Study of Lake Egirdir. In *Proceedings of the 19th International Scientific Conference Engineering for Rural Development* (pp. 140–146). Jelgava, Latvia

Awad M, Khanna R (2015) Support vector regression. In *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (pp. 67–80). Apress

Barzegar R, Aalami MT, Adamowski J (2021) Coupling a hybrid CNN-LSTM deep learning model with a boundary corrected maximal overlap discrete wavelet transform for multiscale lake water level forecasting. J Hydrol 598:126196

Bonakdari H, Ebtehaj I, Samui P, Gharabaghi B (2019) Lake water-level fluctuations forecasting using minimax probability machine regression, relevance vector machine, Gaussian process regression, and extreme learning machine. Water Resour Manage 33:3965–3984

Chen YC, Yeh HC, Kao SP, Wei C, Su PY (2023) Water level forecasting in tidal rivers during typhoon periods through ensemble empirical mode decomposition. Hydrology 10(2):47

Cho M, Kim C, Jung K, Jung H (2022) Water level prediction model applying a long short-term memory (LSTM)–gated recurrent unit (GRU) method for flood prediction. Water 14(14):2221

Danandeh Mehr A, Rikhtehgar Ghiasi A, Yaseen ZM, Sorman AU, Abualigah L (2023) A novel intelligent deep learning predictive model for meteorological drought forecasting. J Ambient Intell Humaniz Comput 14:10441–10455

Dasi H, Ying Z, Yang B (2023) Predicting the consumed heating energy at residential buildings using a combination of categorical boosting (CatBoost) and metaheuristics algorithms. J Building Eng 71:106584

Demir S, Sahin EK (2023) Predicting occurrence of liquefaction-induced lateral spreading using gradient boosting algorithms integrated with particle swarm optimization: PSO-XGBoost, PSO-LightGBM, and PSO-CatBoost. Acta Geotech 18(6):3403–3419

Dorogush AV, Ershov V, Gulin A (2018) CatBoost: Gradient boosting with categorical features support. arXiv

Elman JL (1993) Learning and development in neural networks: The importance of starting small. Cognition 48(1):71–99

Ergüven R, Aydin AB, Avci D (2025) Drought forecasting using standard precipitation index and artificial intelligence models in the mediterranean region of Türkiye. Appl Sci 15(22):12172

Fan Z, Gou J, Wang C (2023) Predicting secondary school student performance using a double particle swarm optimization-based categorical boosting model. Eng Appl Artif Intell 124:106649

Fijani E, Khosravi K (2023) Hybrid iterative and tree-based machine learning algorithms for lake water level forecasting. Water Resour Manage 37(14):5431–5457

Ghafarian Malamiri HR, Rousta I, Olafsson H, Zare H, Zhang H (2018) Gap-filling of MODIS time series land surface temperature (LST) products using singular spectrum analysis (SSA). Atmosphere 9(9):334
Hai TX, Van Nghi V, Hung VH, Tuan DN, Lam DT, Van CT (2019) Assessing and forecasting saline intrusion in the Vietnamese Mekong Delta under the impact of upstream flow and sea level rise. J Environ Sci Eng B 8:174–186
Hammer B (2000) On the approximation capability of recurrent neural networks. Neurocomputing 31(1–4):107–123
Hancock JT, Khoshgoftaar TM (2020) CatBoost for big data: An interdisciplinary review. J Big Data 7:94
Hassani H (2007) Singular spectrum analysis: Methodology and comparison. J Data Sci 5(2):239–257
Herath M, Jayathilaka T, Hoshino Y, Rathnayake U (2023) Deep machine learning-based water level prediction model for Colombo flood detention area. Appl Sci 13(4):2194
Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci 79(8):2554–2558
Ibragimov B, Gusev G (2019) Minimal variance sampling in stochastic gradient boosting. Adv Neural Inf Process Syst, *32*
Jamei M, Ali M, Malik A, Prasad R, Abdulla S, Yaseen ZM (2022) Forecasting daily flood water level using hybrid advanced machine learning based time-varying filtered empirical mode decomposition approach. Water Resour Manage 36(12):4637–4676
Katipoğlu OM, Keblouti M, Mohammadi B (2023) Application of novel artificial bee colony optimized ANN and data preprocessing techniques for monthly streamflow estimation. Environ Sci Pollut Res 30:89705–89725
Kaur A, Sood SK (2020) Artificial intelligence-based model for drought prediction and forecasting. Comput J 63(11):1704–1712
Kim K, Kim DK, Noh J, Kim M (2018) Stable forecasting of environmental time series via long short-term memory recurrent neural network. IEEE Access 6:75216–75228
Kim D, Park J, Han H, Lee H, Kim HS, Kim S (2023) Application of AI-based models for flood water level forecasting and flood risk classification. KSCE J Civ Eng 27(7):3163–3174
Kogekar AP, Nayak R, Pati UC (2021) A CNN-BiLSTM-SVR based deep hybrid model for water quality forecasting of the River Ganga. In *2021 IEEE 18th India Council International Conference (INDICON)* (pp. 1–6). IEEE
Köyceğiz C, Büyükyıldız M (2022) Estimation of streamflow using different artificial neural network models. Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi 5:1141–1154
Le JA, El-Askary HM, Allali M, Struppa DC (2017) Application of recurrent neural networks for drought projections in California. Atmos Res 188:100–106
Liao J, Gao L, Wang X (2014) Numerical simulation and forecasting of water level for Qinghai Lake using multi-altimeter data between 2002 and 2012. IEEE J Sel Top Appl Earth Observations Remote Sens 7(2):609–622
Lin H, Gharehbaghi A, Zhang Q, Band SS, Pai HT, Chau KW, Mosavi A (2022) Time series-based groundwater level forecasting using gated recurrent unit deep neural networks. Eng Appl Comput Fluid Mech 16(1):1655–1672
Louis YHT, Kuok KK, Imteaz M, Lai WY, Derrick KXL (2019) Development of whale optimization neural network for daily water level forecasting. Int J Adv Trends Comput Sci Eng 8(3):354–362
Mahajan P, Uddin S, Hajati F, Moni MA (2023) Ensemble learning for disease prediction: A review. Healthcare 11(12):1808
Masood A, Ahmad K (2021) A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: Fundamentals, application and performance. J Clean Prod 322:129072
Mohammed SJ, Zubaidi SL, Al-Ansari N, Mohammed Ridha H, Dulaimi A, Al-Khafaji R (2023) Application of metaheuristic algorithms and ANN model for univariate water level forecasting. Adv Civil Eng 2023:9947603
Natarajan N, Dharshini S, Vasudevan M (2026) Prediction of standardized precipitation index based on LSTM and variants of LSTM. Physics and Chemistry of the Earth, Parts A/B/C, p 104422
Najm SM, Trzepieciński T, Kowalik M (2023) Modelling and parameter identification of coefficient of friction for deep-drawing quality steel sheets using the CatBoost machine learning algorithm and neural networks. Int J Adv Manuf Technol 124:2229–2259
Nguyen AD, Le Nguyen P, Vu VH, Pham QV, Nguyen VH, Nguyen MH, Nguyen K (2022) Accurate discharge and water level forecasting using ensemble learning with genetic algorithm and singular spectrum analysis-based denoising. Sci Rep 12(1):19870

Pham QB, Yang TC, Kuo CM, Tseng HW, Yu PS (2021) Coupling singular spectrum analysis with least square support vector machine to improve accuracy of SPI drought forecasting. Water Resour Manage 35:847–868

Polomčić D, Gligorić Z, Bajić D, Cvijović Č (2017) A hybrid model for forecasting groundwater levels based on fuzzy C-mean clustering and singular spectrum analysis. Water 9(7):541

Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A (2018) CatBoost: Unbiased boosting with categorical features. Adv Neural Inf Process Syst, *31*

Qian L, Chen Z, Huang Y, Stanford RJ (2023) Employing categorical boosting (CatBoost) and meta-heuristic algorithms for predicting urban gas consumption. Urban Clim 51:101647

Reddy PCS, Yadala S, Goddumarri SN (2022) Development of rainfall forecasting model using machine learning with singular spectrum analysis. IIUM Eng J 23(1):172–186

Rossi F, Lendasse A, François D, Wertz V, Verleysen M (2006) Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. Chemometr Intell Lab Syst 80(2):215–226

Ruma JF, Adnan MSG, Dewan A, Rahman RM (2023) Particle swarm optimization based LSTM networks for water level forecasting: A case study on Bangladesh river network. Results Eng 17:100951

Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, Summers RM, Giger ML (2019) Deep learning in medical imaging and radiation therapy. Med Phys 46(1):e1–e36

Sapitang M, Ridwan WM, Kushiar F, Najah Ahmed K, A., El-Shafie A (2020) Machine learning application in reservoir water level forecasting for sustainable hydropower generation strategy. Sustainability 12(15):6121

Sener E, Davraz A, Sener S (2010) Investigation of Aksehir and Eber Lakes (SW Turkey) coastline change with multitemporal satellite images. Water Resour Manage 24:727–745

Steuer R, Kurths J, Daub CO, Weise J, Selbig J (2002) The mutual information: Detecting and evaluating dependencies between variables. Bioinformatics 18(Suppl 2):S231–S240

Sun M, Li X, Kim G (2019) Precipitation analysis and forecasting using singular spectrum analysis with artificial neural networks. Cluster Comput 22:12633–12640

Tanriverdi I, Batmaz İ (2025) AI-driven US drought prediction using machine learning and deep learning: I. Tanriverdi, I. Batmaz. Climate Dynamics 63(6):249

Tuğrul T, Hinis MA (2024) Improvement of drought forecasting by means of various machine learning algorithms and wavelet transformation. Acta Geophysica. Advance online publication

Tuğrul T, Selek B, Hınıs MA, Selek Z, Oruç S (2025) Data-driven drought prediction by means of machine learning techniques and increasing accuracy with wavelet transform: T. Tuğrul et al. Pure Appl Geophys 182(10):4319–4341

Vionnet V, Fortin V, Gaborit E, Roy G, Abrahamowicz M, Gasset N, Pomeroy JW (2020) Assessing the factors governing the ability to predict late-spring flooding in cold-region mountain basins. Hydrol Earth Syst Sci 24(4):2141–2165

Wardah W, Khan MGM, Sharma A, Rashid MA (2019) Protein secondary structure prediction using neural networks and deep learning: A review. Comput Biol Chem 81:1–8

Wei X, Rao C, Xiao X, Chen L, Goh M (2023) Risk assessment of cardiovascular disease based on SOLSSA-CatBoost model. Expert Syst Appl 219:119648

Yadav B, Eliza K (2017) A hybrid wavelet-support vector machine model for prediction of Lake water level fluctuations using hydro-meteorological data. Measurement 103:294–301

Yücel A, Markovic M, Atilgan A, Rolbiecki R, Ertop H, Jagosz B, Jakubowski T (2022) Investigation of annual lake water levels and water volumes with Şen innovation and Mann-Kendall rank correlation trend tests: Example of Lake Eğirdir, Turkey. Water 14(15):2374

Zhang Q, Wang BD, He B, Peng Y, Ren ML (2011) Singular spectrum analysis and ARIMA hybrid model for annual runoff forecasting. Water Resour Manage 25:2683–2703

Zhang B, Zhang H, Zhao G, Lian J (2020) Constructing a PM2.5 concentration prediction model by combining auto-encoder with Bi-LSTM neural networks. Environ Model Softw 124:104600

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

Authors and Affiliations

Okan Mert Katipoğlu[1] · Veysi Kartal[2] · Türker Tuğrul[3] · Veysel Süleyman Yavuz[2] · Sema Ariman[4]

✉ Veysi Kartal
veysikartal@siirt.edu.tr

Okan Mert Katipoğlu
okatipoglu@erzincan.edu.tr

Türker Tuğrul
turkertugrul@gazi.edu.tr

Veysel Süleyman Yavuz
vsyavuz@hotmail.com

Sema Ariman
sema.ariman@samsun.edu.tr

[1] Department of Civil Engineering, Erzincan Binali Yıldırım University, Erzincan, Türkiye

[2] Department of Civil Engineering, Faculty of Engineering, Siirt University, 56000 Siirt, Türkiye

[3] Department of Civil Engineering, Faculty of Technology, Gazi University, Ankara 06560, Türkiye

[4] Department of Meteorological Engineering, Özdemir Bayraktar Faculty of Aeronautics and Astronautics, University of Samsun, Samsun 55000, Türkiye